A Close Examination of Jo Boaler's Railside Report

Wayne Bishop, Paul Clopton, R. James Milgram

Abstract—**Jo Boaler, an Associate Professor at the Stanford School of Education has just published an already well known study of three high schools that she called Hillside, Greendale, and Railside. This study makes extremely strong claims for discovery style instruction in mathematics, and consequently has the potential to affect instruction and curriculum throughout the country.**

As is the case with much education research of this nature, Prof. Boaler has refused to divulge the identities of the schools to qualified researchers. Consequently, it would normally be impossible to independently check her work. However, in this case, the names of the schools were determined and a close examination of the actual outcomes in these schools shows that Prof. Boaler's claims are grossly exaggerated and do not translate into success for her treatment students. We give the details in the following article.

Other papers where the researchers have refused to divulge such details as the names of the schools to qualified researchers have affected and continue to affect education policy decisions at the school, state and even national levels. Among these papers are *Standards, Assessments – and What Else? The Essential Elements of Standards-Based School Improvement*, **D. Briars - L. Resnick, CRESST Technical Report 528, (2000) which has been cited repeatedly as justification for the adoption of Everyday Mathematics in school districts throughout the country, and** *The impact of two standards-based mathematics curricula on student achievement in Massachusetts.* **D. Perda, P. Noyce, J. Riordan, J. for Research in Mathematics Education, 32(2001), p. 368-398. which has been used to justify the adoption of the mathematics program "Investigations," developed by TERC. It is worth noting that currently about 19% of U.S. elementary students use Everyday Mathematics and between 6% and 9% use Investigations, including many of our inner city schools.**

If we are to reverse the woeful performance of our students it seems crucial that K-12 education research be subject to the same high standards as are the norm in medicine and the sciences. As a key step we believe that the analysis here shows the dangers of accepting the legitimacy of articles such as those mentioned above as long as the results cannot be independently studied and verified.

RECENTLY, Dr. Jo Boaler, an Associate Professor at the Stanford School of Education, gave a standing room only talk at the National Meeting of the NCTM[1] in Anaheim, CA, and received a tremendously enthusiastic response. She had just explained how the results of her NSF[2] funded study of three high schools in Northern California had verified the most controversial of the deeply held beliefs of this country's education schools about mathematics education.

The study followed the cohort of students at each school that started ninth grade in 2000, used a non-standard set of four tests, interviews, and a number of group projects to determine the mathematical competency of these students. The students at two of the schools were said to have been taught mathematics using mainly "traditional methods," but the students at the third were supposed to have been taught using "reform methods" with group learning and detracking, based largely on CPM,[3] a high school mathematics curriculum developed at the University of California, Davis.

The third school, which was called Railside in the study, was claimed to have much poorer demographics than the other two, and the students, when tested at the beginning of ninth grade, did significantly worse than the students at the other two schools. However, by the end of the tenth grade these students were outperforming the students at the other two schools.

Afterwards, at least two mathematicians that attended the lecture expressed deep scepticism though the prevalent response was unbridled enthusiasm.

Unfortunately, the mathematicians are correct. Dr. Boaler kept the names of the schools private and asked that everyone trust that she had faithfully recorded the outcomes of her study. We were able to determine the identities of these schools. Then we studied the considerable amount of data in the California data base relating to these schools, as well as data requested through the Freedom of Information act or the California Public Records Act. This data includes things like school rankings, demographic data, SAT I outcomes, AP outcomes and even student level outcomes. Further, the results of the students from each of these schools on the entry level CSU[4] math skills test are available. The totality of this data does not support her conclusions.

Indeed, there is only one year in the last five where any of these various measures for any cohort of students gives any advantage to the Railside students - the CST[5] Algebra I exams for the ninth grade students in 2003 - and this is the only test data from that California database which is reproduced in Prof. Boaler's report even though these data cannot represent the cohort that is the focus of the report.

We also found evidence that Dr. Boaler obtained her results by focusing on essentially different populations of students

Wayne Bishop is a Professor of Mathematics Emeritus, Cal. State University, LA, e-mail: wbishop@calstatela.edu

Paul Clopton, VAMP, San Diego

R. James Milgram is a Professor of Mathematics Emeritus, Stanford University, USA. e-mail: milgram@math.stanford.edu.

[1] *National Council of Teachers of Mathematics*

[2] *National Science Foundation*

[3] *College Preparatory Mathematics http://www.cpm.org*

[4] *CSU is the California State University system*

[5] *California Standards Test*

at the three schools. At Railside, her population appeared to consist primarily of the upper two quartiles, while at the other two schools the treatment group was almost entirely contained in the two middle quartiles.

Dr. Boaler put her four tests up on her web-site.[6] We studied them using standard methods to verify their content validity and the two post-tests that were relevant to her study were found to be, on average, roughly 3 years below the grade level where they were being administered. Additionally, all four exams had serious mathematical difficulties - including mathematically incorrect problems, problems with hidden assumptions, a multiple choice problem with all answer choices incorrect, and a problem that was incorrectly graded since, even though the problem was not entirely well posed, it was solvable. However, the actual solution was far more difficult than the graders seemed to understand.

Next, we compared the topics covered on these exams with the recommended content in Algebra I and Geometry. We found that very few of the key topics in these courses were covered in even the most superficial ways on her exams. Thus, the content validity (see Appendix 3) of these tests was shown to be extremely small. Moreover, when we compared the results of these tests against standard measures such as SAT I scores, the predictive validity of these tests was also shown to be negligible.

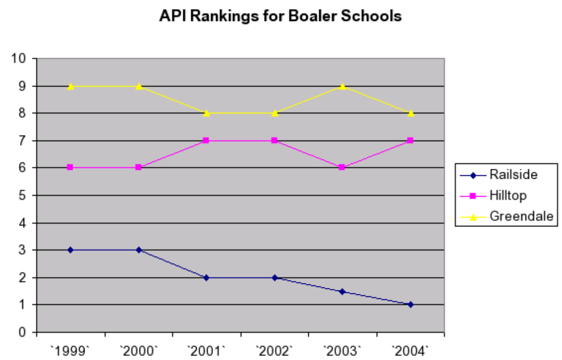
Thus, whatever was being tested by these exams was unlikely to be relevant to student outcomes in the sense of their being able to actually use school mathematics either in their continuing educations or in their daily lives.

Using the style in which the Railside report is written we can summarize by saying

> *Fortunately, there is a clarifying post-script that can be added to the claims of high achievement in mathematics at Railside made in the report. Despite the Railside students' achievement on un-validated tests tailored to favor their program and assessing low-level skills, the students in fact do not perform well on state tests that are more carefully developed and standards-aligned or on AP or SAT exams. This, together with a careful review of the test items used in the study, makes one extremely skeptical of the value of the Railside study's tests for assessing achievement in mathematics. The Railside students show through AP, SAT, and state assessments that they do not have a good understanding of mathematics. This phenomenon speaks more to the flawed nature of the tests used in the study than it does to any claim of adequacy or inadequacy in the reform approach at Railside.*

I. A Discussion of standard measures of student achievement at the three schools

The first measures worth noting are the API and Similar School rankings for the three schools. (API scores are a weighted sum of scores on a number of separate test instruments. A description of these tests and their weights in the API is given in Appendix 5.) 1 is the lowest possible API ranking, and 10 is the highest.



In detail, the California API and similar school rankings for Railside are as follows:

Railside	*API*	*SIM SCHOOL*
1999	3	6
2000	3	2
2001	2	1
2002	2	3
2003	*	*
2004	1	1

The API ranking represents a decile, with 1 representing schools that score in the *lowest* 10 percent of all California high schools on the API exams.[7] The Similar School ranking is also in deciles, but here the comparison is with the 100 most similar high schools in California.[8] So a score of (1, 1) means a school in the lowest 10 percent among all California high schools, and in the lowest 10 percent among all similar high schools. By contrast here are the API results for Greendale and Hilltop:

Greendale	*API*	*SIM SCHOOL*
1999	9	5
2000	9	4
2001	8	2
2002	8	1
2003	9	3
2004	8	4

[6] *http://www.stanford.edu/∼joboaler/*

[7] *In 2003 an insufficient number of students were tested at Railside and the rankings were not recorded. We have interpolated the ranking as 1.5 in the graph.*

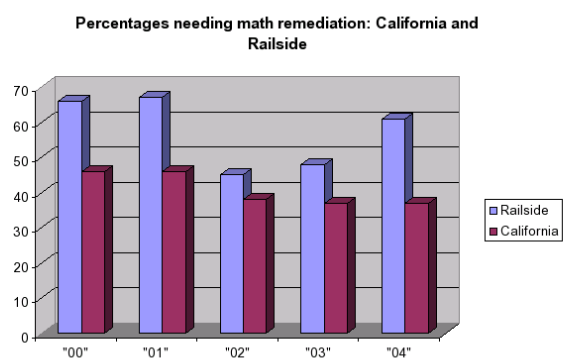
[8] *This is determined by comparing about 70 factors including demographics and teacher backgrounds.*

$$\begin{array}{l} Hilltop \\ 1999 \\ 2000 \\ 2001 \\ 2002 \\ 2003 \\ 2004 \end{array} \begin{array}{c} \begin{array}{cc} API & SIM\ SCHOOL \end{array} \\ \begin{pmatrix} 6 & 7 \\ 6 & 7 \\ 7 & 5 \\ 7 & 8 \\ 6 & 3 \\ 7 & 4 \end{pmatrix} \end{array}$$

REMARK: It is worth looking at the total group of California high schools, (excluding charter schools, alternative schools, and other special needs schools) with rankings of 1 on the API and 1 among similar schools in 2004. There are 16 of them including Railside. Among these schools, Railside had the second highest percentage of highly qualified teachers at 97%, tied for highest level of average parental education, had, by a full 8%, the lowest number of free lunch students and had the fourth smallest number of ESL students.

On the basis of these rankings, there is reason to be concerned that Railside students would not do well in college.

This concern is reinforced by looking at the remediation rates of students at Railside in the California State University system, where they consistently and significantly had much higher mathematics remediation rates than the state as a whole. This was in spite of the fact that the ELM test which is used to determine the need for students to take a remedial mathematics class is set at about a seventh grade level, and that virtually all the Railside students entering the CSU system in recent years had taken "pre-calculus" or "calculus" at Railside according to both Dr. Boaler and the principal at Railside.[9]



Here the blue bars represent the percentages of Railside students entering the CSU system who required remediation in mathematics while the red bars represent the percentages of all entering students in the CSU system that needed remediation. (The 2004 results above - 61% remediation rate for Railside, 37% overall for the state - are the relevant ones for the cohort in Dr. Boaler's study.)

[9] *Specifically, Dr. Boaler says in the introduction that "By their senior year 41% of Railside students were taking calculus compared to around 27% of students in the other two schools." Since admission to the CSU system is limited to the top 33% of graduates of each high school, this implies that virtually all the students from Railside in the CSU system had taken either a course called pre-calculus or one called calculus.*

There is also a dramatic difference in the average scores on the SAT I Math exam.[10] The average score over the last five years on the SAT I Math exam achieved by Railside students, 460, is over .6 standard deviations lower than the scores over this period at Greendale, 524, and Hilltop, 521. Moreover, the average percentage during the last five years of the graduating class taking the SAT I exams at Railside was 39% while 47% of the Greendale students took the exams, as did 45% of the Hilltop students. The detailed data appear in the following table.

	SAT I scores						
	Year	Enrollment	# tested	percent tested	Verbal	Math	percent >1000
Railside	2003-2004	324	130	40.12	416	457	9.6
	2002-2003	316	123	38.9	430	461	10.1
	2001-2002	331	131	39.6	424	455	9.4
	2000-2001	312	113	36.2	427	464	8.3
	1999-2000	304	106	34.9	427	461	10.2
Greendale	2003-2004	264	141	53.4	544	547	39.4
	2002-2003	267	124	46.4	513	523	25.5
	2001-2002	319	170	53.3	514	524	33.2
	2000-2001	286	127	44.4	523	519	27.3
	1999-2000	302	118	39.1	526	509	24.2
Hilltop	2003-2004	430	201	46.7	509	521	26
	2002-2003	376	204	54.3	524	534	36.4
	2001-2002	489	229	46.8	512	519	26.6
	2000-2001	448	168	37.5	506	509	20.1
	1999-2000	470	189	40.2	498	522	21.5

There are also data available on gender differences in performances on the SAT I exams. Before we give them, here is a quote from the Railside article:

> *There were no gender differences in performance in any of the tests we gave students at any level, and young women were well represented in higher mathematics classes. They made up 50% of students in the advanced classes at Hilltop, 48% at Greendale and 59% at Railside.* RAILSIDE ARTICLE, PAGE 11

Unfortunately, this is not supported by the SAT I data. For the three years where gender information is available we have the following outcomes for Railside:

2003-04

	M	F
Enrolled	169	155
Tested	56	74
Verbal	409	421
Math	480	440

2002-03

	M	F
Enrolled	171	145
Tested	62	61
Verbal	437	422
Math	480	441

2001-02

	M	F
Enrolled	170	161
Tested	54	77
Verbal	433	418
Math	471	445

This data, with males generally outscoring females on both the verbal and mathematics portions of the SAT exam - and strongly so in mathematics - is consistent with the scores for

[10] *It is well known that the SAT I scores have a high correlation with success in college, particularly the first year of college.*

the other two schools, though the scores for Railside are much lower.

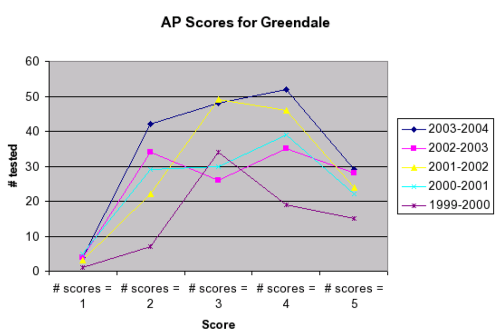
Further the expectation that students from Greendale and Hilltop are likely to outperform students from Railside is confirmed by looking at AP outcomes. These exams are graded with scores of 1, 2, 3, 4, or 5, with 5 being the highest, and 1, 2, being generally considered failing scores. No students at Railside took an AP Calculus exam during the last five years. Greendale was only able to provide us with detailed results for 2003 and 2004, but Hilltop provided us with data for 2000 - 2004. The numbers for the Calculus AB exam are

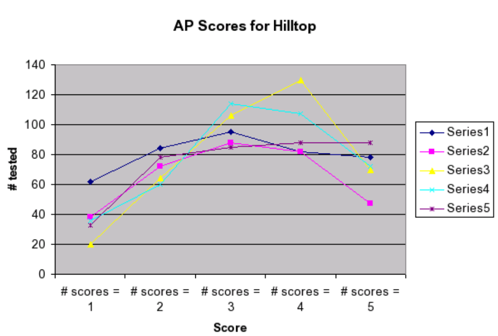
$$\begin{array}{ll} Greendale & \begin{array}{ccccc} 5 & 4 & 3 & 2 & 1 \end{array} \\ \begin{array}{l} 2004 \\ 2003 \end{array} & \begin{pmatrix} 7 & 6 & 6 & 5 & 1 \\ 11 & 1 & 2 & 1 & 0 \end{pmatrix} \end{array}$$

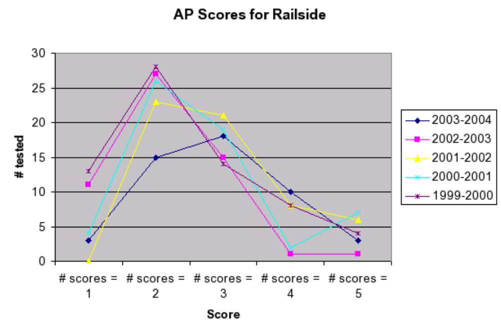
$$\begin{array}{ll} Hilltop & \begin{array}{ccccc} 5 & 4 & 3 & 2 & 1 \end{array} \\ \begin{array}{l} 2004 \\ 2003 \\ 2002 \\ 2001 \\ 2000 \end{array} & \begin{pmatrix} 17 & 10 & 12 & 9 & 2 \\ 8 & 16 & 19 & 7 & 1 \\ 6 & 18 & 9 & 5 & 4 \\ 6 & 9 & 9 & 5 & 14 \\ 5 & 13 & 9 & 15 & 5 \end{pmatrix} \end{array}$$

It is worth pointing out that in 2003 one student took and passed the AP Statistics exam at Greendale, and 32 took this exam with 23 scoring 3 or better at Hilltop. In 2004 fifteen students took this exam, with 9 receiving a score of three or better at Greendale while 35 took the exam at Hilltop with 27 scoring 3 or better. However no students took this exam at Railside from 2000 - 2004.

We also have the overall AP results for the students at these three schools for the last five years. Here are the numbers of students at each school scoring 1, 2, 3, 4, or 5 on the AP exams. Note the differences in the distributions of these scores, and also the number of students taking these exams.







II. Filling in the Missing Data for the Railside Study

> "*The Railside students were able to achieve at comparable levels after a year of algebra teaching, despite starting the course at significantly lower achievement levels. At the end of Year 2 we gave students a test of algebra and geometry, reflecting the content the students had been taught over the first two years of school. By the end of Year 2 Railside students were significantly outperforming the students in the traditional approach (t = -8.304, p < 0.001, n = 512).*" RAILSIDE STUDY, PAGE 8
>
> *By the end of year 4, 41% of Railside seniors were enrolled in advanced classes of pre-calculus and calculus, compared with 30% of seniors at Hilltop and 23% of seniors at Greendale. There were no gender differences in performance in any of the tests we gave students at any level, and young women were well represented in higher mathematics classes.*" RAILSIDE STUDY, PAGE 10

The full data available for these schools shows that this glowing description is not accurate.

It is difficult to identify the cohort of students Boaler was studying from the article. Indeed, the only point where this is indicated is in the title of her Table 6, where she indicates that in 2002-2003 the cohort she studies was in grade 11. *Thus, her cohort took ninth grade algebra in 2000-2001, and graduated in 2004.* Two of the schools, Greendale and Hilltop, were very unstable during this period. Significant changes were taking place in their mathematics curriculums and there was a great deal of faculty discontent at both schools. Some faculty left Greendale as the school reinstated its traditional math courses. At the same time a small number of previously retired faculty rejoined the mathematics staff.[11]

> "*Another curriculum aligned test is the California standards test, taken by students who had completed algebra. These scores show similar performance for all three schools with the Railside students scoring at*

[11] *Greendale had been one of the relatively few schools in California that taught the IMP curriculum exclusively until just before the beginning of Dr. Boaler's study. The first years of her study at Greendale coincided with the reintroduction - due to enormous parental pressure - of more traditional math courses.*

higher levels than the other two schools (see Table 5). Forty-nine percent of Railside students scored at or above the basic level, compared to 33% at Greendale and 41% at Hilltop.

'*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

RAILSIDE ARTICLE, PAGE 12.

This is a key point where the author is being disingenuous. For one thing, her Table 5 reports on *ninth grade Algebra I scores* in 2003. Thus, these are not the scores of the cohort she is following. A second problem is that the outcome reported in Table 5 appears to be an isolated case. Here are the results for 2002 for these three schools:

	Greendale	*Hilltop*	*Railside*
n	128	283	132
Advanced	0	2	1
Proficient	5	12	8
Basic	48	34	36
Below basic	34	38	39
Far below basic	13	13	17

where we see that a solid case can be made for the ninth grade Algebra I students at Greendale outperforming the students at Railside and performing at least comparably with the students at Hilltop. Similarly we have the results for 2004.

	Greendale	*Hilltop*	*Railside*
n	108	250	229
Advanced	1	0	0
Proficient	22	14	10
Basic	34	38	33
Below basic	39	42	48
Far below basic	5	6	8

Here, there can be no doubt that the overall results at Greendale are better than those at either of the other two schools with results at Railside being significantly the weakest of the three.

The author might protest that the students at Greendale and Hilltop have better demographics than the students at Railside, so one would expect them to do better. However, as is standard, if we assume that ability levels are distributed reasonably uniformly, and accepting Dr. Boaler's assertion that the Railside students rapidly recovered from any disadvantages in their previous mathematics educations, we should not see such variations.

But there is a further consideration here. In fact, the populations at the three schools in 2003 that took the ninth grade algebra test appear to be significantly different (and this difference extends to the cohort she was actually studying). More exactly we have the following table for the three years where the data are available which gives the approximate class percentile ranges of the students in the ninth grade classes at these respective schools taking the Algebra I STAR exams.[12]

School	2002	2003	2004
Greendale	31 − 73	28 − 73	37 − 72
Hilltop	32 − 86	36 − 78	35 − 83
Railside	59 − 92	60 − 97	46 − 96

This table is constructed under the assumptions that the students who did not take the STAR mathematics exams together with those students taking the general math exam comprised the lowest percentiles in their respective schools, while the ninth grade students taking the integrated 2 exam, integrated 3 exam, algebra II exam, or the geometry exam comprised the highest percentiles. We also assumed that the ability levels of the students taking the first integrated math exam matched those of the students taking the algebra exam. The data that go into this table are, for example for 2003, given by the following table giving the percentages of the respective ninth grade classes taking each test:

2003	*No test*	*Gen.Math*	*Int.I*	*Alg.I*	*Int.II*	*Geometry*	*Alg.II*
Greendale	17	10	0	46	0	20	7
Hilltop	15	17	7	42	3	15	1
Railside	34	25	0	38	0	3	0

This table also shows that about 27% of the students at Greendale and 19% of the students at Hilltop were taking more advanced courses in grade nine in 2003. This is consistent at these schools according to their STAR test results after about 2000. Here are some details. In 2001 - for the cohort in the Boaler study - about one percent of the ninth grade students at Railside took one of the Algebra I, Integrated I, Geometry, Integrated II, or Algebra II, or Integrated 3 tests. By contrast, at Greendale and Hilltop the numbers were

Greendale

Course	*Percent*	*Score*
Algebra I	31	31.1/65
Integrated 1	14	23.0/65
Geometry	26	44.9/65
Integrated 2	4	37.3/65
Algebra II	1	∗
Integrated 3	0	∗

[12] *Not all students who took the algebra course at Railside - supposedly almost all the ninth grade students - took the algebra exam. 59% of the 2002-2003 freshman class at Railside either took the general math exam or did not take any STAR math test.*

Hilltop

Course	*Percent*	*Score*
Algebra I	42	28.7/65
Integrated 1	3	21.2/65
Geometry	10	46.9/65
Integrated 2	4	28.3/65
Algebra II	1	*
Integrated 3	0	*

In 2000 about 10% of the ninth grade students at Railside took one of these six tests, but in no case were there enough to record the grades. The situation at Greendale and Hilltop was as follows:

Greendale

Course	*Percent*	*Score*
Algebra I	35	*
Integrated 1	5	*
Geometry	6	34.9/65
Integrated 2	4	30.5/65
Algebra II	2	*
Integrated 3	2	*

Hilltop

Course	*Percent*	*Score*
Algebra I	34	*
Integrated 1	22	*
Geometry	14	35.1/65
Integrated 2	*	*
Algebra II	*	*
Integrated 3	*	*

This data is relevant because the key argument Dr. Boaler makes to support her conclusions is that by the end of the 10th grade the students at Railside surpassed the achievements of the students at Hilltop and Greendale on her tests. But these students at Railside appear to be those that had continued with the second year of the Railside mathematics sequence. We reproduce her key Table 2 here.

Table 2
Assessment Results

TRADITIONAL

	Mean score	*Std Deviation*	*n*
Y1 Pre-test	22.23	8.857	311
Y1 Post-test	23.90	10.327	293
Y2 Post-test	18.34	10.610	313

RAILSIDE

Mean score	*Std Deviation*	*n*	t (level of) significance)
16.00	8.615	347	−9.141 ($p < 0.001$)
22.06	12.474	344	−2.040 ($p = 0.04$)
26.47	11.085	199	−8.309 ($p < 001$)

Note the 'n's' - virtually no change in the traditional group, but a drop of 43% in the Railside group. Thus, we can assume that the treatment group for the second post-test at Railside was very close in structure to the group of students that took the Algebra I test in 2003, the upper half of the class, while the students who took the second test at the other schools were essentially the students in the second and third quartiles at Greendale and Hilltop.

The Role of integrated mathematics in the curricula of the schools

On page six of the Railside article the following description of the programs at the three schools was presented.

> "*Classes at Railside were heterogeneous as the school had de-tracked classes in previous years; classes at the other two schools were not. We monitored three approaches in the study - traditional and IMP (as labeled by the two schools) and the Railside approach. However, as only one or two classes of students in Greendale and Hilltop chose the IMP curriculum each year there were insufficient numbers of students to include in our statistical analysis.*

Actually, the situations at Greendale and Hilltop were much more turbulent than this. Both schools, at the beginning of the Boaler study, had faculties committed to reform math instruction. The program at Greendale had been 100% IMP, and until 2003 - well into the Boaler study - Hillsdale had a significant size contingent of students in their IMP sequence. Just before Boaler's study began, parents at Greendale had forced the school to introduce choice between a traditional curriculum and IMP. In 2003 the Hilltop school district forced the school to drop IMP. In both cases, the faculties appeared to be very resentful. There are stories of students being subjected to pressure to rejoin the IMP classes at Greendale. Indeed, for a period of time a preliminary article that Dr. Boaler had written about IMP and traditional students at Greendale appeared on the Greendale web-site. Here is the abstract from that article:

> "ABSTRACT. *The following report gives initial results from the first year of a study of mathematics teaching and learning, conducted by researchers from Stanford University. The study is designed to track and compare students' progress through different mathematics curricular and teaching approaches. In the school discussed in this report, two approaches: a traditional algebra and geometry sequence, and an alternative problem-based integrated mathematics approach, are offered concurrently. This school is a particularly interesting place in which to study mathematics approaches, as the students were taught mathematics in different ways, but they come from the same school population and have had similar mathematics teaching experiences in the past. Results from the first year of this study indicate that there were no significant difference in tested achievement of students taught through the different curricular approaches. Results also show that students in the different programs are demonstrating important differences in attitude and interest towards mathematics. These differences are particularly important because they are likely to affect the students' subsequent learning of mathematics.*"

As regards Hilltop, to this day, if one reads the course descriptions on the Hillsdale web-site, the teacher's attempts to inject IMP style material into courses with traditional names is clear.

In 2003, Hilltop dropped IMP. The article in the local paper reporting on the reasons for dropping it pointed out that more than 90% of the students in the IMP classes scored below or far below basic, with similar scores the year before. Moreover the 2004 students did only slightly better. The article also pointed out that students in the standard math classes averaged more than 40 points higher than students in the IMP program.
The same article inplied that the faculty at Hilltop did not entirely accept this decision. In particular, the teachers were said to be critical of using standardized tests to evaluate students in the IMP course.

Parental pressure at Greendale in the late 1990's resulted in the school being forced to offer ninth grade students a choice between IMP and a more traditional curriculum. Virtually all the students elected the algebra sequence in the first years this choice was available. However, the mathematics faculty was very unhappy with this situation and, with Dr. Boaler observing a number of meetings with parents, they tried to convince many of the parents to place their children back in the IMP sequence. There was some justification for this in that in these years the ninth grade students were generally ill prepared to handle traditional curricula. The reason was that the students at Greendale had taken courses with curricula like Connected Math in middle school and MathLand in elementary school. On the other hand there was at least one letter to the editor of the local paper from a parent complaining that the school was distributing a great deal of information about IMP but saying very little about the reasons for taking a more traditional mathematics sequence.

The IMP program at Greendale is still in place, but a nominal number of students currently take it.

III. THE RELATION OF CURRICULUM TO OUTCOMES AT THE SCHOOLS

In terms of outcomes, Dr. Boaler points out that 41% of the students in the cohort she studied at Railside were taking "pre-calculus or calculus" in their final years at the school. This number was confirmed by Railside's principal in a phone conversation with one of us. The principal also stated that no Railside student had taken an AP Calculus exam during the last five years, and this was confirmed by the actual data.

The head of the math department at Railside responded to our request for clarification of this point as follows. "I would like to quickly respond to one question directly: AP Calculus at *** means the same that it means everywhere: that we follow the AP syllabus. Students may of course choose to take the AP test, but the fact that they do not is probably due to the timing of our block schedule."[13]

Moreover, we have the following course descriptions paraphrased from the Railside course catalog:

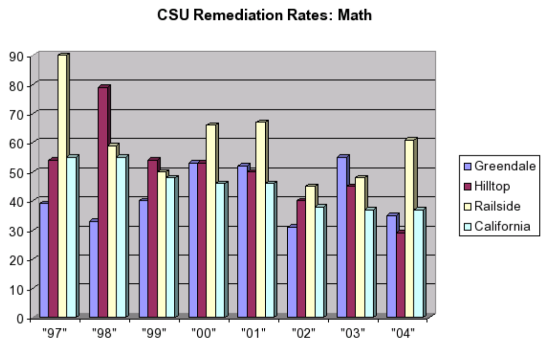
> PRE-CALCULUS:College preparatory mathematics course #5: explore concepts important for Calculus including Euclidian transformations, trigonometry, linear algebra, and complex numbers.
>
> CALCULUS, AP College preparatory mathematics course #6: Use the graphing calculator to learn math of infinity, including advanced graphing techniques, derivatives and integrals.

They appear to be descriptions of relatively low level courses. This is consistent with the remediation rates of the students from Railside in mathematics in the California State University System.

We do not regard the remediation rates for the students at Greendale as entirely relevant to our analysis of Boaler's study because only a small percentage of the students at Greendale who started with algebra I or IMP in grade 9 in 2000 were apt to be in the upper thirty three percent eligible for admission to the CSU system. The remediation rates at Hillside are somewhat more relevant since about 30% of the students there that started with Algebra I or IMP in grade nine would be

[13] *It appears that Dr. Boaler was aware of this problem. The head of the math department at Railside had been awarded a grant to do in-service teacher and curriculum development at Railside working with Prof. Boaler and others. The focus was on preparing students for college. In particular the proposal talked about improving the AP outcomes.*

eligible for admission to the CSU system. In any case, here is a table showing all three schools' remediation rates compared to the state averages. This table shows that none of the three schools appear to be doing a particularly good job of preparing that part of their students who elect to attend the CSU system, though the outcomes for Railside are clearly the worst of the three.



(Note in particular the results for 2004, which are the results for the cohort of students in Dr. Boaler's study.)

IV. The validity of the post-tests used in the Railside study

> "Student achievement data. *In addition to monitoring the students experiences of the mathematics teaching and learning we assessed their understanding in a range of different ways. At the beginning and end of Years 1 and 2 and at the end of Year 3 we administered tests that were carefully written by the research team and considered by the teachers in each approach to make sure they fairly assessed each approach. The test at the beginning of high school was a test of middle school mathematics, as that was the mathematics students would be expected to know at that time. At the end of Year 1 we administered an algebra test. The test was designed to assess only algebraic topics that the students had encountered in common across the different approaches, and we used an equal proportion of question-types from each of the three teaching approaches. At the beginning of Year 2 we administered the same test, giving us a record of the achievement of all students starting Year 2 classes. At the end of Year 2 we wrote and administered a test of algebra and geometry, again focusing only upon content all students had met, using question types from each approach that teachers from each approach reviewed, and repeated this process at the end of Year 3.*" From Page 7 of the Railside study

It is non-trivial to assess test validity, and it is clear from the descriptions of the method of validation given by the author of the Railside study in the last sentence of the paragraph quoted above that no serious attempt was made to validate these tests. See Appendix 3 for a short description of the issues involved in test validation.

Without a complete discussion of all these validation issues, it is very difficult, in fact really impossible, to evaluate research claims based on a test. There are a number of aspects of the tests that we cannot evaluate since the tests were only given once and we do not have access to the student level results on each item in the tests. However, we can discuss the fourth issue, *the extent to which a test measures that which it is supposed to measure.* Since the study does not really use the initial pre-test or the third post-test we will not study these tests. But we will determine the levels of the first and second post-tests, measuring them against the California Mathematics Content Standards. All three authors of this note helped write the current California mathematics standards. One is a professional mathematics educator, one a professional mathematician, and one a professional statistician and they independently evaluated all the parts of each question on the first and second post-tests. They assigned to each part the California mathematics standards that were relevant to that part, and then the grade levels of these assignments were averaged to measure the grade levels of each question.

It was found that the first post-test is, on average, written at the 5.7^{th} grade level while the second post-test is, on average, written at the 6.6^{th} grade level. Hence these tests are approximately 3 years below grade level.

This is somewhat subtle. On superficial examination, it appears that many of these questions are at grade level. However, when one looks at the special assumptions and the particular choices for the given data in each problem, it turns out that very elementary methods can be used to resolve the questions, at least to the point asked for by the examiners. We use a term coined by the mathematician J. Dancis, and popularized by the Washington Post education writer and columnist, Jay Mathews, for problems of these kinds, *pretend algebra* and *pretend geometry*.

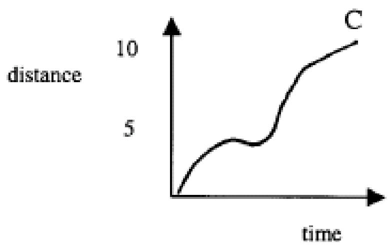
Moreover there are a number of mathematical errors in the problems presented in all four tests, but particularly in the first post-test. One problem, #9 on the first post-test, is particularly problematic though it is not incorrect. A discussion of the difficulties with this problem comprises the entire fourth appendix to this report. The other difficulties with specific parts of problems in the first and second post-tests are discussed in the detailed appendices that follow.

It turns out that there are even more difficulties with the pre-test and the third post-test, though we don't evaluate either of these tests for this study. One particularly egregious error on the pre-test which has an amusing connection with the first post-test follows. The question from the pre-test (which is identical to problem (7) on the first post-test) is

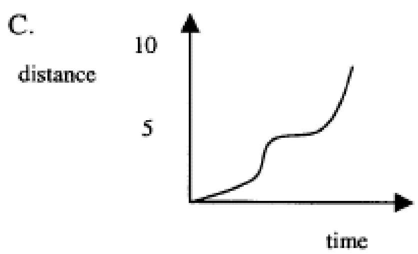
9. Ana is training for a race. She runs 10 miles every day. Below are descriptions of 3 of Ana's training runs. Choose the graph that best represents each of Ana's runs.

Ana ran the first 5 miles very quickly and then slowed down for the last 5 miles. Graph ________	Ana ran a route with 2 hills. She ran down the hills quickly, but up them more slowly. Graph ________	Ana ran at a steady pace for all 10 miles. Graph ________

The answer for (b) in the pretest was given by the following graph



which actually represents turning around at the bottom of the hill and going backwards. It appears that this error was pointed out, because in the year 1 post-test, the graph is replaced by this one



but it is still peculiar. Here Ana descends the first hill at an extremely high rate of speed then goes up the second hill - which must be almost vertical - at a crawl, and seems to be constantly accelerating as she goes down the second hill.

One problem that is incorrect as stated is problem #4 on the third post-test. This problem asks that students find the lengths of the sides and the missing angles for a triangle with a base of length 10 units, area 62 square units, and one angle being 40 degrees.

> *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 wording tends to imply there is only one such triangle. Indeed, this would have been the case if the area had been somewhat larger, say larger than 68.687 square units. However, in all but one case when the area is less than 68.6869 square units there are two non-congruent triangles. To solve the problem we have to first determine the number of non-congruent triangles with one side of length 10, one 40 degree angle, and total area 62 square units. There are two possible cases. The first case is where the given side contains the vertex of the 40 degree angle, and the second case is where the 40 degree angle is opposite this edge. The first case always exists. If we place the 40 degree vertex at the origin and the given side along the positive x-axis, then the coordinates of the missing vertex are $(12.4cot(40), 12.4)$ which is approximately (14.778, 12.4). Consequently, we have the side lengths given as $12.4/sin(40)$ and $\sqrt{(12.4/sin(40))^2 - 240cot(40) + 100}$, while the measure of the angle opposite the 10 unit side is given in degrees as $50 - arctan(cot(40) - 10/12.4)$.

The second case is more delicate. We have to determine that the second case exists. First note that the opposite vertex must lie on the circle of radius $5/sin(40)$ where the given side is a chord on this circle. The triangle with largest area having base of length 10 and opposite angle 40 degrees is the triangle where the opposite vertex lies on the intersection of this circle and the perpendicular bisector of the chord. The height of this point above the chord is $5cot(20)$ which is approximately 13.737. Since this number is greater than 12.4, there are two points on the circle having height 12.4 above the chord. They give congruent triangles, each having the desired area, so we only need to determine one of them. Set the center of the circle at $(0, 5cot(40))$ so the chord is the segement on the x-axis between $(-5, 0)$ and $(5, 0)$. Then the equation of the circle is $x^2 + (y - 5cot(40))^2 = 25/sin^2(40)$ and the coordinates of the opposite vertex are

$$(\sqrt{25/sin^2(40) - (12.4 - 5cot(40))^2}, 12.4).$$

From this the side lengths and the trig functions associated to the remaining angles are directly determined.

Extreme care has to be taken in grading this problem since any student attempting to discuss the triangle where the 40 degree angle is opposite the given side has to realize that the construction is not obtained via a general argument, and has to justify the existence of the triangle. Of course, the triangle where the 40 degree angle is not opposite the 10 unit side is very direct to construct and *is* a consequence of a general argument.

The validity of the measurements provided by the Boaler tests

Since the levels of these two exams were so far below what is expected in California and high achieving countries in algebra and geometry, it was originally thought that we should measure their content validity against the K - 7 California mathematics standards. However, since these tests claim to measure what students should know in the first two years of high school we measured them against the Algebra I standards and the Geometry standards.

THE ALGEBRA I STANDARDS **NOT** COVERED BY THE TESTS. There are 25 distinct standards for Algebra I, and most of them are of major importance. The following list is those standards that **are not covered in any way**, except as noted for 10.0 and 13.0, by any of the questions in these tests.

> 1.0 Students identify and use the arithmetic properties of subsets of integers, rational, irrational and real numbers. This includes closure properties for the four basic arithmetic operations where applicable.
> 2.0 Students understand and use such operations as taking the opposite, reciprocal, raising to a power, and taking a root. This includes the understanding and use of the rules of exponents.
> 3.0 Students solve equations and inequalities involving absolute values.
> 6.0 Students graph a linear equation, and compute the x- and y- intercept (e.g., graph $2x + 6y = 4$). They are also able to sketch the region defined by linear inequality (e.g., sketch the region defined by $2x + 6y < 4$).
> 7.0 Students verify that a point lies on a line given an equation of the line. Students are able to derive linear equations using the point-slope formula.
> 8.0 Students understand the concepts of parallel and perpendicular lines and how their slopes are related. Students are able to find the equation of a line perpendicular to a given line that passes through a given point.
> 10.0 Students add, subtract, multiply and divide monomials and polynomials. Students solve multistep problems, including word problems, using these techniques. *The nearest any problem in the first and second post-tests comes here is #9 of post-test 2 which supposes that a rectangular prism with a square base having side length a, height h and is one-third filled with water having a a total volume of W cubic units. The question then asks for the formula in terms of a, W for h, that is, $3W/a^2$.*
> 11.0 Students apply basic factoring techniques to second and simple third degree polynomials. These techniques include finding a common factor to all of the terms in a polynomial and recognizing the difference of two squares, and recognizing perfect squares of binomials.
> 12.0 Students simplify fractions with polynomials in the numerator and denominator by factoring both and reducing to lowest terms.
> 13.0 Students add, subtract, multiply, and divide rational expressions and functions. Students solve both computationally and conceptually challenging problems using these techniques. *The closest any problem in the first and second post-tests comes to this standard is #4 of the second post-test which is extensively discussed in Appendix 2, and doesn't really come close to the level expected here.*
> 14.0 Students solve a quadratic equation by factoring or completing the square.
> 15.0 Students apply algebraic techniques to rate problems, work problems, and percent mixture problems.
> 16.0 Students understand the concepts of a relation and a function, determine whether a given relation defines a function, and give pertinent information about given relations and functions.
> 17.0 Students determine the domain of independent variables, and range of dependent variables defined by a graph, a set of ordered pairs, or symbolic expression.
> 18.0 Students determine whether a relation defined by a graph, a set of ordered pairs, or symbolic expression is a function and justify the conclusion.
> 19.0 Students know the quadratic formula and are familiar with its proof by completing the square.
> 20.0 Students use the quadratic formula to find the roots of a second degree polynomial and to solve quadratic equations.
> 21.0 Students graph quadratic functions and know that their roots are the x-intercepts.
> 22.0 Students use the quadratic formula and/or factoring techniques to determine whether the graph of a quadratic function will intersect the x-axis in zero, one, or two points.
> 23.0 Students apply quadratic equations to physical problems such as the motion of an object under the force of gravity.
>
>> 24.1 Students explain the difference between inductive and deductive reasoning and identify and provide examples of each.
>> 24.2 Students identify the hypothesis and conclusion in logical deduction.
>> 24.3 Students use counterexamples to show that an assertion is false and recognize that a single counterexample is sufficient to refute an assertion.
>> 25.2 Students judge the validity of an argument based on whether the properties of the real number system and order of operations have been applied correctly at each step.
>> 25.3 Given a specific algebraic statement involving linear, quadratic or absolute value expressions, equations or inequalities, students determine if the statement is true sometimes, always, or never.

THE GEOMETRY STANDARDS COVERED BY THE TESTS. Next we look at the geometry standards that are measured by her tests. As was the case with algebra the California geometry standards are limited in number - there are only 22 - but they are very focused, and virtually all of them are essential for students to know. We now list the extremely few standards that are covered in even the most superficial way in the two Boaler exams:

> 5.0 Students prove that triangles are congruent or similar, and they are able to use the concept of corresponding parts of congruent triangles.
> 6.0 Students know and are able to use the triangle inequality theorem.
> 9.0 Students compute the volumes and surface areas of prisms, pyramids, cylinders, cones, and spheres; and students commit to memory the formulas for prisms, pyramids, and cylinders.
> 10.0 Students compute areas of polygons, including rectangles, scalene triangles, equilateral triangles, rhombi, parallelograms, and trapezoids.
> 13.0 Students prove relationships between angles in polygons by using properties of complementary, supplementary, vertical, and exterior angles.
> 15.0 Students use the Pythagorean theorem to determine distance and find missing lengths of sides of right triangles. *This is represented in an extremely weak way by problem #5 of the second post-test, which can be answered without the use of the Pythagorean theorem if students are aware of the* $(3, 4, 5)$ *right triangle.*

These are all standards that have been, to a large degree, covered in earlier courses. All the basic material that is new to the geometry course, such as ruler and compass constructions, beginning trigonometry, simple proofs, or the content of the last two standards,

> 21.0 Students prove and solve problems regarding relationships among chords, secants, tangents, inscribed angles, and inscribed and circumscribed polygons of circles.
> 22.0 Students know the effect of rigid motions on figures in the coordinate plane and space, including rotations, translations, and reflections.

is simply not represented in these tests.

THE ALGEBRA II STANDARDS COVERED BY THE BOALER TESTS. There are 24 standards in the California Algebra II standards. Here are the standards that these tests cover in any way.

> 2.0 Students solve systems of linear equations and inequalities (in two or three variables) by substitution, with graphs, or with matrices. *Of course it must be noted here that the only methods expected in these tests are substitution and possibly graphing.*
> 22.0 Students find the general term and the sums of arithmetic series and of both finite and infinite geometric series. *This is represented in and extremely weak way by #2 on the pre-test, which is repeated as #7 on the second post-test.*

Since these tests are assumed to generally cover just the material in the first algebra course and the geometry course, we do not regard the lack of any substantive questions about the content of the second algebra course as germane. The remarks above are included only for completeness.

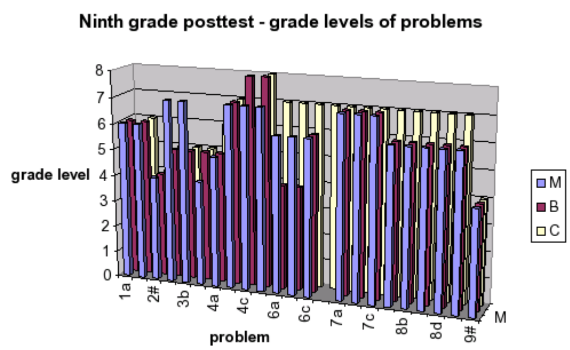
V. CONCLUSION

. We conclude that these tests *cannot measure that which they are supposed to measure*, and consequently, whether or not they are **consistent measurement instruments**, they are of minimal value in determining student achievement in mathematics at the ninth and tenth grade levels.

As a consequence of our analysis we can say that it is unrealistic to try to draw any valid conclusions about student mathematics achievement at the high school level based on the Boaler tests. It is consistent with this conclusion that all the other measures of student achievement we studied gave a dramatically different picture than the Boaler tests did.

VI. APPENDIX 1: THE EVALUATION OF THE NINTH GRADE POST-TEST

The three authors independently reviewed and evaluated the questions on this test against the California Mathematics Standards. All three were involved in writing these standards, and consequently were very familiar with them and the levels of competency expected for each standard. Their individual comments are presented in detail below, identified by the first letter of the authors last name. Here is a summary chart showing the grade levels assigned to each part of each problem. (8 means Algebra I.)



As can be seen, there is substantial agreement among the authors as to the grade levels of the 24 parts of this exam. Since problem #6d was not well posed, two of the three authors did not rate it.

The average grade level of the 9 problems	5.7
Standard deviation for the 9 problems	1.19
Average grade level of all 23 parts	6.03
Standard deviation for the 23 parts	1.13

We now give the detailed evaluation of the problems on this test. Note that a number of the problems are ill-posed or otherwise mathematically inappropriate. This will be noted in the comments.

Problem (1) Simplify each expression

a. $5 - 3 - 4 \times 2 =$
b. $15 - 2(4 + 6) =$

1(a) (M,B) 6AF1.3[14] [Order of operations], but the easiest part of it only.

(C) 4AF1.2. The problem says simplify, but the test's authors probably want it solved. It is rather difficult to peg this item to a standard. Order of operations doesn't show up until 6AF1.3, but by then the student is way beyond what is asked in this problem. One could argue for 3NS2.8 and/or 3AF1.2 for the operations involved. Maybe the fair placement is 4AF1.2 even though parentheses are not involved.

1(b) (M, B, C) 4AF1.3. The illustrative examples are virtually the same.

Problem (2) Using three of these operations $(+, -, \times, \div)$ and some numbers, write an expression that equals 12.

(M) This question is not very well posed - what does the phrase "using three of these operations" mean? I conceive of $(0-0)\times 0+12$ as a possibility, also a nonsense phrase "I am using three symbols but they don't mean anything $+, -, \times$, now I am writing what I want this expression to be, 12."

In any case, this is covered by 4AF1.3.

(B) Easy. This problem could be regarded as third grade level, but if students have never been asked a question of this type before they could have difficulties.

(C) 4AF1.3. This is a fake "real-world" problem - use a specific number of operations and any numbers you want. In any case, the student is very able to do this with 4AF1.3

Problem(3) Here is a rectangle. The sides are $2x+4$ and 6 units.

a. Find the perimeter of the rectangle. Simplify your answer if possible.

b. Find the area of the rectangle. Simplify your answer if possible.

c. Draw and label a rectangle with the same area that you found in part b, but with a different length and width. Show your work.

a. (M) Aside from the fact that one side is given by an expression in a variable $2x+4$, this is evaluated as follows. First the meaning of $2x+4 = y$ is fully covered by 4AF1.5. Then 4MG1.4 covers 3(a), except for the requirement that the resulting value $2(2x+4)+2\times 6$ be "simplified." Of course, simplification is ambiguous. The expression above is pretty simple. Rewrite it as $4x+20$ which is what I suppose is meant. But almost entirely, the simplification requirement is covered by the standards in 7AF.

(B) Grade 7 is more than generous. 5AF1.3 is not keyed but it is "know and use the distributive property in equations and expressions with variables," and perimeter and area are 4MG1.0 unless "understand" doesn't mean "understand and compute" in which case it moves up to Grade 5. And there is still Grade 6 to play with so I'd put it there because of the level of sophistication.

(C) 5AF1.3. The first part of the proble, to write the expression, is 4AF1.1 but writing expressions also shows at 5AF1.2. It is the "simplify" part of the problem that is tricky. The standards don't get to formal simplification of expressions with variables until Algebra I, but formal simplification is clearly expected in 5AF1.3. Note that the problem is not evaluation because the value of x is never stipulated.

b. (M, B) The formula for the area is covered by 4MG1.4, and the simplification requirement is covered by the 7AF1.3.

(C) 5AF1.3. Same argument

c. (M, B) 4MG1.3, 4MG1.4 cover 3(c), except for the explicit method of determining the side lengths. This is ambiguous, however what we think they had in mind was that the area equals $6(2x+4) = 3\times 2\times 2(x+2)$, so this can be written as a product of an integer side length and a side length involving x in a number of ways, for example 3 and $4(x+2)$ or 2 and $6(x+2)$ or 4 and $3(x+2)$ or even 1 and $12(x+2)$. If this is what is expected then the correct standards are 5NS1.4, and 4NS4.1, 4NS4.2, together with the weakest form of 8-2.0 (which isn't really necessary here).

(C) 5AF1.3. Same argument

Problem (4) Solve the folowing equations, check your answers and show your work:

a. $5x-3=101$ *b.* $3x-1=2x+5$ *c.* $x^2=4$

(a) (M,B) $5x-3=101$. The standards for the basic manipulation of such equations are 4AF2.0, 4AF2.1 and 4AF2.2. But the explicit requirement to solve such an equation is 5AF1.2.

(C) 6AF1.1. 5AF1.2 calls for evaluation by substitution but this is solving and probably would not be done by guess and check.

(b) (M) equations of this type, $3x-1=2x+5$, are covered by 7AF4.1, but at a higher level, since in this seventh grade standard the coefficients can be rational numbers.

(B) The phrase "two-step" ought to make it clear but it doesn't because "n-step" is ill defined. Isn't $2x+3=7$ a two-stepper, not just a one step problem as indicated in the example for 6AF1.1?

(C) 7AF4.1. It was difficult for me to classify this one

[14] *The first number in these standards abbreviations is the grade level. Here, 8 means the Algebra I standards, 9 means the Geometry standards, and 10 means the Algebra 2 standards. The California Mathematics Standards have 4 mains strands, Number Sense, abbreviated NS, Algebra and Functions, abbreviated AF, Measurement and Geometry, abbreviated MG, and Statistics, Data Analysis and Probability, abbreviated SD. Thus 6AF1.3 means grade 6, Algebra and Functions, standard 1.3, and 9-7 means Geometry, Standard 7.*

(c) (M) $x^2 = 4$. In 7NS2.4 the illustrative example is a harder case of this type of equation.
(B) Does one get negative square roots in 7NS2.4? How do you sketch the square? This is an extremely simple problem for the standard 8-14.0; i.e., it is an instance of "pretend algebra."
(C) 5NS1.3. The expression $x^2 = 4$ is not solving a second degree equation, it is knowing what a variable is and what squared is. I see some 5NS1.3 and 4AF1.2 here, and also knowing that $2 \times 2 = 4$, which students should know in first grade (1NS2.1).

Problem (5) Use 2 *different methods to find where the following two lines intersect. Show your work.*

$$\begin{aligned} y &= x + 6 \\ y &= -2x + 3 \end{aligned}$$

(M) I do not know what is meant by "two different methods." What determines when two ways of solving these two equations are "different?"
(B) Maybe the methods are meant to be subtraction, substitution, or graphing. Guess and check should not be acceptable for either of them. To be sure we need to see the scoring rubric though?
(M) Leaving aside the issue that the question is not well posed, the equations themselves are sufficiently elementary that they can be solved by guess and check, or graphing. In part, this question is covered by 5AF1.1 and 5AF1.5. It is also covered by 7AF4.1, and 4AF2.0, 4AF2.1, 4AF2.2. [This is a "pretend algebra" problem. It looks like it should represent the true standard 8-9.0, but it does not really rise to the level expected there.]
(B) 8-9.0 is where it has to be put, but with a comment that it does not rise to the expected level.
(C) 8-9.0, but at a very low level. The two ways they are looking for are the fuzzy way (graphic guessing) and tabled guess and check (you have to know that the lines intersect when the same values satisfy both equations). Reluctantly granting Algebra I on this one.

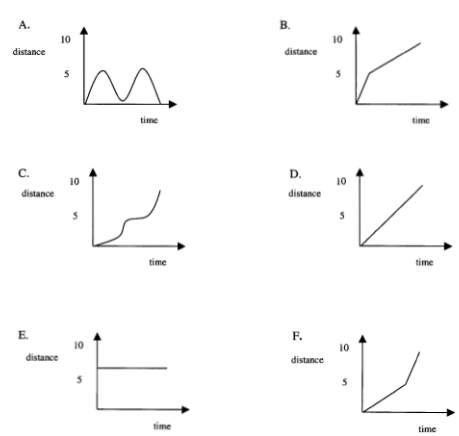
Problem (6) SuperSlide waterslide park charges $10 for the first 3 *rides. Each additional ride costs $2.*

a. *How much would 34 rides cost? Show your work.*
b. *You spent $18 on rides down the waterslide. How many rides did you take?*
c. *What rule tells you the cost of n rides?*
d. *WonderWater park charges $15 for* 3 *rides and $2 for each additional ride. Which park is a better value? Explain why.*

a. (M) 6AF1.2 in a weak form (only one variable) is appropriate for 6(a).
(C) 7AF1.1 and 7AF1.2.
b. (M) 6AF1.1 and 6AF1.2 are the relevant standards for 6(b),
(C) 7AF1.1 and 7AF1.2.
c. (M) 6AF1.2 is again the relevant standard for 6(c).
(C) 7AF1.2.
(B) I have a greater objection to 6(a) and 6(b) in that they can be answered with only number sense at the fourth or even third grade level with no algebra at all. If the order were reversed with an intro as for #8 and the old (a) and (b) becoming "Using (a), conclude that..." But it's not. I'd also prefer "function" to "rule" and with the restriction that n be greater than or equal to 3.
d. (B M) Part (d) is not well-posed. Since there is no "correct" answer for this part we will not consider it.
(C) 7AF1.1, 7AF1.2, 7AF1.5. This is a bad question that we would never allow. One could call it a trick question because there is the implication that one is always better than the other. Note that I pick grade 7 standards because 6AF1.2 calls for evaluation rather than solution.

7. Ana is training for a race. She runs 10 miles every day. Below are descriptions of 3 training runs. Choose the graph that best represents each of Ana's runs.

Ana ran the first 5 miles very quickly and then slowed down for the last 5 miles. Graph ________	Ana ran a route with 2 hills. She ran up the hills slowly, but down them more quickly. Graph ________	Ana ran at a steady pace for all 10 miles. Graph ________



(M) This multiple choice question is appropriate for 5AF1.1. It might also fit under 5SD1.4, but perhaps a better standard here would be 7AF1.5 (*Represent quantitative relationships graphically and interpret the meaning of a specific part of a graph in the situation represented by the graph*). Unfortunately, part (b), as was noted earlier, is somewhat stange, and was badly done in the pre-test.

(B) Seventh grade level.
(C) 7AF1.5 for each of the three parts

Problem (8) **Fishy Expressions** *The Osea Aquarium keeps careful track of the amount of food it feeds its animals. Shown here is a table of variables, or a key, the the aquarium workers use to help them with their records.*

D	*total number of dolphins*
A	*total number of adult sea otters*
Y	*total number of young sea otters*
P	*pounds of fish eaten by one dolphin each day*
Q	*pounds of squid eaten by one dolphin each day*
R	*pounds of fish eaten by one sea otter (both adults and young) (Note: A squid is not a fish)*

Use variables to write an expression that represents each of the following:

a. The total number of otters.
b. The total amound of fish needed each day to feed all of the dolphins and all of the sea otters.
c. There are 10 more adult sea otters than young sea otters.
Write a phrase that explains what each of the following expressions represents.

d. $Q \times D$
e. $A + Y = 4D$

(B,M) All five parts are appropriately covered by the 6AF1.2. They are also covered by 7AF1.1.
(C) Parts (a) through (c) and (e), 7AF1.1 at a very low level

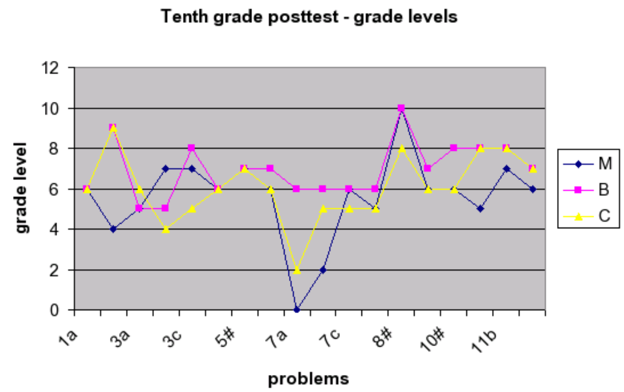
d. 7AF1.1 at a very low level, note that this is reverse of what the standard says but sometimes one does reverse things on a test

Problem (9) Here are two circular disks. A positive integer is written on each side of each disk. All four of the numbers are different. By tossing the disks in the air and adding together the numbers that land face up, you can get totals of 8, 9, 11, or 12.

a. If possible, find the numbers on the other side of each disk.
b. Find a pair of disks that would give you totals of four consecutive integers (for example 2,3,4,5). What numbers are on each side of each disk? Explain how you figured it out.
(M) Supposedly a mathematical reasoning problem. However, the only method students at this level have for solving it is informed guessing. I regard this as one of the weakest forms of math reasoning, and here, since all the numbers are small, so the possibilities for the answer are quite limited in number, the level of reasoning needed to answer the questions as stated is minimal. 4SD2.1, 4MR1.1, 4MR2.3.
(B) 4MR2.6.
(C) 4SD2.1, 4MR2.6.

VII. Appendix 2: The Evaluation of the Second Post-test

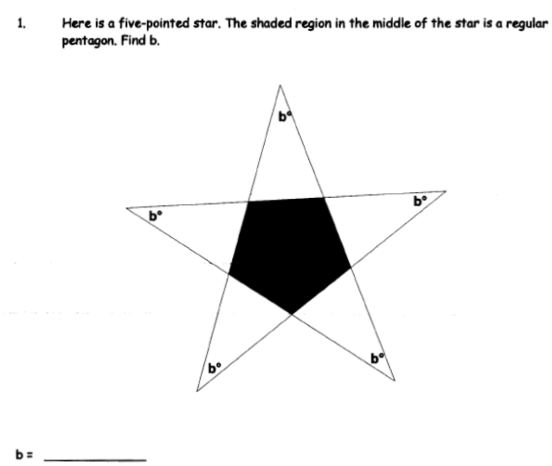
Here is a summary chart showing the grade levels assigned to each part of each problem. (8 means Algebra I, 9 means Geometry, and 10 means Algebra II.)



As can be seen, there is substantial agreement among the authors as to the grade levels of the 24 parts of this exam.

The average grade level of the 12 problems	6.66
Standard deviation for the 12 problems	1.14
Average grade level of all parts	6.20
Standard deviation for the parts	1.78

We now give the detailed evaluation of the problems on this test. Note that a number of the problems are ill-posed or otherwise mathematically inappropriate.
Problem (1).



(M) 6MG2.2. This is pretty close to what is expected in 6MG2.2: Use the properties of complementary and supplementary angles and the sum of the angles of a triangle to solve problems involving an unknown angle.
(B) 6MG2.2, but too difficult for an assessment even at 9-12.0 without similar problems having been discussed in class or done as homework.

(C) 6MG2.2. This problem can be solved with 6MG2.2 but it is quite difficult at that level. The question is very simple at 9-12.0, and there is really no standard in between

Problem (2) Nina tries to make a triangle with three sticks that are 10, 5, and 17 units long. Akim tries to make a triangle with three sticks that are 10, 8, and 13.5 units long. Only one of them is successful. Which one of them is able to make a triangle? Explain your reasoning. You may use diagrams.

(M) Fourth grade level, but does not exactly match any fourth grade standard. In Appendix E of the 2000 framework one finds the following problem in the 4MG section:

3.7 Assume that the sum of the length of any two sides of a triangle is greater than the length of the third side. If the lengths of the sides of a triangle are required to be whole numbers, how many such triangles are there with a perimeter of 14? List all of them.

Thus, one can assume that a problem which rests on knowing the fact that the sum of the lengths of any two sides of a triangle is greater than the length of the remaining side is at fourth grade level.
(B) 9-6.0, but that is really too high because of the formality of that standard as a theorem, not "Explain your reasoning." Without a scoring rubric, it can't be known.
(C) 9-6.0 - even though students will probably solve it by 5MG2.1, the geometry standard itself is explicit and very easy but ths concept just doesn't appear earlier.

Problem (3) a. Solve for a if $a + 5 = 3.5$.

b. Solve for b if $b + 2$ is equal to $2b$.
c. Solve for c if $c^2 + 1 = 10$.

(M) This is virtually identical to (4) on the first post-test. Thus it is possible to repeat the discussion in our evaluation of that question.
a.

(M) 5AF1.2.
(B) 5AF1.2 because the answer is negative.
(C) 6AF1.1.

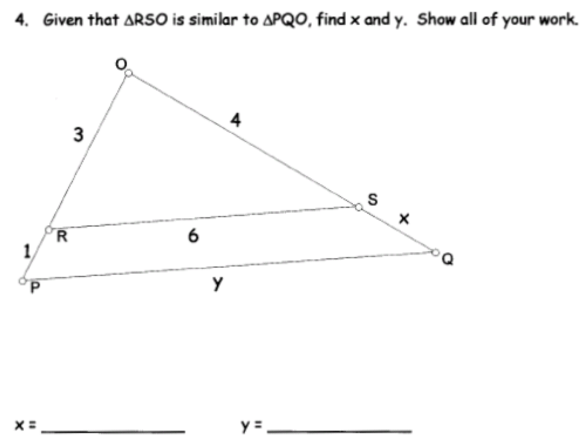
b.

(M) 7A4.1.
(B) 5AF1.2, but it's really a guess and check at about Grade 3.
(C) 4AF2.1.

c.

(M) 7NS2.4.
(B) 8-14.0, but "pretend algebra" because it's so simple.
(C) 5NS1.3.

Problem (4)



(M) 6NS1.3. First comment: A $3, 4, 6$ triangle does not look too much like the picture illustrated - it is considerably flatter with the angle between the sides of lengths three and four larger than illustrated. Indeed, if the first two vertices are (0,0) and (6,0), then the third vertex is $(29/12, (7 \times 65)^{.5}/12)$ which is approximately (2.416666, 1.777561657) But this is irrelevant to the problem. In the problem what is needed is only to understand that any two corresponding sides of similar triangles form a proportion. Thus $4/3 = (4+x)/4 = y/6$. Then one cross multiplies to solve for x and y. The relevant standard is 6NS1.3:

Use proportions to solve problems (e.g., determine the value of N if $4/7 = N/21$, find the length of a side of a polygon similar to a known polygon). Use cross-multiplication as a method for solving such problems, understanding it as the multiplication of both sides of an equation by a multiplicative inverse.

(B) 6NS1.3.
(C) 6NS1.3.

Problem (5) ΔPQR is a right triangle with hypotenuse PQ measuring 10 units. Find one pair of possible values for the lengths of PR and RQ.

$PR =$ ____________________
$RQ =$ ____________________

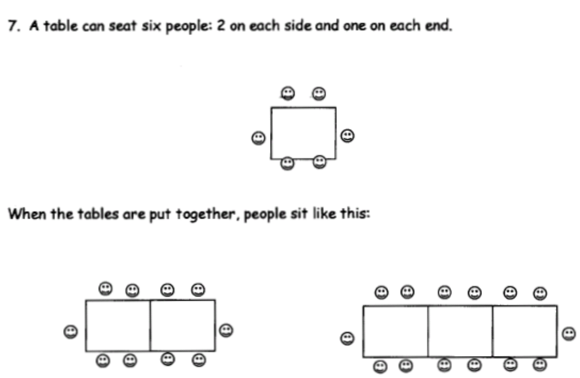
(M) 7MG3.3, (but see comments below). This problem, in full generality, requires knowing the inverse of the Pythagorean theorem (7MG3.3 - Know and understand the Pythagorean theorem and its converse and use it to find the length of the missing side of a right triangle and the lengths of other line segments and, in some situations, empirically verify the Pythagorean theorem by direct measurement). Thus X, Y and 10 are the sides of a right triangle if and only if $X^2+Y^2 = 100$. Hence, setting X equal to any real or rational number less than 10 and greater than 0 one sees that Y is equal to the positive square root of

$(100 - X^2)$. This requires some low level algebra, but is likely NOT to be what the test authors had in mind. Most likely they are assuming that the students know about the 3, 4, 5 right triangle. Hence by similarity the 6, 8, 10 triangle is also a right triangle. In this case, the problem does not really have a grade level. It could easily be solved by a third grade student, and is very likely to be solvable by a sixth grade student. This is thus an example of "pretend geometry."
(B) 7MG3.3, but I don't like the open option. It's a better use of the theorem if one leg is also specified.
(C) 7MG3.3 - if nothing else, teach your students the 3-4-5 Pythagorean triple

Problem (6) Create a trapezoid that has an area of 48 square units. Makes sure that you include labels of all the side lengths that you used to calculate your area.

(M) 6MG3.1. The first issue is the very serious one that there are at least two different meaning for "trapezoid" in common usage. The first is that a trapezoid is a quadrilateral with exactly one pair of parallel sides, and the second is that a trapezoid is a quadrilateral with at least one pair of parallel sides.[15]
Depending on the meaning one chooses for trapezoid one way of solving this problem is to assume the figure is a rectangle. Here the relevant standard is 5MG1.1 or lower.
In the remaining cases, this is still the relevant standard, though students will use the formulas with variables in them for the area of triangles and rectangles and solve for the variables. In this case the relevant standard is 6MG3.1
(B) 7MG2.1, but I don't like the open option. It's too easy - make a rectangle and a right triangle to make up the rest of the area.
(C) 6MG3.1.

Problem (7)



a) How many people can be seated at four tables?
b) How many people cna be seated at 50 tables?

[15] It is amusing to note that according to the Oxford English Dictionary, the correct usage is that a trapezoid is a quadrilateral with no parallel sides.

c) Write a rule for the number of people that can be seated at n tables.
d) How many tables would you need to seat 254 people?

(M) This problem is the repeat of a problem from the Algebra pre-test. It is purely an exercise in reasoning.
(a) (M) KNS1.2. In this case, the response is to draw 4 tables and directly count that 18 people can be seated.
(B) 6AF1.0, specifically 1.1 and 1.2.
(C) 2SD2.1.
(b) (M) 2NS1.1. Some students will draw 50 tables and count 202.
(B) 6AF1.0, specifically 6AF1.1 and 6AF1.2.
(C) 5AF2.2.
(c) (M) 6AF3.0. This requires the construction of the formula 4n + 2 for the number of seats at n tables arranged in a single row. I think that the fifth grade standard 5MG1.1 is what is needed, but others might disagree. So I will set this at 6AF3.0 - Students investigate geometric patterns and describe them algebraically - though it should be understood that problem (7) is at the low end of the expectations for this standard. It could also be regarded as an example of 6AF1.2, though it is again at the low end of the expectations for this standard.

(B) 6AF1.0, specifically 1.1 and 1.2.
(C) 5AF2.2.
(d) (M) 5AF1.2. Presuming that (c) has been done correctly, then this is 5AF1.2 - Use a letter to represent an unknown number; write and evaluate simple algebraic expressions in one variable by substitution. Illustrative example: $3x + 2 = 14$. What is x?
(B) 6AF1.0, specifically 1.1 and 1.2.
(C) 5AF1.2.

Problem (8) Daveeda has three bags of M& M's. She knows that:

a. Bags A and C together weigh 2.6 pounds.
b. Bags B and C together weigh 2.2 pounds.
c. Bags C and B together weigh 1.2 pounds.

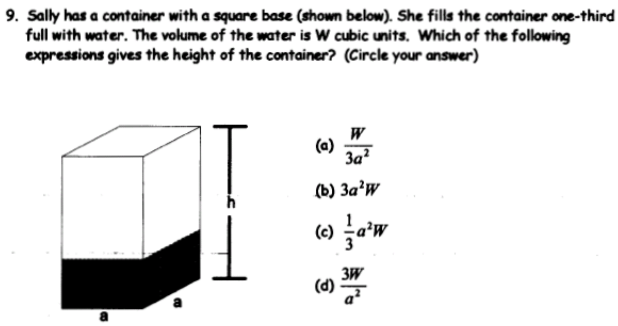
Help Daveeda figure out how much each bag weighs. Show your work.

Bag A =______________
Bag B = ______________
Bag C = ______________

(M) 10-2.0. This is a problem where I would be very interested in knowing the results for the three different schools. It is three linear equations in three unknowns with (easy) decimal coefficients. However, it is easily reduced to two linear equations in two unknowns (8-9.0), but the relevant standard is 10-2.0.
(B) 10-2.0, but ridiculously short of the standard - sort of the first baby step.

(C) 8-9.0, but really 5NS2.1 guess and check, the algebra I standard isn't even exactly correct.

Problem (9)


9. Sally has a container with a square base (shown below). She fills the container one-third full with water. The volume of the water is W cubic units. Which of the following expressions gives the height of the container? (Circle your answer)

(a) $\frac{W}{3a^2}$

(b) $3a^2W$

(c) $\frac{1}{3}a^2W$

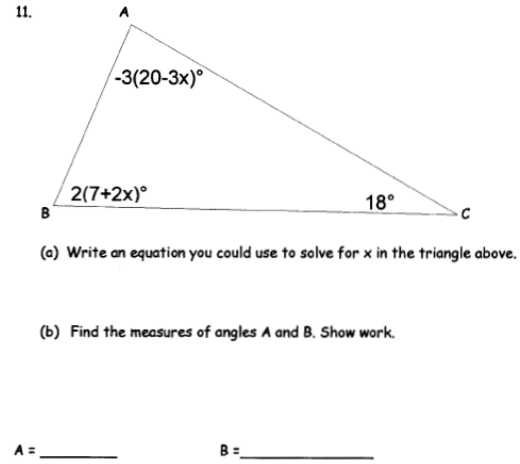
(d) $\frac{3W}{a^2}$

(M) 6MG1.3, 6AF1.2. This is a good - though elementary - problem, except that the "container" should be better specified, perhaps "container which is a rectangular prism with a square base." However, I would suspect that virtually all the students would automatically assume that it's a rectangular prism.
To solve the problem students need to know that the volume of a rectangular prism is Ah where A is the area of the base and h is the height of the prism, or a^2h. Since it is one third full $W = (1/3)a^2h$, so, solving for h, $h = 3W/a^2$. Here are the relevant standards: for the volume, 6MG1.3 Know and use the formulas for the volume of triangular prisms and cylinders (area of base height); compare these formulas and explain the similarity between them and the formula for the volume of a rectangular solid, for setting up and solving the equation 6AF1.2: Write and evaluate an algebraic expression for a given situation, using up to three variables.
(B) 7MG2.1. I actually like the item.
(C) 6MG2.1.

Problem (10) A rectangle of length 78 is cut into 6 congruent rectangles as shown. Each of the smaller rectangles is similar to the large rectangle. Find h. $h =$ ____________________

(M) 7NS2.4, 6NS1.3. This problem amounts to solving the equation $13/h = h/78$. To get this equation requires only that students understand that for similar rectangles the ratio of corresponding sides is a constant. (6NS1.3) To solve this equation involves cross multiplying so $13^2 \times 6 = h^2$ and h (since it is positive) must be $13 \times \sqrt{6}$. The cross multiplication is again 6NS1.3, and taking the square root is 7NS2.4. This has some relation to 7AF2.0, but the latter standard is at a higher level.
(B) 8-20.0. I like this one, too.
(C) 6NS1.3, Find the length of a side of a polygon similar to a known polygon.

Problem (11)


11.

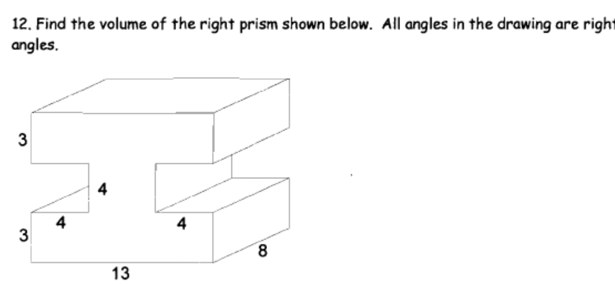
(a) Write an equation you could use to solve for x in the triangle above.

(b) Find the measures of angles A and B. Show work.

A = ________ B = ____________

(a) (M) 5MG2.2. This problem requires that students know that the sum of the measures in degrees of the interior angles of a triangle is 180 degrees. Thus $18 + 2(7 + 2x) - 3(20 - 3x) = 180$. This is given by 5MG2.2.
(B) 8-4.0. But why isn't this standard at Grade 7 or even Grade 6?
(C) 8-4.0, too difficult for 7AF2.1 or 7AF4.1.
(b) (M) 7AF1.3, 7AF4.0. The equation above has to be expanded using the distributive law and negative numbers (the relevant standard is 7AF1.3), obtaining $13x - 28 = 180$ or $13x = 152$, so $x = 11(9/13)$. This is 7AF4.0.
(B) 8-4.0.
(C) 8-4.0.

Problem (12)


12. Find the volume of the right prism shown below. All angles in the drawing are right angles.

(M) 6MG1.3, 4MG1.4. But (B) makes a very good point.
(B) 7MG2.2. This problem has three separate errors. The reader is to assume lots of collinearity that is not given.
(C) 7MG2.1.

VIII. Appendix 3: Test Validity

This is a very serious issue when tests are being used to determine student outcomes, so it might be helpful to give a *brief review of the basic considerations that have to be met in order that a test be viewed as a valid measure.*

Here are the key issues:

> *(1)* ITEM DIFFICULTY*: This is the proportion of students who responded correctly to each item [There is no adequate discussion in this area of how one handles questions that are incorrect or have no correct answer, and the tests in question have both kinds of questions.]*
>
> *(2)* ITEM DISCRIMINATION*: The extent to which an item differentiates between those students with the highest and lowest scores on the total test.*
>
> *(3)* TEST RELIABILITY*: The extent to which a test measures consistently what it is that it measures. There are a number of accepted techniques for establishing the reliability of tests. The usual one for tests that are only given once is the Kuder - Richardson Formula 20, a technique for establishing the internal consistency of a test*
>
> *(4)* TEST VALIDITY*: The extent to which a test measures that which it is supposed to measure. [If it is supposed to measure competency in Algebra I, then how well does it do this?] Within this area there are two sub-topics*
>
> > *4(a)* PREDICTIVE VALIDITY*: this is the extent to which a students performance on a test of unknown validity is related to his or her performance on a test of known validity.*
> >
> > *4(b)* CONTENT VALIDITY*: The extent to which it measures the content that is to be taught.*

IX. APPENDIX 4: AN EXAMPLE OF AN INAPPROPRIATE EXAM QUESTION

Question 9 on the first post-test deserves special comment. This question is not trivial, yet - as stated - could be answered by students with a fourth grade knowledge of mathematics. However, to answer the question in a mathematically complete sense and to understand what is actually going on requires a level of knowledge beyond that which even the strongest high school graduates are likely to have.

> *(9) Here are two circular disks. A positive integer is written on each side of each disk. All four of the numbers are different. By tossing the disks in the air and adding together the numbers that land face up, you can get totals of 8, 9, 11, or 12.*
>
> > *a. If possible, find the numbers on the other side of each disk.*
> >
> > *b. Find a pair of disks that would give you totals of four consecutive integers (for example 2,3,4,5). What numbers are on each side of each disk? Explain how you figured it out.*

AN EXPECTED SOLUTION*: For part a we have the table*

$$\begin{pmatrix} x & y \\ 5 & 7 \end{pmatrix}$$

with $5 + 7 = 12$ and $x + 7$ one of $8, 9, 11$, $x + y$ a second number in the set $8, 9, 11$ and $y + 5$ the third number in this set. Hence, $x = 1, 2$, or 4. Suppose $x = 1$, then $y + 5 = 9$ or 11 and $y = 4$ or $y = 5$ which doesn't work. So suppose $x = 2$ so $x + 7 = 9$. Then $y + 5$ is one of $8, 11$ and $2 + y$ is the other. Since $6 + 5 = 11$ and $6 + 2 = 8$ $x = 2$ works. Solution:

$$\begin{pmatrix} 2 & 6 \\ 5 & 7 \end{pmatrix}.$$

Part b is not well posed. However, if we assume that there is a solution for $2, 3, 4, 5$ as is implied and still require that the numbers be positive integers, then the only way to get 2 is if we have two 1's. In this case a quick check of possibilities shows that the solution is

$$\begin{pmatrix} 1 & 1 \\ 2 & 3 \end{pmatrix}.$$

Otherwise, if we choose to keep the requirement that the numbers all be different we have to abandon the requirement that they be positive integers. If we assume they are non-negative integers, there is a solution:

$$\begin{pmatrix} 0 & 2 \\ 1 & 4 \end{pmatrix}.$$

If the numbers are assumed to be integers, then there are many solutions, and if the numbers are real then there are even more.

Aside from the difficulty coming from the fact that (b) is ill-posed, this is fourth grade level in California as was indicated in Appendix 1.

The method of solution above is characterized as guess and check with only a very limited number of possibilities, so that most students will quickly find solutions, though the question initially looks sophisticated. But suppose that the numbers were much bigger, or we changed the question to the more mathematical question that first does not assume all the numbers are different, second asks for all whole number solutions, and third asks for a demonstration that all the solutions have been obtained, or any of the other variants indicated above.

Then the situation changes, and the question - basically the real question underlying the original problem #9 - becomes both much more interesting and far more difficult. Thus we classify this problem - as it is stated in the context of this test - as a pretend algebra problem . But there is a more advanced context in which the problem is a very good one.

A MATHEMATICAL CONTEXT*: Let us label the disks as follows, so the front sides of the two disks are A and C while the back side of the A disk is B and the back side of the C disk is D. Then there are four possible results giving four linear equations:*

$$\begin{aligned} A + C &= t_1 \\ A + D &= t_2 \\ B + C &= t_3 \\ B + D &= t_4 \end{aligned}$$

Suppose we are given four unordered numbers $\langle a, b, c, d \rangle$. We want to know all the A, B, C, D so that the values of the four sums $A + C$, $A + D$, $B + C$, $B + D$ give the numbers

a, b, c, d in some order. In other words we want to solve the inverse problem for all the (up to 24) different 4-vectors that have the coordinates a, b, c, d in some order.

This is what a great deal of college level linear algebra courses directly address, and the complete solutions for the equations above - both the constraints on the t_i that guarantee solutions exist and all the solutions when they do exist - are easily worked out by students in such a class.

A similar discussion works when we specialize to look at constrained sets such as those in part ((b) where (a, b, c, d) consists of four successive integers, or more generally an unordered set of the form $\langle n, n+1, n+2, n+3\rangle$ for n any real or complex number.

Further, in the situation where the students are the proper level, this can lead to discussions of symmetry - how does the structure of the solutions reflect what happens when we exchange the coins for example.

SUMMARY*: At the fourth grade level this is a good problem to teach exhaustive search and interpretation of data because suitable answers involve small whole numbers and can be found quickly by guess and check methods. A question like this will be seen by people who know only limited amounts of mathematics as entirely appropriate since their tendency is to regard mathematics as lists of disconnected topics, most of which look like combinatoric puzzles. They seem to believe that teaching students to handle (elementary) Putnam style challenge problems is training them in mathematics and problem solving. Thus, such people see including a question like this as the keystone problem in an otherwise un-challenging algebra exam as perfectly legitimate. However, people with a larger grasp of the subject realize that mathematics is incremental, and that solving a problem like this at fourth grade level does not lead to deeper insights. Yet, a problem like this with 2,3,4,5 replaced by integers with appropriate constraints would be a nice problem in a college level linear algebra course.*

X. APPENDIX 5: DESCRIPTION OF THE API SCORE

THE FOLLOWING DESCRIPTION OF THE CALIFORNIA API SCORES IS TAKEN FROM DOCUMENTS PROVIDED BY THE CALIFORNIA DEPARTMENT OF EDUCATION.

The Academic Performance Index (API) is the cornerstone of California's Public Schools Accountability Act of 1999 (PSAA). The purpose of the API is to measure the academic performance and growth of schools. It is a numeric index (or scale) that ranges from a low of 200 to a high of 1000. A school's score on the API is an indicator of a school's performance level. The statewide API performance target for all schools is 800. A school's growth is measured by how well it is moving toward or past that goal. A school's base year API is subtracted from its growth API to determine how much the school improved in a year.

The API score summarizes the results of various indicators (i.e., statewide assessments used in calculating the API). Indicators used in calculating the 2003-04 API reporting cycle include:

Standardized Testing and Reporting (STAR) program

- *California Standards Tests (CSTs)*
  - *English-Language Arts, grades two through eleven, including a writing assessment at grades four and seven*
  - *Mathematics, grades two through eleven*
  - *History-Social Science, grades ten and eleven*
  - *Science, grades nine through eleven*
  - *California Alternate Performance Assessment (CAPA) in English-language arts and mathematics, grades two through eleven*
- *Norm-referenced test (NRT)*
  - *California Achievement Test, Sixth Edition Survey (CAT/6 Survey) in all content areas, grades two through eleven*

California High School Exit Examination (CAHSEE)

- *CAHSEE, grade ten in English-language arts and mathematics*

For the 2003-04 API reporting cycle, the NRT in grades two through eight received 20 percent of the weight in the API, and the CSTs received 80 percent of the weight. The NRT in grades nine through eleven received 12 percent of the weight in the API, the CSTs received 73 percent of the weight, and the CAHSEE received 15 percent of the weight. The weighting demonstrates California's increased emphasis on tests that are closely aligned to state content standards (the CSTs and the CAHSEE) and reflects a commitment towards the full alignment of standards, assessments, and accountability in California public schools.

API Indicator Weights
2003-04 API Reporting Cycle
2003 API Base and 2004 API Growth

	Grades 2-8	Grades 9-12
ELA CST	48%	32%
ELA NRT	12%	6%
ELA CAHSEE		10%
Math CST	32%	16%
Math NRT	8%	3%
Math CAHSEE		5%
Science CST		5%
Science NRT		3%
Social Science CST		20%
Total	100%	100%

To calculate the API, individual student scores from each indicator are combined into a single number to represent the performance of a school. For the NRT, the national percentile rank (NPR) for each student tested is used to make the calculation. For the CSTs, the standards-based performance level

(Advanced, Proficient, Basic, Below Basic, or Far Below Basic) for each student tested is used. For the CAHSEE, a level of pass or not pass is used. The percentages of students scoring within each level are weighted and combined to produce a summary result for each content area. Summary results for content areas are then weighted and combined to produce a single number between 200 and 1000, which is the API for a school.