Reading Educational Research

How to Avoid Getting
Statistically Snookered

Gerald W. Bracey

Foreword by Jay Mathews
Contents

Foreword by Jay Mathews vii
Introduction xi

1 Data, Their Uses, and Their Abuses 1
2 The Nature of Variables 36
3 Making Inferences, Finding Relationships: Statistical Significance and Correlation Coefficients 68
4 Testing: A Major Source of Data—and Maybe Child Abuse 102

Summing Up 171
Resources 173
Works Cited 175
Index 182
Foreword

When we last heard from Gerald W. Bracey, America’s most acerbic educational psychologist, he was losing a part-time job at George Mason University because, it seemed to me, the school couldn’t take the heat that often follows Bracey in his scholarly travels.

I suggested in that column that the annoyance of getting fired would not slow Jerry Bracey down, and I was right. I have just read an advance copy of his latest book, *Reading Educational Research: How to Avoid Getting Statistically Snookered*. It is suitable revenge against his many tormentors, and for people like me still trying to figure out how to make schools better, a must read.

As a popular writer and speaker, with regular columns in two monthly education magazines, the *Phi Delta Kappan* and *Principal Leadership*, and acidic annual reports on the condition of public education, Bracey has been exposing statistics abuse for years. But I have never seen him put together all that he knows as well as he has in this book. It has some of the best explanations of educational numbers manipulation I have ever read, particularly on issues like SAT scores, year-to-year school comparisons, and argument by graph that are most likely to deceive us innocents. The book has Bracey’s deft prose and sure touch with clarifying examples. I also appreciate the fact he trimmed much of his sharp ideological edge, loved by many of his fans, but not by me. He acknowledges several times that no combatant in the bitter education policy wars has an unquestionable grasp on the truth.
(Potential bias alert: Bracey mentions me twice in the book, and unlike his usual treatment of journalists, he does not gut me like a freshly landed trout. On one page he says no larger lessons can be drawn from a story I wrote about a D.C. family using educational vouchers, and on another he says my view of Advanced Placement programs shows that the same AP statistic can have different meanings.)

Here is a good example of the Bracey passion for clarity. He is addressing the difficult concept of correlation, a key to many misunderstandings of educational statistics and to most bad education stories, including some written by me: “We can correlate any two variables. Whether or not the resulting correlation makes sense is another question. Before everyone started wearing jeans, the Dow Jones stock market index correlated with skirt length. Shorter skirts were associated with good economic times and a rising market. Longer skirts were correlated with recessions. To the best of my knowledge, no one suggested raising hemlines as a means to boost the stock market. Similarly, there is a correlation between arm length and shirtsleeve length. Given only a correlation coefficient, though, it makes as much sense to think that increasing sleeve length will make arms grow longer as it does to think that longer arms will mean longer sleeves. In this case other information could be adduced to assist in determining which way the causal relationship would operate.”

Here he is guiding the reader along the twisted path, with charts and other visual aids, that leads to understanding the difference between standard and scaled test scores and how to create the IQ scale: “Now you may be perplexed because I’ve shown the standard scores running from −3 to +3 and they don’t look anything like the SAT that goes from 200 to 800 or an IQ test that would run from 55 to 145. But it’s easy to get from where we are, −3 to +3, to either of these other oft-used scales. Watch closely. I take each standard score, multiply by fifteen, and add one hundred.”

Bracey is a prolific and aggressive critic of No Child Left Behind and the rising use of standardized tests to assess schools and students, but he is too careful an analyst to embrace the most popular alternatives to testing without also giving them the third degree. One favorite of the antitesting movement, portfolios (samples of student work), is seen by Bracey as just another idea with problems. So you have a nice big portfolio envelope, Bracey says. What do you put in it? “Typical work or the best work?” Bracey asks. “Who decides what is best? Teacher or student?” What do you do, he asks, when teachers disagree about the quality of the work?
My favorite part of the book is his look at the National Assessment of Educational Progress (NAEP, rhymes with tape), sometime referred to as the nation’s report card. It is a standardized test given to samplings of children across the country to determine how well American students are doing in math, reading, and other subjects. Its importance has grown as the federal government has gotten into the school-rating business, and some experts have suggested using NAEP, or something like it, to test all U.S. students.

Bracey says that when NAEP was invented in the late 1960s “virtually every education organization in the country rose up in opposition.” Some might say those groups were clairvoyant, because they feared a national test would lead to calls for a national curriculum, which is pretty much what has happened. (Some people, like me, are happy about that.)

The instigators of NAEP, U.S. Commissioner of Education Francis Keppel, and legendary educational researcher Ralph Tyler, wanted to keep the new test simple. They just wanted to describe what students knew, and didn’t know. But in the late 1980s policymakers changed NAEP’s mission to finding how much students knew of what they ought to know. Experts decided what would constitute proficiency in math, English, and the other top-ics, setting a level Bracey and many others think is too high for some grade levels. He quotes a National Academy of Science’s report calling the NAEP achievement level setting procedures “fundamentally flawed.”

This is one of the few places in the book where Bracey cannot resist making a political point. He suggests that making NAEP the definition of proficiency for the nation could be used to make public schools look bad so our education system could be privatized by right-wing zealots. That doesn’t make sense to me, but Bracey is correct in saying that federal education officials of both parties have been consistently bleak about the state of U.S. schools, even in the face of some evidence to the contrary. This may have more to do with bureaucrats seeking money and power from Congress than it does with any master plot against public education, but it is a valid point, and Bracey has made it better and more often than anyone.

Don’t look here for cures for our test-obsessed culture. Bracey makes a more modest case for making tests more reliable, more understandable, and less likely to be taken as the final word on how your child’s school is doing.

His best suggestion is adding courses in what he calls “consumer-oriented probability and statistics” to our curriculums. The first high school principal who offers Bracey a chance to teach such a course will get my vote for administrator of the year, or at least a medal for valor. Bracey is hard to
handle, but his course would be great. For those of us beyond school age, this new book also provides that welcome exposure to Professor Bracey and the many things we still need to know about measuring what is happening to our kids.

Jay Mathews

This originally appeared as an online column on the Washington Post’s website, Dec. 20, 2005.
In seventeenth-century England and France, some people took to collecting numbers they thought reflected the health of the state—births, deaths, marriages, and so on. They came to be called *statists* and the numbers that they collected were first called *political arithmetic* and then *statistics*. State, statists, statistics, an easy progression. Shortly after the creation of statistics, Benjamin Disraeli, prime minister of England during part of Queen Victoria’s reign, declared, “There are three kinds of lies: lies, damned lies, and statistics.” Statistics have had to defend their honor ever since.

The original name, political arithmetic, is telling. People collect statistics with a purpose. Sometimes the purpose is to advance the common good; sometimes it’s to simply advance the people who collected them or their agenda. Addressing an aide, Winston Churchill said, “The first lesson you must learn is that when I ask you for statistics on infant mortality, I want statistics that prove that fewer infants died when I was Prime Minister than when anyone else was Prime Minister” (Wainer and Koretz 2003, 45).

Misleading statistics abound. Some do little harm. “Minutes from Olde Town,” read a lot of real estate ads here in Alexandria, Virginia. Well, my previous homes in Bloomington, Indiana, Littleton, Colorado, and Palo Alto, California, are also minutes from Olde Town. Lotsa minutes, but minutes

1. In America, this statement is often attributed to Mark Twain. Twain certainly popularized it, but he gave Disraeli credit for coining it.

---

Introduction
nevertheless. *Minutes* does not serve as the best unit to describe the distance of those other places, and it might be misleading in some of the real estate ads as well. Similarly, many ads ask us to purchase something for “just pennies a day.” Well, any dollar amount can be rendered in terms of pennies a day. Lotsa pennies, but pennies nevertheless. Readers should be careful to determine if the unit used in any research they read—ethnic groups, socioeconomic status groups, schools, districts, states, nations, whatever—serves as the best unit for understanding what is going on.

An ad in the June 24, 2005, *Mount Vernon Gazette* told me, “You won’t believe that this stunning house is under a million dollars.” Asking price? $999,999. Really. My wife wanted to call and ask if they’d take $999,997 (given the real estate market here, probably not). Darrell Huff would call this gimmick “Much Ado About Practically Nothing.” That’s the title of a chapter in his 1954 classic, *How to Lie with Statistics*. That the book is still in print a half century later might tell us something about the state of the nation’s statistical sophistication.

Huff opens the chapter with an example of two children, one with a tested IQ of 98, the other with an IQ of 101. The average score on IQ tests is 100. We could conclude that one child has a mental capacity below average, the other above average. Huff has a word for a conclusion like this: “nonsense.” Actually, two words: “sheer nonsense.” If you retested the children, this small difference might disappear entirely or reverse direction, with the IQ 98 child scoring above 100. And even if the IQs stayed the same, three IQ points is, well, three IQ points and nothing significant in either the statistical or the practical sense of that word (the difference between practical and statistical significance is explained on pages 68–71).

Many statistics that appear on the national scene are intended to shock and convince and are not so harmless. The Centers for Disease Control and Prevention offer statistics to show that we currently have an “obesity epidemic” in this nation. *New York Times* columnist Paul Krugman wrote that *Health Affairs* reported the “extra costs associated with caring for the obese rose from 2 percent of all private insurance spending in 1987 to 11.6 percent in 2002” (2005b, A17). The documentary *Super Size Me* observed the bodily changes in director-star Morgan Spurlock as he ate three meals a day at McDonald’s for a month. Those changes are themselves presented as statistics—weight, strength, liver function, blood pressure, and so on.

Some argue that the obesity epidemic is just a moneymaking scheme by the pharmaceutical industry. By convincing medical groups to lower
numbers that define such conditions as obesity, high cholesterol, diabetes, and hypertension, they create many more “patients” for whom doctors prescribe the industry’s expensive drugs. On July 4, 2005, the Center for Consumer Freedom (sponsored by Coca-Cola, Wendy’s, and Tyson Foods, among others) published a “Declaration of Food Independence,” which began “When in the course of human events it becomes necessary for freedom-loving people to dissolve the bands which bind them to the Food Cops . . .” (Center for Consumer Freedom 2005).

The center dismisses Spurlock, saying no average person would eat ninety meals in a row in one restaurant. (If I were Spurlock, I’d try the same regimen in a soup-and-salad restaurant.) The center also offers up Don Gorske as a counterexample, a man in the Guinness book of records for having eaten more than nineteen thousand Big Macs, but who weighs in at a muscular 180 pounds and carries a cholesterol level of 155. The center does not mention that Gorske (a) consumed all of those Big Macs over a period of thirty years, (b) does not eat either breakfast or lunch, and (c) has a physically demanding job (Spurlock interviews Gorske in the movie).

The dueling statistics could be pretty funny if the issue were not so serious. I have elaborated on the fat wars for a reason: In many situations an advocate for a policy or program serves up one statistic as the sole and sufficient evidence. In reality, the situation is almost always more complex and more subtle and, to use a word of the moment, nuanced.

For instance, on consecutive days, both David Broder (2005) and William Raspberry (2005) used their columns to call on the Bush administration to abandon vague promises about the war in Iraq and provide what Broder called “Metrics of Success”—statistics—that would go beyond a “fatuous formulation such as ‘We’ll stay until the job is done’” (A25). Both listed many specific economic, security, and military indicators—none of which would alone be sufficient as the sole measure—that should be part of any list of metrics of success (neither columnist, thus far, has been as forthcoming in declaring that would-be reformers should also use metrics of success other than test scores to evaluate schools, teachers, and kids).

America’s principal statistics problem is, as Joel Best put it in Damned Lies and Statistics (2001), “Most of the time, most people simply accept statistics without question” (4; emphasis in the original). One of the reasons we do so is that we are often in no position to do otherwise. We lack the information to challenge the figures. This creates an especially dangerous situation when we’re dealing with what Best calls “mutant statistics”—statistics that
began life as legitimate numbers but suffered mutation by someone or some group. Best observes that in the 1980s a report indicated that some 150,000 women in this country suffered from anorexia. Another study reported that anorexia can lead to death. The mutant—and widely circulated—statistic: 150,000 American women a year die of anorexia.

How many women die of what causes each year is not something most of us keep readily available in memory. Still, says Best, people writing the reports and accepting the statistic should have known better. A simple check of the *Statistical Abstract of the United States*, an enormous annual compendium of statistics from the Census Bureau, would have shown that each year about 8,500 women aged fifteen to twenty-four died of all causes, as did 47,000 women aged twenty-five to forty-four. But anorexia is typically a disorder in young women. Thus, if only 55,500 women aged fifteen to forty-five died each year of all causes, what were the chances that 150,000 women of all ages died of anorexia?

The good news today is that technology now permits many more of us to track down mutant or questionable statistics. Recently, someone mentioned offhand on an Internet listserv that in the original Cinderella story, the slipper was made of fur. In less than five minutes several members of the email list had definitively refuted that claim with information found using the Google search engine. There are, what, dozens, perhaps hundreds, of websites devoted to debunking phony data and urban legends.

Another problem, though, is that people accept many statistics because they lack the tools to think about them critically. Google and the Net cannot offer much if you don’t know how to ask the needed question. The hope of this book is to help develop such tools in the realm of education, to help people become smarter consumers of research and statistics.

In spite of all the new technology, consuming statistics smartly might be harder now than it was when Huff was writing. Huff primarily concerned himself with statistics tossed around by Madison Avenue, then a synonym for the world of advertising. Advertising is a profession of deception, a world committed to making trivial product differences seem large and important and to convincing you that unless you buy a particular product you are incomplete as a human being.

Public schools in Huff’s day were under attack, to be sure, and the reforms concerned curriculum, but there didn’t seem to be enormous ideological wars over such divisions as whole language versus phonics or basic math versus rain forest math (although there was at least a spat about look-say beginning reading instruction versus phonics). Huff’s book appeared the
year after Arthur Bestor’s influential *Educational Wastelands: The Retreat from Learning in Our Public Schools*.

Rudolph Flesch hadn’t yet written *Why Johnny Can’t Read*, nor had Milton Friedman penned *Capitalism and Freedom*, which contained the first contemporary call for school vouchers. Today, alas, there are many “studies” out there that qualify only as pseudoresearch, articles designed to provide something of a scholarly façade for an ideological message. Writing about this subject, Stephen Hegarty of the *St. Petersburg Times* said, “[Jay P.] Greene denies that he has become a ‘Researcher to the Right.’ Yet his work undeniably provides hard data and scholarly cover to policies driven largely by ideology” (2003, 1). I would dispute the “hard data” phrase—a lot of that researcher’s data squish when pressed—but the rest of the comment is dead-on.

The May 18 issue of *Education Week* (Viadero 2005) carried a front-page piece on the debate over whether or not researchers should even release to the press studies that have not yet been peer reviewed (1). Such a practice is taboo in the medical field and should be in education, too. But it isn’t. Peer review is not a foolproof process and it has sometimes been attacked as buddies approving buddies. My experience, though, is that it is more likely to be people in a field reviewing research in the same or a similar field. For peer-reviewed journals, authors are anonymous to the reviewers and the reviewers are anonymous to the authors; I have seldom been able to manage a guess about authorship of manuscripts sent to me by journal editors. Peer review assures a reader that experts in the field have found the methods sound and the conclusions reasonable.

Economist Paul Krugman traces the rise of phony research to Irving Kristol, a conservative thinker of the twentieth century:

In 1978, Mr. Kristol urged corporations to make “philanthropic contributions to scholars and institutions who are likely to advocate preservation of a strong private sector.” That was delicately worded, but the clear implication was that corporations that didn’t like the results of academic research, however valid, should support people willing to say something more to their liking.

Corporations followed his lead, pouring a steady stream of money into think tanks that created a sort of parallel universe, a world of “scholars” whose careers are based on toeing an ideological line, rather than on doing research that stands up to scrutiny by their peers.

There are several reasons why fake research is so effective. One is that nonscientists sometimes find it hard to tell the difference
between research and advocacy—if it’s got numbers and charts in it, doesn’t that make it science? (2005a, A15)

Krugman considers efforts to discount global warming (“lavishly” funded by the energy industry, especially ExxonMobil) and to promote intelligent design as examples of this phony research. In education, readers should beware of “research” emanating from the Hoover Institution at Stanford University, the Heritage Foundation, the Manhattan Institute, the Heartland Institute, the Mackinac Center, the Center for Education Reform, the Thomas B. Fordham Foundation, the American Enterprise Institute, the Paul Peterson group at Harvard, and, soon, the Department of Education Reform at the University of Arkansas. Arkansas is home to the Walton family, and much Wal-Mart money has already made its way to the University of Arkansas, $300 million in 2002 alone. The new department, to be headed by Jay P. Greene, currently at the Manhattan Institute, will no doubt benefit from the Walton presence. The family’s largesse was estimated to approach $1 billion per year (Hopkins 2004), and before his death in an airplane crash, John Walton was perhaps the nation’s most energetic advocate of school vouchers.*

All of the above organizations produce literature whose “data” supposedly support the destruction of the public schools and the establishment of a free-market school system. The pros and cons of this debate are beyond the scope of this book, but the evidence to date does not support contentions that a free-market system would improve education. The National Center for the Study of Privatization in Education operates pretty true to its name. It is not a propaganda factory. Two of its fellows, though, surveyed the various privatization efforts of the 1990s and concluded that there was little evidence that they did work or might work in the future. “Disadvantaged students may not benefit from a free market of choice” (Belfield and d’Entremont 2005).

The media don’t help much. By convention, they present, at best, “balanced” articles, not critical investigative pieces. In a 2004 Wisconsin speech, Robert F. Kennedy Jr. called the White House press corps “a pack of

2. It is interesting to contrast the Waltons’ philanthropy with that of the Bill and Melinda Gates Foundation. Gates has backed the establishment of small public schools in New York City and recently put forth $40 million to establish seventy schools modeled after the innovative and hugely successful Met School in Providence, Rhode Island. The Met is a public school with no curriculum, no courses, and no tests that serves mostly poor students, almost all of whom graduate and go on to college.
stenographers.” Krugman claims that if Bush said the earth was flat, headlines would read “Opinions Differ on Shape of the Earth.” More generally Washington Post ombudsman Richard Harwood revealed that 70 to 90 percent of journalists’ content comes from what people in positions of authority tell them to write (1994, A29).

So you need some savvy about statistics and what they do and don’t, can and can’t tell you. For instance, one unreviewed study that was released to the press and public claimed, “Because these results are statistically significant we can be very confident that the charter schools in our study did have a positive effect on test scores” (Greene, Forster, and Winters 2003). My hope is that by the end of this book you’ll be able to determine that this statement, no matter what the ideology of the people who made it, might very well be false.

You need some language savvy, too. One of these days I’m going to write a book on the language of propaganda. For now, suffice to say that language that seeks to persuade, not merely inform, is suspect in a research report. The standards for such language are lower now than when I attended graduate school forty years ago. In my first report in my first year I presented some data and asserted that they represented “an interesting finding.” The report came back from the overseeing professor with the word interesting struck through in red pencil and this note: “The reader will decide if it’s interesting.”

That might be overkill, but you get the point. Nowadays, researchers commonly call attention to results they find interesting. Given the blizzard of data from research, that eases the job of readers, but it can still mislead. Language that seeks to make up your mind for you or to send your mind thinking in a certain direction is not the language of research. Certainly terms such as Food Cops, dietary despotism, and a growing cabal of menu meddlers, all of which appear in the “Declaration of Food Independence,” do not impress one with their desire to reveal the unvarnished truth.

“Politics and the English Language,” penned in 1946 by George Orwell, two years before he gave us the doublespeak and newspeak of 1984, remains one of the finest essays on language savvy. Decrying “modern” usage, Orwell dismisses speakers of such as automatons:

3. On a whim, I entered “language of propaganda” into Google and got 504 items, some of them quite interesting. A lot of the first 50 or so entries referred to contemporary politics, the Holocaust, or 1984.
When one watches some tired hack on the platform mechanically repeating the familiar phrases—bestial, atrocities, iron heel, bloodstained, tyranny, free peoples of the world, stand shoulder to shoulder—one often has a curious feeling that one is not watching a live human being but some kind of dummy. . . . A speaker who uses that kind of phraseology has gone some distance toward turning himself into a machine. The appropriate noises are coming out of his larynx, but his brain is not involved, as it would be if he were choosing the words for himself. (87)

Orwell developed a few rules for brain-involved speech that are worth repeating even in a book about research and statistics because the rhetoric surrounding statistics is often misleading.

1. Never use a metaphor, simile, or other figure of speech which you are used to seeing in print.

2. Never use a long word where a short one will do.

3. If it is possible to cut a word out, always cut it out.

4. Never use the passive where you can use the active.

5. Never use a foreign phrase, a scientific word, or a jargon word if you can think of an everyday English equivalent.

6. Break any of these rules sooner than say anything outright barbarous. (Orwell 1950)

In this book, the text, which must because of its subject use some scientific and technical words, sometimes leads up to a culminating point, which is then set off as a “principle of data interpretation.” The entire list of such principles follows this introduction as a means of orienting the reader to the book and its contents.

This book is about data, and most of the data about schools these days comes to us in terms of test results, so the book contains an extensive section on testing. The testing section differs a bit from the rest of the book in that it provides some conceptual understanding about testing per se as well as some information about the statistics emanating from tests.

And at the end, the book provides a list of resources for exploring in depth anything mentioned on these pages. Hopefully, many of the works cited in the text will serve that function as well.
1. Do the arithmetic.
2. Show me the data!
3. Look for and beware of selectivity in the data.
4. When comparing groups, make sure the groups are comparable.
5. Be sure the rhetoric and the numbers match.
6. Beware of convenient claims that, whatever the calamity, public schools are to blame.
8. Make certain you know what statistic is being used when someone is talking about the “average.”
9. Be aware of whether you are dealing with rates or numbers. Similarly, be aware of whether you are dealing with rates or scores.
10. When comparing either rates or scores over time, make sure the groups remain comparable as the years go by.
11. Be aware of whether you are dealing with ranks or scores.
12. Watch out for Simpson’s paradox.
13. Do not confuse statistical significance and practical significance.
14. Make no causal inferences from correlation coefficients.
15. Any two variables can be correlated. The resultant correlation coefficient might or might not be meaningful.
16. Learn to “see through” graphs to determine what information they actually contain.
17. Make certain that any test aligned with a standard comprehensively tests the material called for by the standard.
18. On a norm-referenced test, nationally, 50 percent of students are below average, by definition.
19. A norm-referenced standardized achievement test must test only material that all children have had an opportunity to learn.

20. Standardized norm-referenced tests will ignore and obscure anything that is unique about a school.

21. Scores from standardized tests are meaningful only to the extent that we know that all children have had a chance to learn the material which the test tests.

22. Any attempt to set a passing score or a cut score on a test will be arbitrary. Ensure that it is arbitrary in the sense of arbitration, not in the sense of being capricious.

23. If a situation really is as alleged, ask, “So what?”

24. Achievement and ability tests differ mostly in what we know about how students learned the tested skills.

25. Rising test scores do not necessarily mean rising achievement.

26. The law of WYTIWYG applies: What you test is what you get.

27. Any tests offered by a publisher should present adequate evidence of both reliability and validity.

28. Make certain that descriptions of data do not include improper statements about the type of scale being used, for example, “The gain in math is twice as large as the gain in reading.”

29. Do not use a test for a purpose other than the one it was designed for without taking care to ensure it is appropriate for the other purpose.

30. Do not make important decisions about individuals or groups on the basis of a single test.

31. In analyzing test results, make certain that no students were improperly excluded from the testing.

32. In evaluating a testing program, look for negative or positive outcomes that are not part of the program. For example, are subjects not tested being neglected? Are scores on other tests showing gains or losses?
Thank you for sampling this resource.

For more information or to purchase, please visit Heinemann by clicking the link below:


Use of this material is solely for individual, noncommercial use and is for informational purposes only.