Education Policy Analysis Archives

Volume 2 Number 10 **July 11, 1994** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.

Editor: Gene V Glass, Glass@ASU.EDU. College of Education,
Arizona State University,Tempe AZ 85287-2411

On the Academic Performance of New Jersey's Public School Children: Fourth and Eighth Grade Mathematics in 1992

Howard Wainer
Educational Testing Service

hwainer@ets.org

Abstract: Data from the 1992 National Assessment of Educational Progress are used to compare the performance of New Jersey public school children with those from other participating states. The comparisons are made with the raw means scores and after standardizing all state scores to a common (National U.S.) demographic mixture. It is argued that for most plausible questions about the performance of public schools the standardized scores are more useful. Also, it is shown that if New Jersey is viewed as an independent nation, its students finished sixth among all the nations participating in the 1991 International Mathematics Assessment.

Introduction

"...teaching is validated by the transformation of the minds and persons of the intended audience." (Bressler, 1991)

On the Genesis of this Report

In January of this year I was approached by three members of the research staff of the New Jersey Education Association (NJEA) with a proposition. They wanted me to do some research that would show that "New Jersey's teachers were doing a good job." They pointed out that NJEA was an advocacy group and that their principal goal was furthering the interests of their members. I told them I understood this, but that "research" means that we don't already know the answer, and that it was ETS policy that there could be no control exercised by a client over the freedom to publish by its researchers (except for usual concerns about protecting the privacy of

examinees). So whatever I found, good or bad, would be written up and disseminated in the same way. They agreed.

Then we discussed the character of the research. They began with the notion that I somehow collect and summarize information that could be formed into some sort of index of teaching quality. I didn't (and don't) know how to characterize the quality of teaching indirectly. I know that some teachers are justly renowned for their bravura performances as Grand Expositor on the podium, Agent Provocateur in the preceptorial, or Kindly Old Mentor in the corridors. I also knew that these familiar roles in the standard faculty repertoire should not be mistaken for teaching, except as they are validated by the transformation of the minds and persons of the students. Thus I was committed to measuring the success of the schools by the performance of the students.

They agreed and suggested that I compare New Jersey's students with other states on their SAT scores. I pointed out there is a substantial literature (much of it by my favorite author - e.g. Wainer, 1986a,b; 1989a,b) that demonstrates quite conclusively that one cannot make effective comparisons among states using the scores of any college admissions test because of the self-selected character of the sample of students who choose to take it. I suggested that the data collected in the NAEP state assessment would be just the thing. They expressed concern that such an investigation would not have as high a profile since SAT scores are often in the news. I agreed that this might be the case, but that we could begin to change this by making a big deal about the importance of NAEP. I pointed out that NAEP had much to recommend it in addition to it being drawn from a representative sample; that it touched several grades, several different subjects, and allowed international comparisons were three reasons that occurred to me at the moment.

The project was a small one, as these things go at ETS. It could be completed in only a few weeks of my time, since most of the results were already available in one NAEP publication or another. My only contribution was in selecting the results that seemed suitable, figuring out what were the most likely questions these results would be asked to answer, placing them into a form that would allow them to properly answer those questions, and displaying them to make the results accessible.

The report was written and went through the usual peer review process at ETS, pursuant to which it was revised and appeared as an ETS Research Report (RR-94-29). Then, in the June 15, 1994 issue of Education Week there was an extended commentary by Chester Finn, a founding partner of the Edison Project, a branch of the Whittle Corporation that is seeking to privatize public schools. He felt that I had painted a too rosy picture of the performance of New Jersey's schools. He was particularly incensed because I had presented state school performance both before and after statistical standardization to a common demographic mixture. He seemed to feel that allowing any formal statistical adjustment opened the door to all sorts of mischief. His alternative, as I understand it, is to leave the summary scores alone and allow individual users to weigh their component pieces subjectively. This is what Samuel Johnson has referred to as "the ancient method." (Note 1)

It is true that an incorrect statistical adjustment will promote incorrect inferences. It is also true that not adjusting when one ought to will also promote incorrect inferences. The issue that we must address in this instance is what questions are NAEP data most likely to be asked to illuminate. Once this is determined what kind of statistical adjustment, if any, will become clear.

Let me illustrate this issue with a simple, but real example: the 1992 NAEP 8th grade math assessment. Nebraska's average score was 277 New Jersey's average score was 271. On the face of it, it appears that 8th grade students in Nebraska do better in mathematics than their counterparts in New Jersey. We note further however that when we examine performance by ethnic group we find:

	White	*Black*
Nebraska	281	236
New Jersey	283	242

How can this be? Even though Nebraska does better overall than New Jersey, New Jersey's students in both of the major ethnic groups outperform their Nebraska counterparts. This is an example of what statisticians have long called Simpson's Paradox (Wainer, 1986c, Yule, 1903). It is caused by the differences in the ethnic distributions in the two states.

	White	*Black*
Nebraska	87%	5%
New Jersey	61%	16%

Each state's mean score is a product of the mean score within each ethnic group and its proportional representation in the population. Thus Nebraska's mean is composed of the White mean weighted by 87% and the much lower black mean weighted by only 5%. In New Jersey Whites represent a much smaller segment of the population and so are given a smaller weight in the calculation of the overall mean.

If we standardize all states to a common demographic mixture, say the demographics of the United States as a whole, we find that New Jersey's standardized mean is 274 and Nebraska's is 270. Which is the right number? Finn suggests the it is the unstandardized figure.To answer this we have to know what is the question that the number will be answering.

If the question is of the sort, "I want to open a business in either New Jersey or Nebraska. Which state will provide me with a population of potential employees whose knowledge of mathematics is, on average, higher?" The unadjusted mean scores provide the proper answer.

If the question is, "I want to place my child in school in either New Jersey or Nebraska. In which state is my child likely to learn more mathematics?" The standardized scores give the right answer. One can see this immediately by imagining a sequence of questions that someone trying to help the parent phrasing the above question might ask."Does your child have a race? If your child is White, he/she is likely to do better in New Jersey. If he/she is Black, he/she is likely to do better in New Jersey." Presenting the data in a disaggregated way allows these sorts of questions to be answered specifically, but if a single, overall number is needed to summarize the performance of a state's children, for questions like this, one must standardize.

I contend that NAEP data are gathered to illuminate the performance of schools. Different schools face tasks of differing difficulty depending upon the particular mix of students that attend. If we want to make comparisons among schools that are about the schools and not about the mix of students we must standardize. Not doing so is wrong and misleading, exactly the opposite, if I may be permitted the obiter dictum, of what Mr. Finn suggests.

The Report in Question

The most critical measure of any educational system is the performance of its students. But what yardstick should be used to accomplish this measure? The fact that modern education has many goals suggests that we must measure the extent of its success in a variety of ways. This report describes the first of a series of researches that will attempt to characterize the performance of New Jersey's public school system. We will do this through comparative and

absolute measures, the primary instrument of which will be the data gathered during the course of the National Assessment of Educational Progress (NAEP).

NAEP is a Congressionally mandated survey of the educational achievement of American students and of changes in that achievement across time. Although this survey has been operational for nearly 25 years, it was only in 1988 that Congress authorized adding state level surveys to the national assessment. This was begun on a trial basis with states participating on a voluntary basis. In 1990 37 states, two territories and the District of Columbia participated in the first Eighth Grade Math Assessment. In 1992 seven more states joined the state assessment yielding 44 jurisdictions. The 1992 assessment was expanded to also include the Fourth grade. In this report we shall focus attention only upon the 41 states in the assessment. Guam, the Virgin Islands and the District of Columbia will be explicitly excluded because they are sufficiently different from the states in their size, character and composition so as to distort most comparisons. The assessment methodology is technically sophisticated. Through the use of linking items and item response theory, the performance of all students participating in the assessment can be placed on the same numerical scale. Measuring students' growth is thus straightforward. Subtracting 4th grade scores from 8th grade scores is the growth obtained. Consequently the expansion of the assessment to the fourth grade provides us with two important bits of information. First, is a measure of how much mathematics Fourth graders know. Second, a measure of how much mathematics is learned between 4th and 8th grade. Note that having a measure of the gain obtained (about 49 NAEP points on average) helps us to interpret the NAEP scale. It tells us that if one state trails another by about 12 points this is about the same as the average gain in one year of school. Thus, when we compare California's mean 8th grade NAEP score of 261 to New Jersey's score of 273, we can interpret the 12 point difference as indicating that the average California 8th grader performs about the same in mathematics as the average New Jersey 7th grader would have. This helps give additional meaning to the numerical scale.

More meaning still for the eighth grade math assessment is yielded by comparing performance on it with performance of 13 year old students in the 1991 International Assessment. Because the NAEP Math assessment was coordinated with the International Assessment both sets of scores can, with reasonable confidence, be placed on the same scale. This was accomplished by having a common sample of examinees for both assessments. As we shall see, the performance of New Jersey's students compares favorably with those from the developed nations.

The Mathematics assessment contained tasks for the students drawn from the framework provided by the Curriculum and Evaluation Standards for School Mathematics, developed by the National Council of Teachers of Mathematics. The content and the structure of the assessment has been widely praised as being representative of the best that current knowledge and technology allows. A full description of the 1992 Mathematics Assessment is found in NAEP 1992: Mathematics Report Card for the Nation and the States (Mullis, Dossey, Owen, & Phillips; 1993).

The NAEP State Assessment Sample

Within each state 100 public schools are carefully selected to be representative of all public schools in that state. Within each school at least 30 students are chosen at random to be tested (in larger schools this number can be as large as 90). Students (usually of foreign birth) whose English language proficiency is deemed to be insufficient to deal with the test, are excluded from the sampling frame.

The Results

All results are reported on a uniform scale that can meaningfully characterize the

performance of students over a very wide range of proficiency. This scale can be used in a normative manner, for example comparing one state with another, or one state with the nation as a whole. Or it can be used as an absolute measure, since expert judges have provided a correspondence between score levels and specific proficiencies. These proficiencies are denoted Basic, Proficient and Advanced and what is required to perform at each of these levels obviously increases as the student progresses through school, but are always referred back to the five NAEP content areas. These are: (1) numbers and operations, (2) measurement, (3) geometry, (4) data analysis, statistics, and probability, and (5) algebra and functions.

For example, a score of 211 is characterized as "Basic Level" fourth grade performance. "Basic Level" is defined as "showing some evidence of understanding the mathematical concepts and procedures of the five NAEP content areas." The second level is called "Proficient" and is located at score 248 and reflects being able to "consistently apply integrated procedural knowledge and conceptual understanding to problem solving in the five NAEP areas." The highest performance level is termed "Advanced," is located at score 280 and reflects the ability to "apply integrated procedural knowledge and conceptual understanding to complex and nonroutine real-world problem solving in the five NAEP areas."

The mean performance of all participating states for the 8th grade assessment is shown in Figure 1.

Figure 1. A stem & leaf display of the 1992 NAEP State Assessment in 8th Grade Mathematics. These results are the raw (unstandardized) means from each state. New Jersey ranks 14th among all participants.

NAEP 1992 Mathematics Assessment Overall Proficiency-8th Grade Mathematics (unstandardized)

283	Iowa North Dakota
282	Minnesota
281	
280	
279	
278	Maine New Hampshire
277	Nebraska Wisconsin
276	
275	
274	Idaho Wyoming Utah
273	Connecticut
272	Colorado Massachusetts
271	New Jersey Pennsylvania
270	Missouri
269	Indiana
268	
267	Michigan Oklahoma Virginia Ohio
266	Delaware
261	Kentucky
260	California South Carolina
259	Florida New Mexico Georgia
258	West Virginia Tennessee North Carolina
257	Hawaii
256	
255	Arkansas
254	

253
252
251 Alabama
250
249 Louisiana
248
247
246 Mississippi

The results shown in Figure 1, while accurately reflecting the actual mean performance within each state, may not be appropriate for certain kinds of state-by-state comparisons. The student populations of each state differ in their demographic make-up. As such, some states face more difficult challenges in educating their populations than others. One obvious example of such a situation occurs in states like California, New Jersey and Florida that have large immigrant populations whose children, even if they do not participate directly in the assessment, require a larger share of instructional resources than native English speakers. In addition, the various subpopulations of students in each state often perform very differently from one another. For example, in Figure 2 are displayed the mean performance of students in different parts of the country broken down by race/ethnicity.

Figure 2 has two important messages:

1. There are very large differences in performance by ethnic group. These differences are much larger than the geographic variation observed.
2. New Jersey's students perform better than the national average and all regional averages for all groups. Thus although it is true that New Jersey's African-American and Hispanic students do worse than White students, they do better than African-American and Hispanic students in any region.

Figure 2. A stem & leaf depiction comparing the performance New Jersey's students, broken down by race/ethnicity, with similar groups from all other parts of the country. (Samples of "Asian/Pacific Islanders" were insufficient to obtain accurate estimates for any other regions than the West.)

(Note to Reader: The horizontal spacing of the entries in this figure are significant. If you view this figure in a proportional font such as Times or New York it will be distorted. The proper spacing is retained in Courier.)

NAEP 1992 Trial State Assessment Subgroup comparisons of NJ with other parts of the Nation

Race/Ethnicity

Grade 8 Mathematics	White	Asian/Pacific Islander	Hispanic	African-American
297		NJ		
296				
295				
294				
293				
292				
291				
290				

289
288
287 NATION
286 West
285
284
283 NJ
282
281
280 Central
279 Northeast
278
277 West
276 NATION
275
274
273
272
271
270
269 Southeast
. . . .
. . . .
. . . .
. . . .
247 NJ
246 Central & West
245 NATION
244
243
242 NJ
241 Northeast
240
239 Southeast Northeast & Central
238
237
236 NATION
235
234 West
233 Southeast

In addition to the widely different performances of the various demographic subgroups, the distribution of these groups is not uniform across all states. A brief summary of these distributions is shown in Table 1. As is evident, New Jersey's racial/ethnic distribution is rather close to that of the nation as a whole. The central states are the most deviant in the sense that they have a substantially larger proportion of their student population that is White.

Table 1. The national and regional racial/ethnic distribution.

NAEP 1992 Trial State Assessment Percentage Race/Ethnic Representation in NJ Compared to that in other parts of the Nation

	Race/Ethnicity

Mathematics	*White*	*Black*	*Hispanic*	*Asian*	*Other*
NATION	69	17	10	2	2
Northeast	68	20	9	2	1
Southeast	63	29	5	1	2
Central	79	11	7	1	2
West	65	11	16	5	3
NJ	67	14	13	5	1

If we wish to use such data to draw inferences about the relative efficacy of a state's schools, it is considered good practice to statistically adjust for the demographic differences. Why is it helpful to make such adjustments? It is beyond the goals of this report to investigate fully why there are differences in performance by demographic group, although there is a rich literature of fact and conjecture that attempts to do so (Note 2).

However, to understand why we need to make a statistical adjustment it is important to provide some explanation. To do so requires that we draw the important distinction between education and schooling. The school is only one agency among many -- family, church, neighborhood, mass media -- that provides children with windows on the world. Mass schooling was invented because families could no longer perform essential educational functions. "Once upon a time schools could proceed on the only partly fictitious assumption that, in their efforts to teach children, they were supported by relatively stable families, and by neighborhoods that enforced elementary standards of civility." (Note 3) Not only is this much less true now than in the past, but also it is less true within some demographic groups than others. NAEP measures education and not just schooling, but inferences about the differences among states are explicitly about schools. To add some validity to those inferences we must try, statistically, to place all states on a level playing field with respect to their children's nonschool educational opportunities. Adjusting for differences in demographic groupings is a crude beginning.

What follows is a methodology that recognizes that such differences exist, and a statistical technology that attempts to partition state differences due to demographics from those due to differences in school performance.

Interpretable comparisons through statistical standardization

The between-state comparisons that are implicit in Figure 1 can yield misleading inferences if one is not acutely aware of the differences in the demographic make-up of all of the constituent units. This is of such a complex nature that it is impossible to keep things straight without some formal adjustment. One accepted way to accomplish this adjustment is termed "standardization" (Mosteller & Tukey, 1977, p. 223). The basic idea of standardization is to choose some specific demographic proportions as the standard and then estimate each state's mean proficiency on that specific mixture. In this instance it is sensible to choose the configuration of the entire United States as the standard mixture. Thus the estimated score for each state will be the answer to the question, "What would the national average be if all children went to school in this state?"

How is this adjustment accomplished? It is very simple in theory, although sometimes, because of peculiarities in sampling weights, a bit tricky to execute. We take the mean score obtained in a state for a particular subgroup and multiply it by that subgroup's proportional representation in the standard (national) mixture. Do this for all subgroups and the resulting score is the adjusted one. So far we have reported New Jersey's scores for four racial/ethnic groups. As we have seen, because New Jersey's demographics are so much like the national standard this

sort of adjustment would have little effect. A greater effect would be on more disparate areas (i.e. the central US,). Is it sufficient to adjust simply on the basis of this one demographic variable? No, although if we adjust on too many variables, and so include some irrelevant ones, no damage will be done, since irrelevant variables will typically not show differences in performance. In this paper we adjust on three variables. These are:

1. Race/ethnicity - Five categories: White, African-American, Hispanic, Asian/Pacific Islander, Other
2. Type of community - Four categories: Extreme Rural, Advantaged Urban, Disadvantaged Urban, Other.
3. Limited English Proficiency - Two categories: Yes, No

This resulted in dividing each state up into forty pieces corresponding to the forty possible combinations (5X4X2) and calculating the mean proficiency within each of those 40 groups. These 40 means were then weighted by their representation in the entire nation yielding a standardized score for each state. The results of this standardization are shown in Figure 3.

Figure 3. After standardization New Jersey ranks fourth among all participating states in the 1992 8th grade mathematics assessment.

NAEP 1992 Trial state Assessment (Standardized for demographic differences)

Grade 8 Mathematics	
278	North Dakota
277	
276	
275	Iowa Minnesota
274	New Jersey Maine New Hampshire
273	Idaho
272	Connecticut
271	Massachusetts Wisconsin
270	Nebraska Wyoming Utah
269	Texas
268	Colorado Pennsylvania New York Virginia Missouri
267	Arizona California Maryland Indiana Michigan
266	NATION
265	South Carolina Oklahoma Ohio
264	Delaware New Mexico Florida
263	Rhode Island
262	Georgia
261	
260	Kentucky
259	North Carolina
258	
257	Hawaii Tennessee
256	
255	Alabama Louisiana Arkansas West Virginia
254	Mississippi

After standardization to national demographic norms we find that although New Jersey's

mean score has not changed much, ten other states, with more homogeneous student populations (e.g., Massachusetts, Wisconsin, Iowa, Idaho, North Dakota) that had previously been slightly higher are now ranked equal to or below New Jersey.

What is the point of standardization? There are many reasons. So far we have mentioned just one -- making comparisons between states on the basis of their children's performance on the same tasks and not on the differences in the demographic structure of their population. A second, and oftentimes more important use of standardized scores is in easing the difficulties in making inferences about changes that occur within a state across time. When changes do occur the standardized scores assure that the change reflects changes in the students' performance and not changes in the demographic structure of the state. We expect that as time goes on this will be the aspect of greatest value of the standardization. (Note 4)

A natural question to ask is, "At what age do the differences observed among the states manifest themselves?" If we see the same difference between two states in 4th grade as we do in 8th, it implies that the lower scoring state needs to place more emphasis on learning in lower grades. If the difference observed in 4th grade grows proportionally larger in 8th it means that the deficit is spread throughout the years of school and a more systemic change is needed for improvement. Trying to make inferences of this sort based on just two time points is risky, but is certainly instructive. Shown in Figure 4 are the standardized scores for the 1992 4th grade math assessment. A comparison with the 8th grade rankings shown in Figure 3 indicates that the positions established in 4th grade are maintained and the differences observed between states increase. The range of 24 NAEP points observed between the relatively extreme states of North Dakota and Mississippi in 8th grade was 13 NAEP points in 4th grade. One way to interpret this is that the average Mississippi 4th grader was a year behind the average North Dakota 4th grader in math, and by the time they both reached 8th grade this deficit had increased to two years.

Figure 4. The standardized scores for the 41 states in the 1992 4th grade math assessment.

NAEP 1992 Trial state Assessment (Standardized for demographic differences)

Grade 4 Mathematics

227	New Hampshire
226	Maine
225	
224	Connecticut
223	New Jersey Iowa Wisconsin
222	North Dakota Pennsylvania
221	Minnesota Texas Wyoming Virginia Massachusetts
220	Nebraska Missouri New York
219	Maryland Georgia
218	Colorado Idaho Indiana Michigan Delaware Oklahoma
217	NATION Utah Ohio Arizona
216	South Carolina
215	New Mexico North Carolina
214	Rhode Island Florida
213	Kentucky
212	California West Virginia
211	Hawaii
210	Tennessee Alabama Arkansas Louisiana
209	Mississippi

By subtracting the scores shown in Figure 4 from those in Figure 3 we obtain estimates of the average growth exhibited in each state. This result is shown in Figure 5 below. All states' scores are standardized to the demographic structure of the nation as a whole. Thus were these results longitudinal rather than cross- sectional, we would be able to interpret the changes as due entirely to growth and not demographic changes. As they are now constituted these changes in scores are due to differences in performance and not to demographic differences in the two grades.

Figure 5. Standardized estimates of change in mathematics performance seen by state between 4th and 8th grade in the 1992 assessment. New Jersey's gain was the seventh largest.

Gain in Mathematics Proficiency from 4th to 8th grade
(Scores are standardized to entire US population)

56	North Dakota
55	California Idaho
54	Minnesota
53	Utah
52	Iowa
51	New Jersey
50	Arizona Colorado Florida Massachusetts Nebraska
49	NATION Indiana Michigan New Mexico Rhode Island Wyoming South Carolina Ohio Texas
48	Connecticut Maine Wisconsin New York Maryland Missouri
47	Kentucky Oklahoma Virginia Tennessee New Hampshire
46	Delaware Hawaii Pennsylvania
45	Alabama Arkansas Louisiana Mississippi
44	North Carolina
43	Georgia West Virginia

International Comparisons

As mentioned previously, the 1991 International Assessment contained enough NAEP items to allow accurate comparisons. The most newsworthy result was that the United States finished near the bottom in this assessment, finishing ahead of Jordan but behind all of the participating developed nations. This was (properly) viewed with alarm. But, as we have seen in the preceding figures, there is tremendous variation within the United States. Shown in Figure 6, are the results of this assessment augmented by the inclusion of New Jersey (standardized to national demographics). As is evident, New Jersey's students' performance was sixth among all nations participating in the assessment. Further details of the International Assessment can be found in Salganik, Phelps, Bianchi, Nohara, & Smith (1993).

Figure 6. Placing New Jersey explicitly into the 1991 International Assessment shows that its students performed above the average level of most developed nations.

International 1991 Mathematics Assessment
(Predicted Proficiency for 13 year olds)

285	Taiwan

284
283 Korea
282
281
280
279 Soviet Union Switzerland
278
277 Hungary
276
275
274 New Jersey
273 France
272 Italy Israel
271
270 Canada
269 Ireland Scotland
268
267
266 Slovenia
265
264
263 Spain
262 United States
.
.
.
246 Jordan

Interpretation of this figure is helped by remembering that, on average, students advance roughly 12 NAEP points a year. Thus the average student in Taiwan and Korea is about a year ahead of the average New Jersey student, who is within a month or two of the other developed nations.

Thus we see rather dramatically that because of the great diversity within the United States looking at just an overall figure for the entire country provides an incomplete and, for some purposes, misleading picture. Because New Jersey's score is standardized to the demographic structure of the entire nation one can interpret this result as what the nation's location would have been if all of the states' educational systems performed as well as New Jersey's.

Within State Variation

We have seen that the variation among states (roughly 30 NAEP points from highest to lowest) makes interpretation of a national mean of limited value. In the same way, the variation within states dwarfs the variation between them. In most states the average score obtained by the lowest 10% of the students is more than 90 points below the score obtained by the top 10%. (Note 5) 90 points is an enormous gulf. Before trying to understand the reasons for this great disparity (with an eye toward developing strategies for ameliorating it) it will be useful to continue this series of comparisons for one important segment of the population --the very top.

In Figure 7 is a comparison of the performance of the top 5% in the 1992 8th grade math assessment with the top 5% of the various OECD countries. We see immediately that New Jersey's top 8th grade students compare favorably with their counterparts throughout the world.

The United States as a whole has also improved relative to the other countries, but still lags the other developed nations by from 3 to 12 months.

Figure 7. New Jersey's top students rank third in the world in the 8th grade math assessment.

International 1991 Mathematics Assessment
(Predicted Proficiency for 95th %ile of 13 year olds)

345	Taiwan
	.
	.
	.
335	Korea
334	
333	
332	
331	
330	
329	
328	New Jersey
327	
326	Hungary
325	
324	Soviet Union
323	
322	Switzerland
321	
320	
319	France
318	
317	Israel Italy
316	Ireland
315	Scotland Canada
314	
313	
312	United States
311	Slovenia
310	
309	
308	
307	
306	Spain
	.
	.
	.
296	Jordan

Summary & Conclusions

This report is the beginning of a series that examines of the performance of New Jersey's school children relative to other children within the United States and world-wide. The measure of performance used was the 1992 National Assessment of Educational Progress mathematics exams and their linked versions used in the 1991 International Assessment. This is done in the ardent belief that the efficacy of schools must be measured by the performance of their students. We chose NAEP for several reasons, three of which were:

1. It is composed of test items that satisfy the best of current wisdom with respect to both their content and their form.

2. The psychometric model underlying the scoring of NAEP yields a single scale on which not only can the fourth grade and eighth graders be characterized, but also the 13 year olds from the OECD countries from around the world.

3. The students sampled by the NAEP are drawn in a principled way from the populations of interest. This in sharp contrast to the sorts of self-selected samples that are represented by state means of such college admission tests as the SAT and the ACT. It is well known that trying to draw inferences of useful accuracy from such self-selected samples is impossible (Wainer, 1986a, b; 1989a, b).

We concur with prevailing expert opinion that of all broad-based tests NAEP provides the most honest and accurate estimates of the performance of the students over the broad range of jurisdictions sampled.

We found that, based on the unstandardized results of the 1992 Mathematics Assessment, New Jersey was among the highest performing states. Once these results were standardized to reflect a single (national) demographic composition New Jersey's rank among the participating states increased to fourth. The United States finished next to last when the performance of its students was compared with that of the students in the other 14 participating OECD nations in the 1991 International Assessment. New Jersey's students were ranked sixth on the same assessment when their performance was placed on the same scale. However New Jersey's best students, its top 5%, when compared with the performance of the top 5% of all other OECD nations, ranked third; trailing only Taiwan and Korea.

Notes

This research was supported by the New Jersey Education Association. I am pleased to acknowledge their help. Furthermore I am grateful for the advice and help of John Fremer, Gene Johnson, Philip Leung and John Mazzeo.

1. Samuel Johnson "To count is modern practice, the ancient method was to guess"; but even Seneca was aware of the difference-"Magnum esse solem philosophus probabit, quantus sit mathematicus."

2. The Coleman report (Coleman et al., 1966) remains the most encyclopedic of such investigations, summarizing, as it does, the performance of more than 645,000 children in 4,000 public schools. It arrives at the not surprising conclusion that family and economic variables drive educational achievement.

3. This quote and much of the surrounding logic comes from Marvin Bressler's delightful and wise 1992 essay, "A teacher reflects."

4. A caveat: Big changes as a result of a statistical adjustment tell us that great care must be exercised in making inferences. A careful comparison of Figure 1 and Figure 3 reveals that most states do not change very much. This is evidence that the standardization is generally behaving as it ought, for if one disagrees with the structure of the adjustment one can still be content that it

isn't changing anything drastically. A notable exception to this would be the District of Columbia, whose small size and atypical demographic structure would combine to yield an enormous shift. Inferences about the meaning of its standardized location ought not be the same as those drawn about the states. For the more important purpose of tracking changes in a jurisdiction's performance over time, it is probably prudent to develop a special standardization for each of the four most unusual jurisdictions (DC, Hawaii, Guam, Virgin Islands).

5. In all but one of the OECD countries this gulf between the 10th percentile and the 90th is somewhat smaller, about 70 points. Taiwan is the lone exception a difference of 96 points.

References

Bressler, M. (1991). Reflections on teaching. In Teaching at Princeton. Princeton, NJ: Princeton University.

Bressler, M. (1992). A teacher reflects. Princeton Alumni Weekly, 93(5), 11-14.

Coleman, J. S. et al (1966). Equality of Educational Opportunity. Washington, D.C.: U.S. Office of Education.

Finn, C.E. (June 15, 1994). Drowning in Lake Wobegone. Education Week, P. 31,35.

Mosteller, F., & Tukey, J. W. (1977). Data analysis and regression. Reading, MA: Addison-Wesley.

Mullis, I. V. S., Dossey, J. A., Owen, E. H. & Phillips, G. W. (1993). NAEP 1992: Mathematics Report Card for the Nation and the States. Report No. 23-ST02. Washington, DC: National Center for Education Statistics.

Salganik, L. H., Phelps, R. P., Bianchi, L., Nohara, D., & Smith, T. M. (1993). Education in States and Nations: Indicators Comparing U.S. States with the OECD Countries in 1988. NCES Report No. 93-237. Washington, DC: National Center for Education Statistics.

Wainer, H. (1986a). Drawing inferences from self-selected samples. New York: Springer-Verlag.

Wainer, H. (1986b). Five pitfalls encountered while trying to compare states on their SAT scores. Journal of Educational Measurement, 23, 69-81.

Wainer, H. (1986c). Minority contributions to the SAT score turnaround: An example of Simpson's paradox. Journal of Educational Statistics, 11, 229-244.

Wainer, H. (1989a). Eelworms, bulletholes & Geraldine Ferraro: Some problems with statistical adjustment and some solutions. Journal of Educational Statistics, 14, 121-140 (with discussions). Reprinted in Shaffer, J. P. (Ed.) (1992). The role of models in nonexperimental social science (pps. 129-148). Washington, D.C.: American Educational Research Association & American Statistical Association.

Wainer, H. (1989b). Responsum. Journal of Educational Statistics, 14, 187-200. Reprinted in Shaffer, J. P. (Ed.) (1992). The role of models in nonexperimental social science (pps. 195-207). Washington, D.C.: American Educational Research Association & American Statistical Association.

Yule, G. U. (1903). Notes on the theory of association of attributes of statistics. Biometrics, 2, 121-134.

About the Author

Howard Wainer

Howard Wainer received his Ph.D. from Princeton University in 1968, after which he was on the faculty of the University of Chicago. He worked at the Bureau of Social Science Research in Washington during the Carter Administration, and is now Principal Research Scientist at the Educational Testing Service. He was awarded the Educational Testing Service's Senior Scientist Award in 1990 and was selected for the Lady Davis Prize. He is a Fellow of the American Statistical Association. His latest book, *Visual Revelations*, will be published by Copernicus Books (a division of Springer-Verlag) in April of 1997.

Aimee Howley
ess016@marshall.wvnet.edu

Craig B. Howley
u56e3@wvnvm.bitnet

William Hunter
hunter@acs.ucalgary.ca

Richard M. Jaeger
rmjaeger@iris.uncg.edu

Benjamin Levin
levin@ccu.umanitoba.ca

Thomas Mauhs-Pugh
thomas.mauhs-pugh@dartmouth.edu

Dewayne Matthews
dm@wiche.edu

Mary P. McKeown
iadmpm@asuvm.inre.asu.edu

Les McLean
lmclean@oise.on.ca

Susan Bobbitt Nolen
sunolen@u.washington.edu

Anne L. Pemberton
apembert@pen.k12.va.us

Hugh G. Petrie
prohugh@ubvms.cc.buffalo.edu

Richard C. Richardson
richard.richardson@asu.edu

Anthony G. Rud Jr.
rud@purdue.edu

Dennis Sayers
dmsayers@ucdavis.edu

Jay Scribner
jayscrib@tenet.edu

Robert Stonehill
rstonehi@inet.ed.gov

Robert T. Stout
aorxs@asuvm.inre.asu.edu