Chapter 13
The Implications of Understanding That PISA Is Simply Another Standardized Achievement Test

David C. Berliner

It occurred to me one day that despite all the excitement, and both the satisfaction and handwringing engaged in by some nations after scores are released, that the Program for International Student Assessment, PISA, is merely another Standardized Achievement Test. Almost all Standardized Achievement Tests (SATs) try to adhere to certain principles of design, have similar correlates, and have similar limits on the interpretations of the results obtained. Neither the popular press nor most politicians ordinarily understand these realities and their implications. Opinion makers are unaware that many of the Standardized Achievement Tests we commonly use do not have the powers attributed to them. It is not far from the truth to call the scores derived from some of these assessments "talismanic" (Haney et al. 1987). That is, for many people, test scores have special powers, particularly of prophesy, a bit like the Kabbalah of the middle ages. Scores from SATs are a part of the metrification associated with the modern world, no doubt aided by a global market place in which business leaders and technologists, instead of humanists and educators, have garnered political power. Contributing to this trend toward metrification has been the ascendance of economics as an influential discipline throughout the world (Lingard et al. 2015). But in my opinion, economists, journalists, and politicians too often seek in metrics powers that are more illusionary than they are real.

D. C. Berliner (✉)
Division of Educational Leadership and Policy Studies, Mary Lou Fulton Teachers College,
Arizona State University, Tempe, AZ, USA
e-mail: berliner@asu.edu

G. Fan, T. S. Popkewitz (eds.), *Handbook of Education Policy Studies*,
https://doi.org/10.1007/978-981-13-8343-4_13

13.1 What Do We Know About High-Quality SATs and PISA

PISA is simply another SAT, so we have knowledge with which to criticize it, because over the decades, we have learned what constitutes a high-quality SAT. Well-designed SATs should have items that have been written carefully, been scrutinized well, tried out, demonstrate little gender or cultural bias, and contribute to test reliability. The total test must also provide convincing arguments about its validity for particular purposes. Not all SATs meet such standards in an exemplary manner, and PISA is no exception.

13.1.1 Language and PISA Items

PISA is being used cross-nationally. Thus, every item in this SAT must have the same meaning in each country to insure that each country has items that are not positively or negatively associated with the all important passing rates for items. If this condition cannot be met, interpretation of this SAT may be seriously compromised. Many scholars, myself included, are not sure that these basic criteria, for this particular SAT, can be met, although PISA designers say that they can do just that. Common sense and supporting research challenge PISA's claims.

For example, here, alphabetically, are just a few of the countries that take the *same* items: Argentina, Australia, Austria, Azerbaijan, Belgium, Brazil, Bulgaria, Canada, Chile, Colombia, Croatia, the Czech Republic, Denmark, and so on. It is quite likely that it is quite difficult to have test items mean precisely the same thing in each of these nations, and thus be an item of equivalent difficulty in languages of each of these nations.

Some of us find it hard to believe that *item equivalence* can be assured for the 65 nations, 65 cultures and their sub-cultures, 65 dialects and languages, which participated in the 2012 PISA test. My colleague Gene Glass (2012), one of the most distinguished educational researchers in the world, asks:

> How do you write a reading test in English and then translate it into Swedish (or vice versa) and end up confident that one is not intrinsically more difficult than the other? I insist that the answer to that question is that you can't. And to claim that one has done so merely sweeps under the rug a host of concerns that include grammatical structure, syntax, familiarity of vocabulary, not to mention culture of the students taking the test.

Years ago, Gerald Bracey (1991) pointed to one international test where 98% of Finnish students, but only 50% of American students, scored correctly on a vocabulary question. The students were asked to indicate whether "pessimistic" and "sanguine" were antonyms or synonyms. Because "sanguine" does not exist in the Finnish language, the word "optimistic" was substituted, making the question much easier to answer.

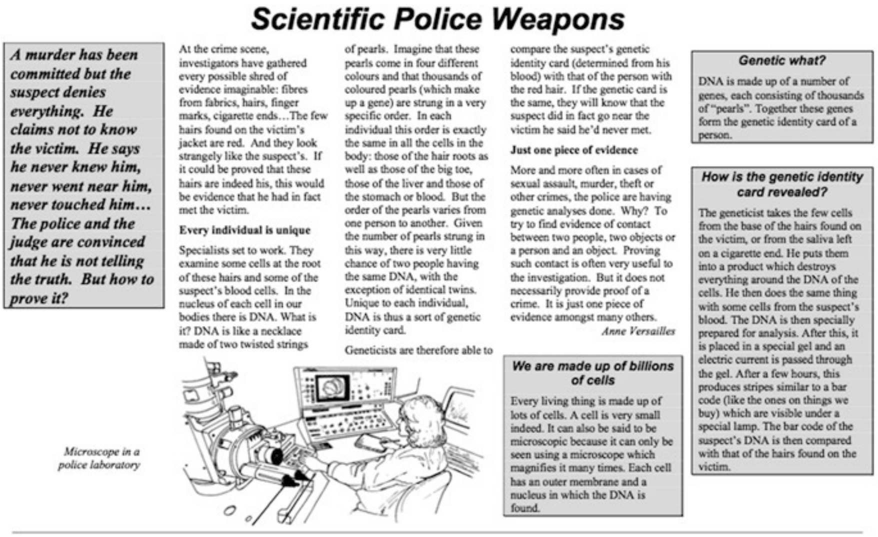
So, common sense leads many of us to believe that no one can produce two non-trivial passages of text, in two different languages, and make them, and the questions derived from them, of equal cognitive difficulty. Figure 13.1 presents an item from the 2006 PISA reading test (Organization for Economic Cooperation and Development 2009). It illustrates this concern. It is a passage in the reading assessment intended to tap comprehension. Please read a little of it. It sets up this passage this way: "A murder has been committed but the suspect denies everything. He claims not to know the victim. He says he never knew him, never went near him, never touched him … The police and the judge are convinced that he is not telling the truth. But how to prove it?"

It is hard for me to believe that such an item is of the same cognitive and emotional character in scientifically advanced countries with appreciation for the police, and scientifically less advanced countries, with fear of the police.

Figure 13.2 is also from the 2006 PISA exam (Organization for Economic Cooperation and Development 2009). It is an example of a problem-solving context designed to assess the quality of scientific thinking by 15-year olds. Please read a bit of it (see Ruiz-Primo and Li 2015).

Although PISA developers claim otherwise, it is hard to believe that the substantial amount of reading required in these two items is likely to be interpreted the same in, say, Hungary, Denmark, and Korea. Nor are the questions associated with each of these contexts likely to yield equal pass rates. Glass (2012, response to BLOG comments) says this:

> It is not a matter of the fidelity of a translation. It is a matter of producing psychometric equivalence right down to percentage points of difficulty between two items. Even small differences in item difficulty between two items in different languages accumulated across several items could produce differences between two nations of the magnitude observed for



Scientific Police Weapons

A murder has been committed but the suspect denies everything. He claims not to know the victim. He says he never knew him, never went near him, never touched him… The police and the judge are convinced that he is not telling the truth. But how to prove it?

At the crime scene, investigators have gathered every possible shred of evidence imaginable: fibres from fabrics, hairs, finger marks, cigarette ends…The few hairs found on the victim's jacket are red. And they look strangely like the suspect's. If it could be proved that these hairs are indeed his, this would be evidence that he had in fact met the victim.

Every individual is unique

Specialists set to work. They examine some cells at the root of these hairs and some of the suspect's blood cells. In the nucleus of each cell in our bodies there is DNA. What is it? DNA is like a necklace made of two twisted strings of pearls. Imagine that these pearls come in four different colours and that thousands of coloured pearls (which make up a gene) are strung in a very specific order. In each individual this order is exactly the same in all the cells in the body: those of the hair roots as well as those of the big toe, those of the liver and those of the stomach or blood. But the order of the pearls varies from one person to another. Given the number of pearls strung in this way, there is very little chance of two people having the same DNA, with the exception of identical twins. Unique to each individual, DNA is thus a sort of genetic identity card.

Geneticists are therefore able to compare the suspect's genetic identity card (determined from his blood) with that of the person with the red hair. If the genetic card is the same, they will know that the suspect did in fact go near the victim he said he'd never met.

Just one piece of evidence

More and more often in cases of sexual assault, murder, theft or other crimes, the police are having genetic analyses done. Why? To try to find evidence of contact between two people, two objects or a person and an object. Proving such contact is often very useful to the investigation. But it does not necessarily provide proof of a crime. It is just one piece of evidence amongst many others.

Anne Versailles

Genetic what?

DNA is made up of a number of genes, each consisting of thousands of "pearls". Together these genes form the genetic identity card of a person.

How is the genetic identity card revealed?

The geneticist takes the few cells from the base of the hairs found on the victim, or from the saliva left on a cigarette end. He puts them into a product which destroys everything around the DNA of the cells. He then does the same thing with some cells from the suspect's blood. The DNA is then specially prepared for analysis. After this, it is placed in a special gel and an electric current is passed through the gel. After a few hours, this produces stripes similar to a bar code (like the ones on things we buy) which are visible under a special lamp. The bar code of the suspect's DNA is then compared with that of the hairs found on the victim.

We are made up of billions of cells

Every living thing is made up of lots of cells. A cell is very small indeed. It can also be said to be microscopic because it can only be seen using a microscope which magnifies it many times. Each cell has an outer membrane and a nucleus in which the DNA is found.

Microscope in a police laboratory

Fig. 13.1 A PISA reading item. (Source: OECD (2009). PISA Released Items—Reading. Retrieved from: http://www.oecd.org/pisa/38709396.pdf)

A copying machine for living beings?

Without any doubt, if there had been elections for the animal of the year 1997, Dolly would have been the winner! Dolly is a Scottish sheep that you see in the photo. But Dolly is not just a simple sheep. She is a clone of another sheep. A clone means: a copy. Cloning means copying 'from a single master copy'. Scientists succeeded in creating a sheep (Dolly) that is identical to a sheep that functioned as a 'master copy'.

It was the Scottish scientist Ian Wilmut who designed the 'copying machine' for sheep. He took a very small piece from the udder of an adult sheep (sheep 1). From that small piece he removed the nucleus, then he transferred the nucleus into the egg-cell of another (female) sheep (sheep 2). But first he removed from that egg-cell all the material that would have determined sheep 2 characteristics in a lamb produced from that egg-cell. Ian Wilmut implanted the manipulated egg-cell of sheep 2 into yet another (female) sheep (sheep 3). Sheep 3 became pregnant and had a lamb: Dolly.

Some scientists think that within a few years it will be possible to clone people as well. But many governments have already decided to forbid cloning of people by law.

Fig. 13.2 A PISA science item. (Source: OECD (2009). PISA Released Items—Science. Retrieved from https://www.oecd.org/pisa/38709385.pdf)

> many of the nations in these international rankings. To place one's trust in the PISA scholars to have solved a problem so fraught with complexities as equalizing the cognitive load in two different languages … strikes me as naive.

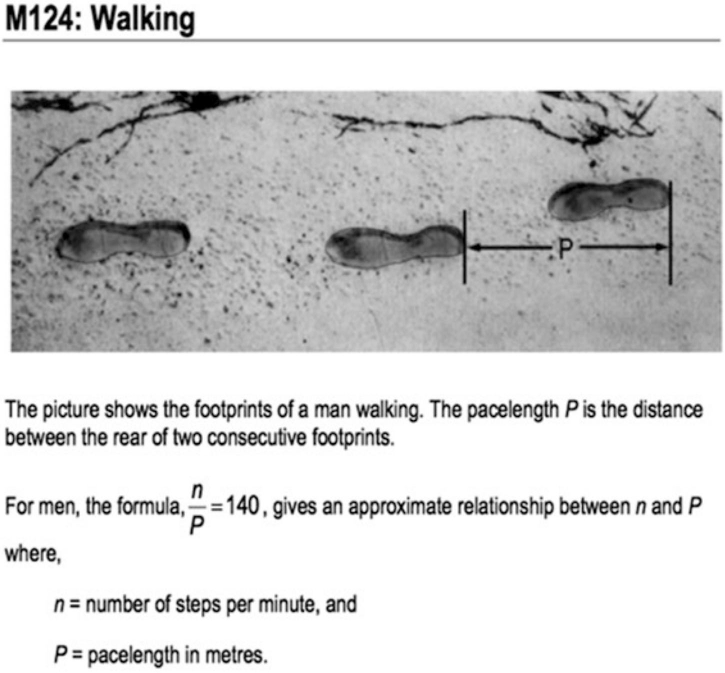
13.1.2 Context and PISA Items

In fact, the common sense about this issue is fully supported by the research of Ruiz-Primo and Li (2015). They point out that items on PISA are intended to tap deeper learning than do multiple-choice items. To do that requires the design of a context, such as that given in Fig. 13.2. The context is to be read and understood before the actual questions designed to tap problem-solving skills are asked. These kinds of questions are in contrast to straightforward multiple-choice questions, designed to assess memory for factual knowledge, and where text providing background contexts are usually not necessary.

Ruiz-Primo and Li believed that these context-dependent questions might be understood differently in different countries, a perfectly reasonable hypothesis that was confirmed. They found differential student performance, *by nation*, on PISA, associated with the contexts in which items were presented. They also found evidence that item contexts across countries affected male and female respondents differently.

13.1.3 Illustrations and PISA Items

We need also to remember that items designed to tap problem-solving, not simple memory, often have illustrations associated with them, as well as contexts. An illustration for a PISA math item is given in Fig. 13.3 (Organization for Economic



M124: Walking

The picture shows the footprints of a man walking. The pacelength *P* is the distance between the rear of two consecutive footprints.

For men, the formula, $\frac{n}{P} = 140$, gives an approximate relationship between *n* and *P*

where,

n = number of steps per minute, and

P = pacelength in metres.

Fig. 13.3 Illustration from which competency in mathematics is assessed on PISA. (Source: OECD (2009). PISA Released Items—Mathematics. Retrieved from https://www.oecd.org/pisa/38709418.pdf)

Cooperation and Development 2009). But Solano-Flores and Wang (2015) discovered that items with illustrations are interpreted differently in different countries.

These investigators say that cultural differences in the interpretation of illustrations *significantly affected the scores obtained by nations*.

Commenting on this particular figure, Sjøberg (2007) says:

> If the marked foot**step** is 80 cm (as suggested in other information that is given), then the foot**print** is 55 cm long! A regular man's foot is actually only about 26 cm long, so the figure is extremely misleading! But even worse: From the figure, we can see (or measure) the next foot**step** to be 60% longer [than the first foot**step**]. Given the formula above, this also implies a more rapid pace, and the man's acceleration from the first to the second foot**step** has to be enormous!

After other criticisms, Sjøberg says:

> The situation is unrealistic and flawed from several points of view. Students who simply insert numbers in the formula without thinking will get it right. More critical students who start thinking will, however, be confused and get in trouble! (see also Sjøberg 2015).

13.1.4 Construct-Irrelevant Variance and PISA Tests

Another criterion for good SATs is to minimize construct-irrelevant variance. If construct-irrelevant variance affects the scores obtained, the interpretation of the test is more difficult, and the inferences that we want to make from the test may be suspect.

Test items should be related to the constructs under investigation, in this case—reading, science, and mathematics knowledge. But we see that in cross-national testing, there are potential sources of construct-irrelevant variance due to differences in languages, and in the interpretations of contexts and illustrations. And how do we avoid irrelevant variance in science assessment when the USA, Libya, and Myanmar continue to reject the metric system? Or, because the reading load is quite heavy when illustrations and contexts are employed, it seems a sure bet that reading ability is a source of construct-irrelevant variance in the assessment of both mathematics and science.

13.1.5 Raw Scores and Imputed Scores Derived from PISA Items

I have little doubt that PISA technicians are skilled. PISA has employed some of the best measurement people in the world. But common sense and research now suggest that even small differences in national raw scores due to small differences in the interpretation of items associated with language, illustrations, contexts, and construct-irrelevant variance, make PISA interpretations quite a bit more problematic than we have been led to believe. Because of these factors, there is certainly a high likelihood of small national differences at the level of individual items. These small differences become magnified when sophisticated statistical models are used to put the national scores into a metric with a mean of about 500 and a standard deviation of about 100, from which the ranks of nations on PISA are determined. As I understand it, total scores are imputed from the characteristics of each of the items passed, and we now know that these items are likely to reflect national differences in language and culture, not simply student achievement. In fact, raw scores among nations hardly differ, while the scaled scores and ranks used in interpreting PISA scores differ quite a bit. This situation arises because of the predictions of total scores from the small sample of items given to each student in a PISA sample (Table 13.1).

As can be seen, nations with the *same raw scores*, say Slovenia and the USA, have scale scores that differ quite dramatically. Slovenia is given the scale score of 501, while the USA's scale score is determined to be 481. This is hard for the common person to understand: The same raw score, but scale scores that differ by 20 points, and producing a difference of 15 ranks! Note also that Finland and Israel differ by 2 raw score points but by 52 scale score points, and by 29 ranks.

If each test form of about 30 items is a purposeful sample of the 109 math items used in 2012, then many nations are performing quite similarly up until the imputation scheme, where the sampling design is used to determine the scale scores and

Table 13.1 PISA 2012 raw scores, scaled scores, and ranks for selected countries

Nations	PISA 2012 Raw Score in Math	PISA Scaled Score in Math	Rank of Nations
Finland	13	519	12
Poland	13	518	14
Vietnam	13	511	17
Austria	13	506	18
Ireland	13	501	20
Slovenia	12	501	21
France	12	496	25
Iceland	12	493	27
Norway	12	490	30
Spain	12	485	33
United States of America	12	481	36
Croatia	11	471	40
Israel	11	467	41

Selected national raw scores, scaled scores, and ranks for PISA 2012 NB: Number of items in mathematics in 2012 = 109; Students in each country take about 30% of these items on different forms of the test. PISA scores for countries are based on the unadministered items, as well as the administered items, with scores on the unadministered items predicted/extrapolated from the items that were administered, using weights based on student samples in each nation ranks for a nation. What we get, a bit magically for most of us, is a lot of imputed (or plausible) scores for a nation's students.

The psychometric procedure, used to determine these scores from the samples of items, but not the whole test, uses the well-known and quite brilliant Rasch model. But as I understand it, this model only works if the questions PISA uses in each country have the same difficulty level. And we have just seen that because of language, the use of illustrations, and the need for descriptive contexts, equality of difficulty across nations is unlikely, and perhaps impossible.

This suggests that we can expect some wide ranges of values in the scale scores determined from the forms that a nation used, and the plausible or imputed values determined from those forms. And that is exactly what we get. For example, according to the 2006 reading rankings, Canada could have been positioned anywhere between the 2nd and 25th ranks, Japan between the 8th and 40th, and the UK between the 14th and 30th (Kreiner 2011). Such variation in scores and the associated ranks suggests that the reliability of PISA scores and rankings is more questionable than consumers of PISA have ever been led to believe (Kreine and Christensen 2014).

13.1.6 PISA Reliability

Standardized Achievement Test designers pride themselves on having high reliability, so that the possibility of valid inferences can be made from the scores obtained. But the PISA designers may not always meet that criterion as well as they might hope to do. For example, in two provinces of Italy (Bratti and Checchi 2013), the opportunity arose to retest students a year later with the same forms of the PISA test that they originally took the previous year. This study was concerned with the value added by the students' schools. They chose to use PISA as the Standardized Achievement Test from which the school's added value would be calculated. In one province, tested in Italian, the year-to-year student scores were quite highly correlated, as might be expected when using a well-designed SAT. But in another province, French speaking, the correlation of the students' scores from year to year was quite low, in fact, near zero! This is not very reassuring! The differences, it seems, were due to different attrition rates over the single year, which meant that in the low reliability district, a slightly different cohort took the test the second time. Since PISA is given every 3 years, and different 15-year-old cohorts are used in each nation, the stability of scores over the 3 years between assessments is quite likely to be less than is desirable for the design of national education policies that depend heavily on reliable trends. The trends derived from these data may, therefore, be quite suspect.

13.1.7 Sampling Issues and PISA Testing

The trustworthiness of the raw scores, especially of the imputed scores, clearly depends on the sampling schemes devised by PISA. That too is not perfect. Loveless (2013) has shown that the extremely high PISA scores obtained by Shanghai in 2012 were obtained, in part, by leaving out tens, if not hundreds of thousands of children of migrants. These migrants are often rural Chinese families without government permission to work in Shanghai. The children of these illegal or undocumented families are not always permitted to go to school, or they may be purged from public school by age 14, just before the samples for the following year's PISA assessments are determined. The sampling errors were well known to PISA, though apparently ignored by them, and ignored as well by newspapers around the world that discussed how good the Shanghai schools appeared to be.

Similarly, and quite convincingly, Carnoy and Rothstein (2013) have identified PISA sampling problems in the USA. For example, the 2009 PISA sample had 40% of the participating American students coming from schools where half or more of the students were eligible for free and reduced lunch programs. But the percent of US students actually in schools with such high rates of poverty is much lower. Carnoy and Rothstein (2013) determined that if the 2009 sample had been correct, the rank of the USA on PISA would have gone from 14th to 6th in reading and from 25th to 13th in mathematics.

13.1.8 Validity and PISA Testing

All SATs depend on convincing evidence of validity to justify both their use and their costs.

To claim *Content Validity* for PISA would require evidence that the PISA assessments of 15-year olds today be related to the real-world tasks that are required of adults in their work and home lives in, say, 10–15 years from now. PISA explicitly seeks assessment tasks that are representative of the skills needed *in the future*, making it impossible to judge PISA on the adequacy of its content validity in the way that we can judge other tests, TIMSS, for example, which attempts to assess contemporary curricula.

A *construct validity* argument would find that scores on PISA, PIRLS, TIMSS, and certain national tests (NAEP and NAPLAN, for example) are moderately or strongly correlated in each of the content areas assessed. There is evidence that this is true, so the construct validity argument can be made, but not as strongly, perhaps, as might be desired. Score and rank order differences that arise as a function of taking these different tests promote the argument that the mathematics, science, and reading knowledge constructs being measured in different nations may be different. This results in difficulty in test score interpretation for a particular test in a particular country. The USA, for example, does extremely well on the PIRLS test of reading. We do quite well on the TIMSS science and mathematics tests. But we do not do as well on PISA. How shall we judge our national level of achievement when these tests of similar constructs yield such different estimates of U. S. achievement?

The *consequential validity* argument has already been alluded to—newspapers and politicians each go mad with the PISA results, either attributing credit to national governments for things they may have had nothing to do with, or blaming institutions and people for results they do not like, even though those institutions and people may not have had much influence on the results. We know that in the USA the variance attributable to teachers and schools from results of virtually all SATs is quite small, compared to the variance attributable to social class, income, neighborhood, educational level of the mother, etc. So, interpreting the results of PISA in ways that laud or condemn teachers and schools makes little sense. While Finland and the USA differ in scores on PISA, they also differ on childhood poverty rates. Finland's childhood poverty rate is about 4%, while the poverty rate for children in the USA is likely to be over 20%. Though politicians and journalist may blame schools and teachers, it is certainly the case that the social systems of the two nations have real effects on PISA scores (Condron 2011). Clearly, in almost all PISA countries, the tests have consequences. Results are attended to in both appropriate and highly inappropriate ways. In the USA, valid inferences drawn from PISA data are a rare experience.

One more type of validity needs to be addressed, *predictive validity*. PISA really is about predicting a nation's fate as a function of the test scores generated by its school systems. The economists Hanushek and Woessmann (2010), along with the OECD and many politicians, make the case that a substantial rise in PISA scores for nations would mean trillions of dollars in increased business activities. Their argument is that as nations set about to improve their curriculum, their schools, and the quality of their teachers, they will soon have higher PISA scores, and that will inevitably make their national economies hum. The data used by Hanushek and Woessmann (2010) to make these oft-repeated claims for predictive validity have been seriously challenged, and now appear to be indefensible (Komatsu and Rappleye 2017; Lauder 2015).

So, PISA is taken quite seriously as an omen; the scores are talismanic objects. But for me, the logic of this is closer to that of the cargo cults of the early twentieth century than the realities associated with modern nation states.

I find at least three things wrong with the economic benefits argument. The first is that Standardized Achievement Tests only weakly show any effects of curriculum, schools, and teachers. Thus, improving these aspects of schooling will meet with very limited success in influencing PISA scores. Teachers and schools simply do not affect the variance in Standardized Achievement Tests very much. Thus, all policies derived from SATs, such as PISA, NAEP, TIMSS and others, that are designed to improve schools *without* improving the economic and social conditions of the children and families in those schools are doomed to deliver tiny benefits!

PISA obtained its magical powers, becoming endowed with predictive validity in part, because, of a search for an index that assessed the potential of our globalized economies (Rizvi and Lingard 2011). It does not do this well at all. For example, based on the results of previous administrations of PISA, at a time when Japan did especially well on PISA, theirs was the economy to watch. We did. Japan's economy failed and after over 10 years of strong PISA scores, it is still failing. On the other hand, the 2000 administration of PISA provided a rude shock for Germany as it garnered a relatively low score. Their economy, however, moved on to become the strongest in the European Union. In 2000, Finland also received a rude shock (Sahlberg 2011). Because of its high PISA scores, it became the fantasy land for Western nations. And although it has fallen off a bit in recent PISA testing, it is still acknowledged as a world leader. But what about the Finnish economy? It has not been doing well for a number of years, despite all of Finland's PISA talent. Not long ago the Finnish Prime Minister said that his high scoring country is in a "lost decade." Their economy has fallen behind its Nordic neighbors and its European peers. Pay increases have been on hold, government debt almost doubled from 2008 to 2014, taxes are up a few percent, and the jobless rate not long ago was about 9%.

And at the same time that Germany worried, and Finland was surprised, the USA and Israel did relatively poorly on PISA. Yet each of their economies has thrived in the years since. The USA, in fact, with modest performance on PISA, has won the distinction of being number one on the 2014, 2016, and 2017 GEDI Index (Acs et al. 2016). The GEDI is the World Global Entrepreneurship and Development Index. While PISA gets attention as a predictor of future economic prosperity, despite no proof that it actually has predictive validity, it might be more reasonable to expect the GEDI to be such a predictor instead. Few US pundits said anything about the release of the GEDI, though they rarely miss a chance to wring their hands over PISA scores. Yet, the GEDI researchers were associated with the Imperial College Business School, the London School of Economics, the University of Pécs, and George Mason University. Researchers at these institutions studied entrepreneurship in 120 nations. They have found consistently that the No. 1 country in the world was the USA. It strikes me that entrepreneurship among adults is much more likely to be a predictor of a nation's future economy than is PISA. This is especially true when that test assesses 15-year-old American kids who know that the test scores count for nothing, and that the results are never seen by their teachers or their parents. Korean youth may take the PISA for the honor of their country. Youth in the U SA take the test because they are ordered to. I would not predict much based on that kind of sample!

Contrary to the despair over PISA scores in the USA, the GEDI authors say:

> Entrepreneurship plays a crucial role in the US economy and as result, policy initiatives are created to encourage entrepreneurial behavior. This, coupled with the culture of determination and motivation, makes the US a great place to be an entrepreneur.

Moreover, the researchers say, the gulf between the United States and other countries is large and appears to be widening, not narrowing. In addition, the 2014 GEDI compared the experience of female entrepreneurs, for the first time, to reflect the increasing participation and importance of women in entrepreneurship around the world. The researchers determined that the USA is also the world's leader in female entrepreneurship.

Furthermore, while the US PISA scores lead many to predict doom and gloom for our economy, there is also the Global Innovation Index. It too appears likely to predict future economic activity for nations better than would a SAT like PISA. The Global Innovation Index is put together each year by two prestigious universities and a UN committee and it uses 82 different metrics to determine rankings. The 2016 index is presented as Table 13.2 (Dutta et al. 2016).

In the first column, I entered the Global Innovation Index ranking for the top 15 nations. As you can see, the USA ranks 4th which is not bad at all. Then I entered the combined ranks of these nations on PISA 2015 and correlated the two measures. The correlation is negative! Pisa seems not to predict innovativeness of a nation. I think, therefore, that claims of predictive validity for PISA remain unsubstantiated.

There is one more issue that is both a concern about reliability and validity for PISA. It is the number of false positives and false negatives that show up at the item level. New Zealand researchers (Harlow and Jones 2004) studied items that students had gotten wrong and right on an international SAT of science and did a version of dynamic testing with the students. They administered the items individually to the students and probed whether the students who got the items wrong really did know

Table 13.2 Global Innovation Index in 2016 and PISA 2015 Science Rank

Country/Economy	Score (0–100)	Rank	Income	PISA2015 Science Rank
Switzerland	66.28	1	HI	18
Sweden	63.57	2	HI	28
United Kingdom	61.93	3	HI	15
United States of America	61.40	4	HI	25
Finland	59.90	5	HI	5
Singapore	59.16	6	HI	1
Ireland	59.03	7	HI	19
Denmark	58.45	8	HI	21
Netherlands	58.29	9	HI	17
Germany	57.94	10	HI	16
Korea, Rep.	57.15	11	HI	11
Luxembourg	57.11	12	HI	33
Iceland	55.99	13	HI	39
Hong Kong, China	55.69	14	HI	9
Canada	54.71	15	HI	7
Japan	54.52	16	HI	2
France	54.04	18	HI	27
Australia	53.07	19	HI	14
Austria	52.65	20	HI	26

Source: Dutta et al. (2016). The Global Innovation Index 2016 Winning with Global Innovation. http://english.gov.cn/r/Pub/GOV/ReceivedContent/Other/2016-08-15/wipo_pub_gii_2016(1).pdf. OECD, PISA 2015 Database, Tables I.2.12a-b, I.3.1a-c and I.3.10a-b the answer if probed a bit. And they probed to see whether the students, who got the items correct, had the knowledge required for a correct answer. Their research revealed that many of these New Zealand test takers were false negatives on the test items, that is, they did know their science but got the item wrong. Many others were false positives; they did not know their science well but got the item right, anyway. More work, like this, needs to be done since these results suggest problems with both the reliability and validity of PISA and other SATs.

13.2 A Conclusion About PISA Quality

PISA is clearly better supported, has better personnel working on it, and its technical characteristics are better than many other Standardized Achievement tests. But like all other SATs, it has faults and is not above criticism. I have criticized the assumption of comparability of the test across nations because of the real possibility of differences in the cognitive complexity of items and item understanding in the different languages and cultures of each nation. It is more than just translation that is of concern. The contexts and illustrations from which PISA items are derived were found *not* to be equivalent across countries. And it is likely that they can never be made equivalent until we all speak Esperanto!

The difference in scaled scores and ranks, associated with identical raw scores among nations, may stem from slight differences in item difficulty by country. And if that is true, then the requirements of the Rasch model are not met and imputation of PISA scores and their associated ranks is seriously flawed.

We now know as well that samples drawn in each nation are not always as accurate a representation of the entire population as they should be, and this too makes the imputation of scores from the samples problematic. Looking at reliability also revealed some rough spots for PISA. That is, sampling procedures and cohort differences from administration to administration make trends much more difficult to trust than is acknowledged. In addition, the predictive validity of PISA in the economic realm appears to be quite overstated. And finally, the rates of false negatives and false positives, at the item level, on one PISA administration, have been found to be of considerable magnitude. Both reliability and validity depend on the magnitude of these occurrences being quite small. But that may not be the case.

Technically, I doubt if any organization can do a better job than PISA in designing a cross-national test that is an appropriate starter of conversations about education. But the national angst, joy, and subsequent policies derived from either low or high scores on PISA assessments are misplaced. Because of its inherent design flaws—*not unlike every other Standardized Achievement Test I know*—PISA results at best might initiate conversations about each nation's visions for childhood, schooling, and economic vitality. PISA should *not* be a catalyst for change without considerable time spent in conversations about one's own national education system in a globalized world. I would limit PISA's influence not just because of the technical problems I have just reported; it will always struggle with those. Rather, I would limit its influence in the USA because PISA’s biggest flaw is unacknowledged, and that is, that the test is picking up the cruel realities of contemporary American policies about income distribution and housing, medical care, jobs, wages, and so forth.

PISA is a Standardized Achievement Test and as such it is a reflection of our society much more than it is a reflection of our curriculum, teachers, schools, and students. That is, the biggest problem with all SATs is the same: In our times, too many inferences about the quality of life in our schools are being drawn, while too few inferences are being drawn from these tests about the quality of life for our families and in our neighborhoods.

13.3 The Limits of PISA and Other SATs in Providing Information for Policies About Teachers and Schools

I will make a bold statement: There are no SATs—neither state, national nor international—whose scores cannot be very well predicted from demographic data. The SATs are notoriously insensitive to teacher and school effects, and powerfully influenced instead by cohort and neighborhood effects, and by family social class, particularly level of poverty. SATs are reflective of sociological variables much more than they are reflective of instructional and educational variables. The evidence for this is overwhelming and avoided by most of those who use PISA data to design policy. Note that what I say suggests that every policy derived from PISA (and other SATs) concerned with the improvement of schools and classrooms is doomed to small effects. This is best described by Haertel (2013), who reviewed the literature on SATs and offers the analysis, as shown in Fig. 13.4. Teachers account for about 10% of the total variance, schools also account for about 10% of the total variance. Error (unexplained variance) accounts for about 20% of the total variance. The

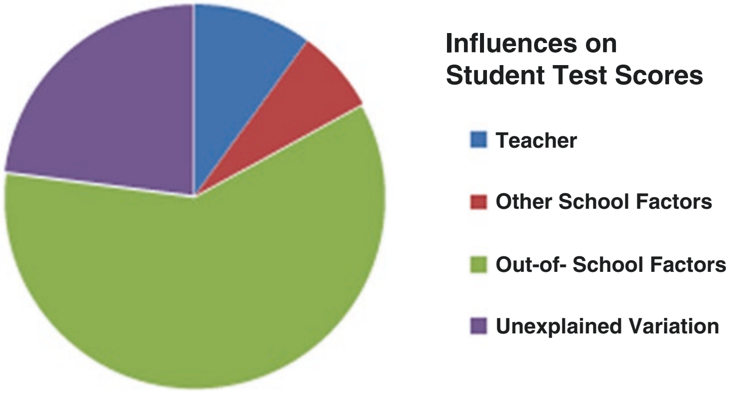


Fig. 13.4 Variance accounted for on SATs.

majority of the variance in scores on SATs, about 60%, is accounted for by out-of-school factors such as family, neighborhood, and income.

Here is the most important point of this figure: Policies designed to affect teachers that are derived from SATs will usually affect only about 10% of the variance we see in students' test scores. And policies designed to affect curriculum, leadership, scheduling, time usage, homework, or other school level factors will also affect only about 10% of the variance in SAT scores. It is the outside of school variables that affect SAT scores the most. I have identified a set of out-of-school factors known to affect the test scores produced by schools (Berliner 2009; Wilkinson and Pickett 2010). These all affect what occurs inside the school and inside the classroom.

- Percent of low birth weight children in the neighborhood
- Inadequate medical, dental, and vision care in family and neighborhood
- Food insecurity in the family
- Environmental pollutants in home and neighborhood
- Family relations and family stress
- Percent of mothers at the school site that are single and/or teens
- Percent of mothers at the school site that do not possess a high school degree (or have not finalized secondary education)
- Language spoken at home
- Family income
- Neighborhood characteristics
    - Rate of violence
    - Drug use
    - Mental health
    - Average income
    - Mobility rates of families
    - Availability of positive role models
    - Availability of high-quality early education
    - Transportation to get to jobs

Concern for these variables, more than PISA test scores, seems much more likely to affect student achievement. Our contemporary thinking about these out-of-school issues begins with the Coleman report (Coleman et al. 1966). That report shocked our democracy 50 years ago as it convincingly argued that teachers and schools were not nearly as powerful as we thought in breaking the cycle of poverty. Although that fact has been understood for a long time now, it is too often ignored by policy makers and researchers alike (Powers et al. 2016).

Borman and Dowling's (2010) re-analysis of the Coleman data, using more modern and more powerful statistical measures, makes two claims that are important for the argument being made here. First, they claim that teacher effects, compared to composition effects, *are a minor predictor of student scores*. Second, they claim that the *peer and compositional effects on achievement test scores* are about twice as strong as is the racial or social class standing of the students themselves. Who you go to school with matters a lot!

Because, in the western world, we almost always live in socioeconomic and racially homogenous enclaves, our schools are often segregated by race and class. In terms of school achievement, the race and class of those individuals matters, but only a little, until those racial and social class characteristics influence the peer, or cohort, or composition of students at the school site. Some peer groups and cohorts promote high achievement, and some do not; but what must be remembered is that the aggregate scores obtained on SATs given to those classrooms and schools are substantially independent of the effects of teachers and schools on those students.

The point is that teachers, who currently get so much blame for the outcomes of our schools, are probably accounting for only about 10% of the variance in those aggregate outcomes. And the schools, frequently the recipients of blame when PISA or other SAT scores are low, also account for about 10% of the variance in SAT outcomes.

The most recent support for this claim is from the American Statistical Association. Their position paper was on value-added models of teacher evaluation (American Statistical Association 2014), in which a pre- and a post-SAT is used to judge the value added to student scores by a particular teacher. They say, "Most VAM studies find that teachers account for about 1–14% of the variability in test scores, and that the majority of opportunities for quality improvement are found in the system-level conditions."

13.4 Conclusions Considering the Limits of SATs for Policies

So, outside-the-school factors are often three times more powerful in affecting SAT scores than are inside-the-school and inside-the-classroom factors; Put differently, outside-the-school factors are six times more powerful than are teachers and six times more powerful than are the schools when the influence on SATs is analyzed. Policies, dealing with teacher and school improvement that are derived from SATs like PISA, can only have limited success.

A few million independent anecdotes about how teachers affect individual students are proof enough of their power to influence individuals. In my life, they made a big difference in what kind of person I became, and some of my teachers also affected the habits of mind I bring to my work and to my personal life. My children have also been positively affected by some of their teachers. And I am sure that readers of this chapter have similar stories to tell, the vast majority of which are about positive effects, although teachers have the power to negatively affect individual children, as well.

This is a paradox and like all paradox's a bit confusing: Teachers and the schools attended by a nations children affect the individual students in their classes enormously: teachers really do touch eternity (Barone 2001). But teachers and schools affect the SATs ordinarily used to judge teachers and schools only a little. PISA is merely a SAT. It measures demographic characteristics quite well and is almost useless for suggesting policies that affect teachers and schools that will affect the scores on those SATs.

13.5 Summing Up

At the start of this chapter, we saw that PISA, like every other SAT, struggles with technical problems, including that most important criterion for any SAT, its meaning! What is PISA valid for? I would argue that PISA is perfectly wonderful for starting conversations about schooling; the outcomes desired for a nation's youth; the curriculum to achieve those outcomes; culture and childrearing practices and their effects on school achievement; income distribution and its effects on youth behavior and school achievement; the design of trust relationships between educators, parents, and policy makers; discussions about whether a metric really can be created for everything that a community wants to assess; and so forth.

The distinguished British comparative educator, Robin Alexander (2012), is likely to agree with a good deal of what I say in this chapter. He has some remarkable insights into the madness that attends to PISA scores because of their inappropriate use. For example, he notes how a team headed by Michael Barber wrote a report for the multibillion dollar management corporation McKinsey & Co., that is inane (my description!). The report, titled *How the World's Best-Performing Education Systems Come Out on Top*, was almost universally praised by policy makers throughout the western world. Its authors concluded from PISA 2003 that "Three things matter most: (1) getting the right people to become teachers, (2) developing them into effective instructors, and (3) ensuring that the system is able to deliver the best possible instruction for every child." (Barber and Mourshed 2007: 2). Well, duh! I might have written such banality well before anyone ever heard of PISA. This really is not high-level thinking, especially given the cost of the report! But in addition, as I have argued above, this report is not merely ordinary in its conclusions: it is also wrong!

Policies aimed at teachers, schools, and school systems will have little effect on national school system achievement as measured by SATs, because SAT scores are reflections of other things—income inequality, housing policy, cohort effects, culture, and so forth (Sjøberg and Schreiner 2005). This expensive and lauded report has no clue about what makes for an SAT score!

Alexander (2012) cites others who also would have found Barber's McKinsey report ridiculous. Ernest Boyer, an influential educator and policy analyst of the 1960s once said: "Schools can rise no higher than the communities that support them" (Boyer 1983: 6). Not long after, in 1970, the well-respected British social scientist Basil Bernstein said, "Education cannot compensate for society." (Bernstein 1970: 344). I would like to end this chapter with the insights of a scholar who works for PISA (Andreaus Schleicher 2009), and one who I believe to be much wiser, though he wrote 115 years earlier (Michael Sadler 1900: 50; see Alexander 2012). Schleicher, after examining PISA data, says the ideal school might have little bits of

Finland, Japan, England, Israel, Norway, Canada, Belgium, and Germany in the way it develops individual strengths of students, gets teachers to cooperate, sets clear performance targets, celebrates discourse and helps students to learn from their mistakes, and so forth. But Sadler says this:

> In studying foreign systems of education we should not forget that the things outside the schools matter even more than the things inside the schools, and govern and interpret the things inside … No other nation, by imitating a little bit of German organization, can thus hope to achieve a true reproduction of the spirit of German institutions … All good and true education is an expression of national life and character … The practical value of studying in a right spirit and with scholarly accuracy the working of foreign systems of education is that it will result in our being better fitted to study and understand our own.

That is what PISA and other SATs are good for. They are capable of providing data for conversations about schooling in each society. PISA has no magic. Its scores are not talismanic. It is a starting place for conversations not for the immediate design of policy (see Sellar et al. 2017).

Policies about teachers and schools, like those promoted by McKinsey and company, are both misleading and useless if they are expected to produce large changes in the scores on an SAT. PISA is simply another SAT, with some added and unusual technical problems, and all the usual insensitivities to teaching and schooling that characterize all standardized achievement tests.

References

Acs, Z., Szerb, L., Autio, E., & Lloyd, A. (2016). *Global entrepreneur ship index, 2017*. Washington, DC: Global Entrepreneurship and Development Institute.

Alexander, R. J. (2012). ‘International evidence, national policy and classroom practice: Questions of judgment, vision and trust’ (Keynote address to the 3rd Van Leer International Conference on Education, Jerusalem, May 2012), *Educational Echoes*, 105–113 (In Hebrew).

American Statistical Association. (2014). *Value-added models to evaluate teachers: A cry for help*. Retrieved from http://chance.amstat.org/2011/02/value-added-models/.

Barber, M., & Mourshed, M. (2007). *How the world’s best-performing schools come out on top*. McKinsey& Co. Retrieved from http://mckinseyonsociety.com/how-the-worlds-best-performing-schools-come-out-on-top/.

Barone, T. (2001). Pragmatizing the imaginary: A response to a fictionalized case study of teaching. *Harvard Educational Review, 71*(4), 734–741.

Berliner, D. C. (2009). Are teachers responsible for low achievement by poor students? *Kappa Delta Pi Record, 46*(1), 18–21.

Bernstein, B. (1970). Education cannot compensate for society. *New Society, 344*.

Borman, G. D., & Dowling, M. (2010). Schools and inequality: A multilevel analysis of Coleman’s equality of educational opportunity data. *Teachers College Record, 112*(5), 1201–1246.

Boyer, E. L. (1983). *High school: A report on secondary education in America*. New York: Harper & Row, Inc.

Bracey, G. (1991). Why can’t they be like we were? *The Phi Delta Kappan, 73*(2), 104–117.

Bratti, M., & Checchi, D. (2013). *Re-testing PISA students one year later. On school value added estimation using PISA—OECD*. Retrieved from www.iwaee.org/papers%20sito%202013/Bratti.pdf.

Carnoy, M., & Rothstein, R. (2013). *All else equal: Are public and private schools different?* London: Taylor and Francis.

Coleman, J. S., et al. (1966). *Equality of education opportunity*. Washington, DC: U.S. Government Printing Office.

Condron, D. J. (2011). Egalitarianism and educational outcomes: Compatible goals for affluent societies. *Educational Researcher, 40*(2), 47–55.

Dutta, S., Lanvin, B., & Wunsch-Vincent, S. (2016). *The Global Innovation Index 2016 Winning with Global Innovation*. Retrieved from http://english.gov.cn/r/Pub/GOV/ReceivedContent/Other/2016-08-15/wipo_pub_gii_2016(1).pdf.

Glass, G. V. (2012). *Among the many things wrong with international achievement comparisons. Education in two worlds*. Retrieved from http://ed2worlds.blogspot.com/2012/02/among-many-things-wrong-with.html.

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement, 11*(1), 1–18.

Haney, W., Madaus, G., & Kreitzer, A. (1987). Charms talismanic: Testing teachers for the improvement of American Education. *Review of Research in Education, 14*(1), 169–238.

Hanushek, E. A., & Woessmann, L. (2010). How much do educational outcomes matter in OECD countries? *Economic Policy, 26*(67), 429–491.

Harlow, A., & Jones, A. (2004). Why students answer times science test items the way they do. *Research in Science Education, 34*(2), 221–238.

Komatsu, H., & Rappleye, J. (2017). A PISA paradox? An alternative theory of learning as a possible solution for variations in PISA scores. *Comparative Education Review, 61*(5), 269–297.

Kreine, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika, 79*(2), 210–231.

Kreiner, S. (2011). A note on item–Rest score Association in Rasch Models. *Psychological Measurement, 35*(7), 557–561.

Lauder, H. (2015). Human capital theory, the power of transnational companies and a political response in relation to education and economic development. *A Journal of Comparative and International Education, 45*(3), 490–493.

Lingard, B., Martino, W., Rezairashti, G., & Sellar, S. (2015). *Globalizing educational accountabilities*. London: Routledge.

Loveless, T. (2013). *Attention OECD-PISA: Your silence on China is wrong*. Washington, DC: Brookings Institute. Retrieved September 1, 2017 from www.brookings.edu/research/attention-oecd-pisa-your-silence-on-china-is-wrong/.

OECD (Organization for Economic Cooperation and Development). (2009). *PISA data analysis manual: SPSS, Second Edition*. Retrieved from https://doi.org/10.1787/9789264056275-en.

Powers, J. M., Fischman, G. E., & Berliner, D. C. (2016). Making the visible invisible: Willful ignorance of poverty and social inequalities in the research-policy nexus. *Review of Research in Education, 40*(3), 744–776.

Rizvi, F., & Lingard, B. (2011). Social equity and the assemblage of values in Australian higher education. *Cambridge Journal of Education, 41*(1), 5–22.

Ruiz-Primo, M. A., & Li, M. (2015). The relationship between item context characteristics and student performance: The case of the 2006 and 2009 PISA science items. *Teachers College Record, 117*(1), 36.

Sadler, M. (1900). How can we learn anything of practical value from the study of foreign systems of education? In J. H. Higginson (Ed.), *Selections from Michael Sadler: Studies in world citizenship* (p. 50). Liverpool: Dejall and Meyorre.

Sahlberg, P. (2011). PISA in Finland: An education miracle or an obstacle to change? *Center for Educational Policy Studies Journal, 1*(3), 119–140.

Schleicher, A. (2009). Securing quality and equity in education: Lessons from PISA. *Prospects, 39*(3), 251–263.

Sellar, S., Thompson, G., & Rutkowski, D. (2017). *The global education race: Taking the measure of PISA and international testing*. Edmonton: Brush Publishing.

Sjøberg, S. (2007). PISA and “Real Life Challenges”: Mission impossible? In S. T. Hopmann, G. Brinek, & M. Retzl (Eds.), *PISA zufolge PISA – PISA According to PISA*. Berlin: LIT Verlag.

Sjøberg, S. (2015). PISA and global educational governance—A critique of the project, its uses and implications. *Eurasia Journal of Mathematics, Science and Technology Education, 11*(1), 111–127.

Sjøberg, S., & Schreiner, C. (2005). Perceptions and images of science and science education: Some simple results from ROSE—A cross-cultural study. In M. Claessens (Ed.), *Communicating European Research 2005: Proceedings of the Conference, Brussels, 14–15 November 2005* (pp. 149–156). Dordrecht: Springer.

Solano-Flores, G., & Wang, C. (2015). Complexity of illustrations in PISA-2009 science items and its relationship to the performance of students from Shanghai-China, the United States, and Mexico. *Teachers College Record, 117*(1), 18.

Wilkinson, R., & Pickett, K. (2010). *The spirit level: Why equality is better for everyone*. London: Penguin.

David C. Berliner is a Regents’ Professor Emeritus in the Mary Lou Fulton Teachers College in the division of educational leadership and policy studies at the Arizona State University. He has taught at the Universities of Arizona and Massachusetts, and Stanford University, as well as at universities in Australia, the Netherlands, Spain, and Switzerland. Professor Berliner is a member of the National Academy of Education, a fellow of the Center for Advanced Study in the Behavioral Sciences, and a past president of both the American Educational Research Association (AERA) and the Division of Educational Psychology of the American Psychological Association (APA). He is the recipient of awards for distinguished contributions from APA, AERA, and the National Education Association (NEA). He received the AERA’s Outstanding Public Communication of Education Research Award in 2016. Professor Berliner has authored more than 200 published articles, technical reports, and book chapters. E-mail: berliner@asu.edu.