14 Conversations About Three Things

Howard Wainer

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2010 35: 5
DOI: 10.3102/1076998609355124

Invited Article

Journal of Educational and Behavioral Statistics
February 2010, Vol. 35, No. 1, pp. 5–25
DOI: 10.3102/1076998609355124

14 Conversations About Three Things

Howard Wainer
National Board of Medical Examiners

In this essay, the author tries to look forward into the 21st century to divine three things: (i) What skills will researchers in the future need to solve the most pressing problems? (ii) What are some of the most likely candidates to be those problems? and (iii) What are some current areas of research that seem mined out and should not distract us from working on more critical areas? In the course of this discussion, the author identifies 14 different things that fall into one of these areas.

Keywords: *technical skills needed; future problems; worn out topics*

I say not that it is, but that it seems to be:
As it now seems to me to seem to be.

—Hubert N. Alyea

I. Introduction

All independent researchers are faced with three questions: (1) What areas will I investigate? (2) How will I go about doing the investigations? and (3) What tools/skills will I need to carry them out? If a poor choice is made in question (1), it does not matter how cleverly question (2) is addressed nor how thoroughly we prepare to satisfy question (3). However, despite this, most of graduate training is concerned with the second and third questions, and approaches to dealing with the first one, too often are left implicit. In this essay, I will nonetheless follow this same form, focusing initially on the knowledge and skills I feel a researcher in the near and midterm future must have. But then I will go on to describe my current view of what seem to be worthwhile areas to investigate and then conclude with some areas whose time may be past and are perhaps best left alone. First, I try to explain how to pick an area for profitable investigation or at least how I pick them; the choices themselves exemplify the approach.

The unsolved problems of today form the basis of the problems of the future, and so any assessment of what the future will bring must be rooted in a careful examination of current practice. Seventeen years ago (Wainer, 1993), I used this

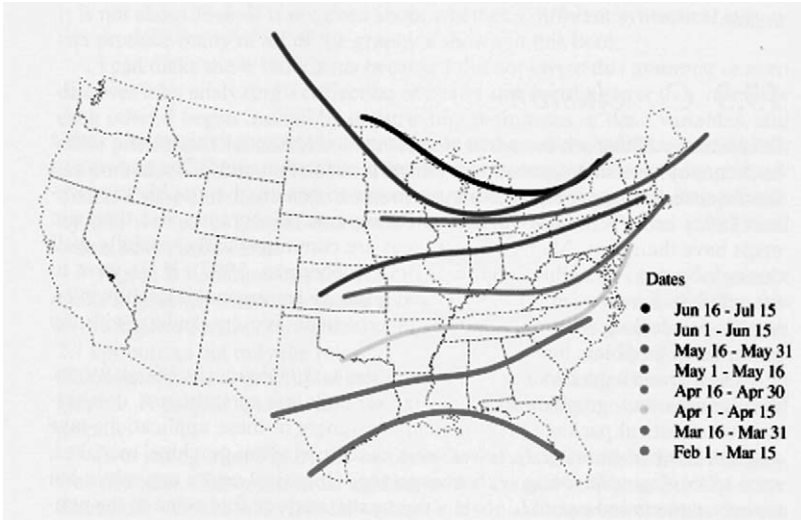

FIGURE 1. *The northern migration pattern of Monarch butterflies.*

technique (described a century ago by David Hilbert, 1902) to lay out what I thought would be the problems of the 21st century (an unexpurgated version was published 2 years later; Wainer, 1995). Today with almost two decades of additional experience, and perhaps a little more wisdom, let me try an update.

Much of current measurement is affected intensely by modern technology. The computations for standard statistical methods such as factor analysis and item response theory (IRT) are done trivially, whereas they once posed serious commitments of time and resources. Indeed some methods (e.g., the bi-factor method of Holzinger & Swineford, 1937) were developed as shortcuts, because what was thought to be the preferred method was computationally impractical. The computationally intensive maximum likelihood approach to factor analysis, made famous by Jöreskog (1969), was first laid out by Lawley (1943). Indeed, today's ubiquitous Monte Carlo procedures were birthed in 1947 in Stanislaw Ulam's sick bed, long before computing facilities were available that would make it practical. Thus, it would not be surprising if profitable future paths for research involve the heavy use of computation. However, modern technology is not confined to just faster computing; it also provides remarkable opportunities for data gathering that were either impossible or impractical even a decade ago.

For example, consider the data shown in Figure 1 (from Wilkinson, 1999), which represent the migration of the Monarch butterfly. Each line connects the location of where Monarch butterflies were seen on the date indicated. In the past, the data gathering effort to produce such a result would require an army of scientists spread across the country for months. This was done using school children and the Internet.

It is clear that profitable future research must anticipate access to data that could only have been dreamed of in even the recent past.

With this brief introduction, let me begin my list of topics with, first, those tools researchers must have to be successful in tackling the problems that loom in the near and more distant future. Next, I try my hand at picking those problems in the greatest need of improved solution. And finally, I conclude with some research areas that seem, to me at least, to be mined out. I will, however, illustrate how sometimes moribund areas are reactivated through new needs.

II. Six Necessary Tools

1. Bayesian methods—Since the 1764 posthumous publication of Thomas Bayes's famous essay, the logic of Bayesian inference has been attractive to anyone who must deal with uncertainty. However, all but the most dedicated were deterred by the practical difficulties involved in implementing it on problems of reasonable scale. These have now been largely overcome and anyone can, in Jimmie Savage's famous words, "eat the Bayesian omelet: if they are but willing to break the Bayesian egg" (Savage, 1954). Bayesian methods allow us to do easily what would be hard otherwise (see Wang, Bradlow, Wainer, & Muller, 2008, for a compelling example). Facility with them is a must for anyone who intends to make contributions to measurement in the future. And so, if the concepts associated with such terms as conjugate prior, jumping kernel, inverse-gamma distribution, and the Metropolis-Hastings algorithm are not close to your soul, get busy. You can learn a little about it in chapter 15 of our testlet book (Wainer, Bradlow, & Wang, 2007) and the references therein, especially Gelman, Carlin, Stern, and Rubin (2003) and Gilks, Richardson, and Spiegelhalter (1996).

2. Causal inference—For centuries, philosophers pondered the difficulties involved in finding the causes of effects. However, statisticians had more success by focusing on measuring the effects of causes. It is this latter problem that is the goal of much of modern science, and its understanding provides important insights into many contemporary problems. Although the start of this modern view goes back at least to Jerzy Neyman (1923; Fisher, 1935, too was very close when he described the importance of randomized experiments as "the reasoned basis for inference"), its current manifestation is due to Don Rubin (Rubin, 1974). Perhaps, the most accessible, fully developed, current description is Paul Holland's (1986) essay, which should be on every scientist's "must read" list. Harold Gulliksen was fond of Kurt Lewin's (1951, p. 169) observation that there is nothing so practical as a good theory. And Rubin's model for causal inference is a very good theory indeed. See how Holland and Rubin (1983) use it to disentangle a paradox that had confounded statisticians since Lord proposed it in 1967.

More recently, it was used by Rubin, Stuart, and Zanutto (2004) to lay out what is perhaps the most difficult challenge facing those who would use value-added modeling to assess the contributions of teachers to their students' performance.

3. Missing data—Dealing with missing data is, quite simply, the most important practical problem facing researchers. A shining example is a study done by Dunn, Kadane, and Garrow (2003), who asked the deceptively simple question, "How much damage is done to student performance by absenteeism?" Their design was simple: give tests at the beginning and end of the school year and measure the relationship of the change in score to the number of absences. We would expect, as indeed they found, that the more classes missed the smaller the gain. However, before they could do the obvious analysis, they had to deal with the inevitable missing data. Some students missed the pretest, some the posttest, some students' records were confused and so absenteeism was hard to calculate. The missingness had many causes—entering or leaving in midyear, illness, and many other possibilities, some not benign. How the analysis accommodates the missingness can have profound consequences. For example, if missing scores were dealt with by imputing the average score of the students who had them (a common approach), it is easy to imagine how a scheming school administrator might game a system that rated a school by its average student gain. He could have a field trip for the best students at the beginning of the school year on the day that the pretest was given. Their imputed scores would then be low. Subsequently, he could have a parallel field trip for the less accomplished students in the spring, when the posttest was administered. Their imputed scores would be high. The estimated gains would then be overestimates, with the extent of the bias determined by the variance of test scores at the school and the number of children who went on the field trips. Note that the gaming was made possible by the method chosen to deal with the missing data.

We can profitably think of many situations in terms of missing data. For example, in regression, we can think of the independent variables x and dependent variable y as observed and the regression coefficients β as missing—and it is our job to estimate them. This way of thinking about regression provides a real insight when we think about such procedures as factor analysis in which we observe y and the xs and βs are missing. This conception makes it clear why we must make such strong assumptions to obtain a solution (and why solutions are not unique).

Causal inferences are also profitably thought of as missing data problems. For example, to assess the causal effect of the treatment x on the outcome y as compared to the outcome y^* under the control condition x^*, we merely subtract; $y - y^*$ is the causal effect. However, we can never observe both y and y^* on the same subject. We must treat one as missing. How we estimate the value of the missing observation is obviously crucial to our estimate of the causal effect. Randomization is a powerful tool for allowing us to make credible assumptions about the

average values of both y and y^*, hence the importance of randomized experiments. Yet every randomized experiment is an observational study waiting to happen. Consider one plausible source of placebo effects: we run a randomized experiment in which one group gets a drug and the other a placebo. Suppose that the drug, on average, has a small positive effect and the placebo has none. However, there is variability. Subjects who receive the placebo and who do not feel better have a larger likelihood of dropping out of the study. The ones who remain showed a more positive result. The placebo effect, as described here, is merely a consequence of the nonignorable missingness.

The only thing we know for sure about a missing data point is that it is not there, and there is nothing that the magic of statistics can do to change that. The best that can be managed is to estimate the extent to which the missing data have influenced the inferences we wish to draw. We accomplish this by using methods that measure the sensitivity of outcome to various assumptions about missingness. The place to start learning about how to deal with the inevitable problem of missing data is the book by Little and Rubin (1987) and the references therein. Closely allied is Rubin's (1987) book on using multiple imputations to assess the variability due to nonresponse in surveys; also Rosenbaum's (2002) masterwork on observational studies should not be missed.

4. Picturing data—A graph of data is the best way to find something that you were not looking for. It is too easy to do analyses that are so complex that you cannot tell if you made an error. It is absolutely imperative that through various kinds of exploratory analyses, mostly graphical, you first get to know what you have done. Then confirmatory analyses can tell you how well you seem to have done it. For a start on exploratory analysis, there is no place better than the original (Tukey, 1977). For graphical display of results, we have a wealth of wonderful books. Jacques Bertin's *Semiology* (1973/1983) is the most thorough, but sometimes heavy going. Lee Wilkinson's (2005) *Grammar* is a *tour de force* and a guide to anyone who would like to think about data display systematically. Edward Tufte's four marvelous books (1983/2000, 1990, 1997, 2006) are both an education and a treat. My own attempts (Wainer, 1997/2000, 2005, 2009) meld graphic display with statistics and history in a way I believe that helps broaden understanding. And last, Cambridge University Press has recently republished William Playfair's original books in which he invented graphs (Playfair, 1801a, 1801b). These are an inspiration to read and provide examples of how well-prepared data displays can become works of art. It also allows one to get a book for $40 that previously was only available on the rare book market for upwards of $7,000.

5. Writing clear prose—In my years as a journal editor, I became acutely aware of how few journal submissions are written clearly. Part of the blame for this is surely that many authors are writing in a second (or third!) language. This

is understandable, but not forgivable. If you do not feel comfortable writing English, collaborate with someone who does.[1] Alas, I have seen ample evidence that being a native English speaker does not prevent one from writing turgid, unclear prose. In a technical presentation, in which mathematics is important, we must continue to remind ourselves of T. S. Eliot's admonition, "I gotta use words when I talk to you."[2]

Happily there are short, pointed instructional manuals on how to write, which should be taken seriously. The classic *The Elements of Style* by Strunk and White (1959) is a jewel. Let me provide three (slightly modified) samples:

> "4. Write with nouns and verbs. Write with nouns and verbs, not with adjectives and adverbs. The adjective hasn't been built that can pull a weak or inaccurate noun out of a tight place (p. 57)."

> "19. Do not take shortcuts at the expense of clarity. Do not use acronyms . . . unless you are certain the acronyms will be readily understood. Write things out. Not everyone knows that IRT means Item Response Theory, and even if everyone did, there are babies being born every minute who will some day encounter the term for the first time. . . . Many shortcuts are self-defeating: they waste the reader's time instead of conserving it (p. 67)."

Writing technical prose has its own challenges, but the famous mathematician Paul Halmos (1970) has offered some valuable help.

Finally, remember that what you write may live long after you are gone, and you will want to leave as good an image of yourself as you can. Two thousand years ago, the Roman poet Horace spoke not only for himself but also for many scholars and literary people: *Exegi monumentum aere perennius*.[3]

6. A deep understanding of Type I and Type II errors and how they affect strategies for research—too often we think only in terms of Type I errors (false negatives) and ignore the impact of Type II (false positives). This omission is often catastrophic. Let me offer an example drawn from one of the tactics of the U.S. government's "War on Terror" (taken from Savage & Wainer, 2008, but see also Meehl & Rosen, 1953).

The government's use of wiretaps to trap terrorists has been the subject of much debate. However, the debates have been principally focused on ethics not efficacy. Widespread wiretaps is a tactic that, regardless of its legality or its morality, is so unlikely to bear fruit that it should never have been used.

To evaluate the use of wiretaps, we must consider both the chance that we will correctly identify a true terrorist if we have one on the other end of the line, and also the probability that an innocent person will be incorrectly identified as a terrorist. This latter probability depends crucially on the overall prevalence of terrorists in the population that we are listening in on.

Let us start with the question: How many terrorists are currently in the United States? Not thugs, murderers, or rapists but hard-core terrorists intent on mass murder and mayhem. I have no idea, but for the sake of our model, let us assume that among the 300,000,000 people living in the United States, there are 3,000 terrorists; if there were twice as many as this or half as many, the conclusions of my argument would be unchanged. Or, in other words, one person in 100,000 is a terrorist.

Now consider a magic bullet for this threat; unlimited wiretapping tied to advanced voice analysis software on everyone's phone line that could detect would-be terrorists within the utterance of three words. The software would automatically call in the Federal Bureau of Investigation (FBI) as required. Let us assume that the system was 99% accurate. That is, if a true terrorist was on the line, it would notify the FBI 99% of the time, while for nonterrorists, it would call the FBI (in error) only 1% of the time. Although such detection software probably could never be this accurate, it is instructive to think through the effectiveness of such a system if it could exist.

When the FBI gets a report from the system, what is the chance that it has identified a true terrorist?

To answer this question, we must realize that when the FBI gets a warning, it either has the correct report of a true terrorist or the false report of a nonterrorist. Of the 3,000 true terrorists, 99% or 2,970 would be correctly identified. Of the 299,997,000 nonterrorists (300 million minus the 3,000 terrorists), only 1%, or 2,999,970 would be falsely reported.

Thus, the probability of correctly identifying a true terrorist is only about one chance in a thousand, even with a 99% accurate detector. If there were fewer than 3,000 terrorists, this probability would decrease still further. And even if the number of terrorists went up 10-fold to 30,000, the chances of a correct identification would still be only one in a hundred. What looked at first like a magic bullet does not look as attractive once we realize the number of innocent people who would be falsely accused.[4]

Is this probability algebra limited to just the War on Terror? Consider what would be the likely outcome if we had universal AIDS testing. Because the test is far from perfect, we would surely find that the number of false positives would dwarf the number of AIDS cases uncovered. And how much of the agony associated with receiving such an incorrect diagnosis would compensate for finding an otherwise undetected case?

And the problems with errors of this type do not stop here. It is well established that any survey question is answered with error. Sometimes it is large, as when you ask questions like, "when was your last chest X ray?" where an answer of "six months" could easily be off by 100% (e.g., it was a year ago). Sometimes it is small, as when people are asked about sex or ethnicity. An often-confirmed rule of thumb is that no question has less than 3% error. So, let us calculate what happens if, for example, we are interested in making

comparisons between, say Black and White students. For the sake of this example, let us assume that we have a sample of 1,000 of whom 80% are White. Using the 3% error rule means that 6 Black students (3% of 200) misidentify themselves as White. In parallel, 24 White students (3% of 800) misidentify themselves as black. Our inferences will be made on one group that has 782 students identified as White (800 – 24 + 6), but 0.8% of them were, in fact, Black; hardly enough to induce large errors of estimate. But of the 218 students who identify themselves as Black (200 – 6 + 24), fully 11% are white. This can cause a noticeable effect on inferences. Of course, if the ratio of the sizes of the two groups gets more extreme, the effect gets magnified apace.

It is probably impossible to eliminate error and so the best we can do is (i) be aware of the problem and (ii) take steps to minimize it. This is why sampling statisticians insist that minority populations be massively oversampled.

III. The Future of Measurement

The psychometrics of today is both more extensive and better than we need. By my rough estimate, fully 80% of all the testing work done at the Educational Testing Service during my 20 years could have been accomplished well enough with the wisdom contained in Gulliksen's (1950) *Theory of Mental Tests*. Almost all of the remaining 20% could be dealt with using the IRT technology in Lord and Novick (1968), not withstanding the sweetness of elegance in the Bayesian formulation of the recently developed theory of testlets (Wainer et al., 2007). There is nothing that I have seen in 40 years in this field that suggests otherwise.

Let me begin an elaboration of this remarkable conclusion with an analogy. In (American) football, the goal is to move the ball to the very end of the field and score a touchdown. An intermediate goal is to move the ball 10 yards in four tries. If a team is successful in doing so, they get to keep control of the ball and get four more tries. This is called "getting a first down." A team's success at getting a first down is often determined by a referee unpiling a sizable group of very large men, taking the ball from one of them, who is invariably at the bottom of the pile, and plunking it down on the field in a place that represents, to the best of his ability, the forward progress of the person with the ball. Then a chain is brought out from its resting place on the side of the field. The chain is 10 yards long and is anchored at both ends with poles. One pole had been placed where the previous first down had been achieved; the second pole was placed 10 yards down the field. When a measurement is called for, a referee picks up the chain from its middle, where it passes over one of the lines that cross the field every five yards and runs out onto the field. Two other officials carry the poles. When they get to the part of the field where the ball lies, the referee with the chain lays it down on the cross line and the two ends of the

chain attached to the poles are stretched out. If any part of the ball is beyond the farther pole, it is ruled a first down, the sticks are relocated on the sideline to reflect this, and play continues. The measurement is made to the millimeter despite the grossness of the actual placement of the ball.

Psychometrics is analogous to the chain; ball placement's analog is test construction. If we want to improve the practice of testing, there is much more bang for the buck to be had in improving tests than improving test theory. How might this be done?

> Evidence may not buy happiness,
> but it sure does steady the nerves.
>
> —Paraphrasing what Satchel Paige said about money

7. Evidence-based[5] test design (sometimes also referred to as evidence-centered design and thus designated as ECD)—These exciting new methods for designing tests were pioneered by Linda Steinberg, Bob Mislevy, and their colleagues (e.g., Almond, Steinberg, & Mislevy, 2002, 2003; Mislevy, 2006; Mislevy, Steinberg, & Almond, 2003; Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002; and references therein) and represent a true breakthrough in how tests can be built to provide much more information than they currently do. The basic ideas are so fundamental that it hardly seems revolutionary.[6]

ECD begins by recognizing the fundamental idea that assessment is about inferences related to the state of examinee proficiency, which is unobservable. These inferences require evidence. ECD is a way of defining the meaning of a score before the test is implemented so that the inferences drawn are explicitly supported with evidence. Understanding what the evidentiary requirements are for supporting the claims of interest allows a rational, considered approach to evaluating costs and benefits of a particular implementation of these requirements.

To understand how ECD works, let us first consider how traditional assessments are designed. Typically the content domain is divided into very detailed outlines as provided in standards documents. "Items are authored to be 'about' various topics, leading to a score that can only be interpreted to mean that the student knows 'something' about the domain—but one that cannot be interpreted to support more specific claims. The fact that no collection of inferences relevant to the purpose and audience of the assessment, or the evidence to support them, is designed up front allows assessments to be used more easily for invalid purposes."[6]

In contrast, ECD organizes the content domain into a collection of arguments relevant to the purposes of, and audiences for, the assessment. "The development of this collection of arguments requires the imposition of argument structure on the domain; that is, the understanding of domain content in terms of the elements of an argument and the establishment of the necessary relationships. Items are authored with the sole intent of eliciting pre-defined evidence to support

inferences of interest."[6] By doing this, we are able to make much more direct statements about what a particular score implies about the examinee's proficiency, and we can point to the evidence that supports this inference.

An assessment argument has three parts:

i. The claims—statements about the examinee's unobservable proficiency, formulated to suit the purpose and the audience of the assessment. Example: The examinee can read.
ii. The evidence—what we need to attend to in the examinee's work, behavior, and performance. "Evidence lives objectively in the world and comes to us through physical perception. Data become evidence when they bear on specific inference(s) of interest. Evidence (1) consists of a set of features and characteristics of examinee work, behavior or performance; (2) is directly observable; (3) has 'THE WORK' as its subject and uses nouns and adjectives (this work can be generated by the examinee or selected by the examinee)." Examples of evidence that the examinee can read might be the recognition that the work contains identification of main ideas; relation of supporting details to relevant ideas; recognition of multiple points of view; and extrapolation of main idea to a new context.[7]
iii. The tasks—"Tasks are the situations we construct to afford the examinee an opportunity to produce the evidence we have defined as required. Tasks exist ONLY in service to the generation of evidence. Tasks consist of sets of characteristics of stimulus material (e.g., length and difficulty of reading passage) as well as sets of other features that shape and focus the examinee response."

"The ECD process is iterative and moves through successive stages of knowledge engineering to produce artifacts appropriate for use by various participants in the assessment/curriculum development process (e.g., domain experts, test developers, authors of professional development materials). The evidentiary arguments constructed for summative assessment are useful and reusable not only in designing formative assessments but also essential in providing learning goals for the curriculum."

W. Edwards Deming (1900–1993) told a story that can serve as an instructive parable here. It seems that one day, there was unusually high absenteeism on an automobile assembly line, but the assembly line's speed was left unchanged. This yielded an increase in the number of errors. Happily, these errors were caught at the final checking station and the cars affected were shunted aside and the errors corrected. Correcting errors take a while because the car needs to be disassembled a bit to redo whatever the problem was. A manager, watching the whole process from high above the factory floor, saw the backup at the checking station and seeking to relieve congestion ordered some workers off the assembly line to move to the checking station to help out. Of course, this magnified the problem. This story has been used to illustrate many important lessons. It does not take much insight to see that moving tax money from education to law enforcement

is an immediate analog. It also illustrates the pitfalls of local optimization. But my point here is directly analogous to Deming's point, which he used to enforce his view that you cannot inspect quality into a product, you must build it in in the first place. Similarly, when you try to draw inferences from test scores, the elegance of your psychometrics cannot extract information that was not built into the test in the first place. Evidence-based test design forces us to make explicit the goals of the test and the specific inferences we wish to draw from its scores. How could anyone disagree with this? The only question I have is why has it taken so long to do?

The most impressive work being done on evidence-based test design is now underway at the College Board where Kristin Huff and her colleagues are using it to redesign the entire Advanced Placement program.

8. Value-added models (VAMs)—The term VAM refers to a family of under-researched but overused statistical models, which are used to make inferences about the effectiveness of educational units, usually schools and/or teachers. They are characterized by their focus on patterns in student score gains over time, rather than on student status at one point in time. In particular, they attempt to extract from the data on score trajectories, estimates of the contributions of schools or teachers to student learning.

Interest in such models has been fueled by an increasing emphasis in the United States on holding the public education system accountable for student learning. In 2001, the U.S. Congress passed a law, the No Child Left Behind (NCLB) Act (NCLB, 2002) that requires states receiving federal funds to establish standards for proficiency in reading and mathematics in Grades 3 through 8 based on performance on standardized assessments. Moreover, states must set yearly targets for the proportions of students in each school achieving those standards.[8] NCLB marks a decisive shift away from evaluating districts and schools on the basis of inputs (e.g., per-pupil expenditures) to judging them on the basis of outcomes (e.g., student performance).

The increasingly stringent targets (for each grade) that a school must meet are termed "adequate yearly progress," or AYP. Schools that fail repeatedly to meet their AYP requirements are subject to sanctions. There have been a number of concerns raised with respect to the technical aspects of AYP (Linn, 2004, 2005). One is that AYP involves a comparison in the proportions proficient between different cohorts (e.g., students in Grade 4 in the current academic year in comparison to students in Grade 4 in the previous year.) Accordingly, the apparent change in effectiveness of the school is confounded with intrinsic differences between the cohorts.

Perhaps the most serious objection, however, is that judging schools on the basis of student status at the end of an academic year without regard to their status at the beginning of the year can be unfair to schools serving student populations either with substantial deficits on entry or with high migration rates, or both. Although some of

these schools may be helping their students make excellent progress, too few students reach the proficient level, causing the school to be sanctioned. By contrast, some schools serving advantaged student populations may meet the requirements of AYP despite the fact that the students are learning at a slow pace. Finally, schools may have little or nothing to do with the observed progress that students make.[9]

To remedy this situation, a number of policymakers and measurement specialists have argued that it would be both fairer and more helpful if schools were judged on the basis of how much growth their students achieved. The critical point is that to evaluate growth properly, each student must be tested twice—once at the beginning of the academic year and once at the end. More realistically, students are likely to be tested at the end of each academic year.[10] Recall that the testing provisions of the NCLB Act mandate such a regular testing regime in reading and mathematics in Grades 3 through 8. As a result, states are building databases containing longitudinal student records—precisely what is required for the application of VAMs. At first blush, the prospect of evaluating schools on the basis of an appropriate, aggregate measure of their students' growth is appealing, and it has advantages over the current AYP regulations. Nonetheless, difficulties remain (see Braun & Wainer, 2007, for a fuller discussion). Some districts in Tennessee and the city of Dallas, Texas, are using value-added approaches to evaluate schools. Other states are planning to follow suit.

Although much of the technical work on VAM has been completed—model specifications, estimation, and so on—there is a great deal to be done before they can be safely used to accomplish their stated purposes—and there is precious little time to do it.

Here are three problem areas in which rich rewards surely await anyone who can make a dent in them.

a. How can we make credible causal inferences from the results of a VAM? The typical user of VAMs wants to be able to both estimate the gain a student makes over a particular time period *and* partition the cause of that gain among the student's teachers, school, district, and the student himself or herself. The partitioning is easy—making it causal is a lot tougher.
b. How sensitive are the parameter estimates of VAMs to missing data? Any longitudinal data gathering effort will have missing data. If their missingness cannot credibly be thought of as missing-at-random, we must assess how much bias, and in what direction, does the nonignorably missing data have on the inferences we will draw from the VAM parameters.
c. What do the change parameters of the VAMs refer to if the tests taken change over time? Al Beaton's old saw that "if you want to measure change, don't change the measure" is a valid concern. As students progress through school, the subject matter changes and so too must the tests that measure mastery of that material. What can we infer from change scores based on different material? And, more difficult still, if VAMs are to be applied among students that take different courses, what meaning can they have? Is my gain in physics as large as yours was in French?

9. New kinds of data—The proliferation of computer-based communication devices has made it possible to gather huge amounts of data that until recently were either completely unavailable or were so limited that it was not practical to study their use. An obvious example is response time data. In computer-administered tests, we can trivially obtain response times for individual items. What are we to make of them? It seems obvious that if one examinee gets items correct with equal likelihood as another, but does so faster, there is a strong likelihood, ceteris paribus, that that student is more able. But how can we use the enormous amounts of response time data that are now available? Naively looking at the relationship between response time and proportion correct will yield nothing but misunderstanding. Such dopey analyses will strongly suggest that we can improve student performance by giving them less time. Surely a bad idea! To learn what is the relationship will require more clever data gathering in which we experimentally assign examinees at random to differing amounts of time and then see how scores vary. Such a paradigm is not hard to do, even within the restrictions of large-scale standardized tests (Wainer, Bridgeman, Najarian, & Trapani, 2004).

10. Integrating computerized adaptive testing (CAT), diagnostic testing, and individualized instruction—The promise of CAT has yet to be fully realized. So far, when it has been applied, it has been used as a mechanical horse, not doing much more than could have been done with paper and pencil testing except that it is faster (a little) and more expensive (a lot). The future must be brighter for such an ingenious method. One obvious place is in diagnostic testing, perhaps coupled with individualized instruction (intelligent tutoring systems are one version of this that cries out for such technology). Currently, large-scale tests are given by states for many reasons (the requirements of NCLB are the principal ones). Such tests are often sold to the public as aids to guide instruction. This is rarely true. When a test is given in March and the score reported to the teacher, the following September it is of little help for individualized instruction, as the students are no longer in that teacher's class. Large-scale tests in general take a while to score. However, if a computer gave such tests adaptively, they could provide immediate detailed information about each child. It seems to me worthwhile to do some medium-scale pilot studies that would provide some sort of cost-benefit analysis. Such uses would have to be widespread to amortize the considerable costs of item pool development over a lot of examinees. But the pools would need to be large to span enough subareas with enough detail to yield reliable estimates of an examinee's weaknesses with enough detail to guide instruction. Yet the usual reason for large item pools, item security, would not be operative if the tests were used for diagnosis; no one cheats on an eye test.

Such tests could prove powerfully useful for instruction and, because the stakes are low and the feedback immediate, could be popular. The technology to do this

is already well established, although I suspect that proficiency estimation is likely to be done in a multidimensional space, and Bayes nets (e.g., Almond, 1991, 1995) will probably come in handy.

IV. Enough, for Now

Earlier, I suggested that we already had enough psychometrics for current purposes, and efforts in other directions would be more likely to bear fruit. This does not mean that no one should work in these areas, but only that the primary focus of the field should, in my view, shift in other directions. In addition, I most expressly do not mean that we should not apply these methods to current problems, only that further research into their expansion and continued development should be of lower priority.[11] Some areas that could profit from some neglect are:

11. Differential item functioning (DIF)—There are, fundamentally, two approaches developed for studying DIF; observed score methods, of which the Mantel-Haenszel statistic (Holland & Thayer, 1988) is both the best known and the best performing, and model-based methods, in which likelihood-ratio tests provide the probability of DIF (Thissen, Steinberg, & Wainer, 1988, 1993). These two approaches are more than enough to suit virtually any occasion. Journal editors I have spoken with admitted to the same feeling I had as an editor when a new submission arrived on yet another simulation showing the sensitivity of some DIF method to one variation or another of parameter distributions. Enough already!

An area that could use further investigation is the use of DIF models to study change; simply substitute "before" and "after" for "focal" and "reference" and the same models can provide a delicate measure of the likelihood of a change having occurred.

12. The Rasch model (and IRT in general)—What more do we need to know about IRT to be able to use it well? We have variations for dichotomous models, polytomous models, and mixtures. We can estimate their parameters with least squares, with maximum likelihood, with Markov chain Monte Carlo (MCMC). Admittedly, multivariate IRT (MIRT) is still murky—both technically and epistemologically—and so someone should work on this. This is one area of IRT whose popularity seems to nicely match our need for more in it. Only a small proportion of researchers are spending time on it. In addition, there are a couple of variations on IRT that could still use a little work, principally on applications, to see how well they work. Two of my favorites are

i. Charlie Lewis's zero-parameter logistic model (0-PL) in which all items have the same difficulty and so each examinee's performance is binomial based solely on θ. I like this as a null model. If you fit a 0-PL and then a 1-PL, say, and find that

the fit does not improve significantly, you have shown that there is no evidence to support that the items have different difficulty. This will help control folks who want to make inferences from insubstantial samples.

ii. Jim Ramsay's multiplicative Rasch model in which the probability of getting an item right is the product of the examinee's ability and the item's difficulty. For example, $P(x = 1|\theta) = \exp(b\theta)/[k + \exp(b\theta)]$, where the notation is the usual one and k is the number of distracters. Thus, with a five-choice item, the denominator would be $[4+ \exp(b\theta)]$. θ is defined on the positive real half—line. This model has the usual property of asymptoting at one as θ grows larger and having a lower asymptote of 0.2 (in this case) as θ goes to zero. What I especially like about this (other than the neat way it accommodates guessing) is that it provides a natural zero for proficiency. So, if you see someone lying dead in a roadside ditch you can say, with some assurance, "that person has zero proficiency." By providing a natural zero, it might also yield measurement of proficiency on a ratio scale. Some more work on this would be worthwhile.

13. Factor analysis/path models—Karl Jöreskog's pathbreaking work in the 1970s ushered in a new era of factor analysis. With the introduction of a rigorous statistical foundation, factor analysis became modern. It then developed extensions through various kinds of methods to help us understand covariance structures. Each family of models had its own name, derived from the software that implemented it (e.g., LISREL, Mplus, Amos). As it grew, it acquired a reputation for powers beyond what any sensible person could believe. The analysis of covariance structure became "causal modeling" that David Rogosa (1988) corrected through a vowel movement to yield "casual modeling" and composed a song in its honor. Enough already!

Let me expand a little, lest I be misunderstood. The popularity of these complex factor-analytic models in a broad range of applied fields strongly suggests that methods like these are badly needed. Statisticians have been pretty good about exposing the shortcomings of these methods, especially with regard to their ability to make causal inferences—or even unambiguous descriptions. But the field of statistics has been less forthcoming in providing viable alternatives. Research in these new directions seems to me important. Such research is hindered when so much talent is focused on polishing up methods that do not seem up to the task.

14. New measures of reliability—More than a half-century ago, Lee Cronbach (1951) proposed a lower bound on reliability that he called α. He did this with the anticipation that there would be improvements and so left the rest of the Greek alphabet available for improvements. Subsequent work, for the most part, did not improve matters; it just explained things more fully. For example, Novick and Lewis (1967) provided details explaining when α was actually the reliability of the test and not just a lower bound. Important work, and I am glad we know it. However, we do not need any more on it. Cronbach himself abandoned

reliability, and in 1972, he and his colleagues provided us with a much more powerful substitute, generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). His adapting of variance components to dissect the observed variability of test scores into their component pieces represented an important advance. Continued work on refining traditional reliability seems akin to improving abacus bead design to speed up computation.

V. Caveats and Opportunities

I began this essay with a quotation from the famous Princeton chemist Hubert Alyea (1903–1996). It was meant as a caution, a hubris destroyer, to remind us that regardless of our thoughts at the moment, new, mind-changing evidence may appear tomorrow. All conclusions are, to some extent, tentative. Rarely is that more true than in the sorts of subjective judgments I made in this essay. I am more confident in some of these judgments than in others. The tools for the future are pretty certain. No one's career will be hurt because he or she has a deep understanding of causal inference or can handle the subtle ideas that lie behind modern Bayesian statistics. I have reasonable faith that the various topics I selected as "hot" will in fact garner a great deal of positive attention for anyone who manages to put a substantial dent in any of them. I am less confident in my selection of moribund topics. This is not because I fear that it is likely we will suddenly need a new DIF method or that another study of the efficacy of the Rasch model will suddenly turn around the practice of testing. But only that some topics only appear dead and are, in fact, just hibernating. As an example, consider the statistical topic of multiple comparisons. I believed that we knew all we needed by 1953 with the circulation of Tukey's monograph on the subject. With what was in there, and the Bonferroni (1936) inequality, we could handle pretty much everything that came our way. And so for more than 40 years, we could. However, by the mid-1990s, a new area of application arose: genetic research and microarrays. Now, we might have a study with two groups of, say, 50 subjects and we needed to compare those two groups on each of 100,000 genes. Larger *n*s were expensive and dividing α by 100,000 made it difficult, if not nearly impossible, to find anything significant. A different method of multiple comparisons to control error rates was required. Happily, Benjamini and Hochberg (1995) and their false discovery rate opened our eyes to new opportunities and the possibility of useful research on multiple comparisons awoke. So too it might be in the future when suddenly a new kind of instrument is developed for which a new Rasch model was needed. Who can tell? Never say never.

> I don't know anything "perfectly well," Mr. Danby,
> and I mistrust those who say they do.
>
> —John Galsworthy (1924)

Notes

1. No less an authority than George S. Kauffman characterized the value of collaboration with others as *"Geld by association."*

2. From Sweeney Agonistes.

3. Translation: I have built a monument more lasting than bronze. Also from Strunk and White (1959, p. 67): "20. Avoid foreign languages. The writer will often find it convenient or necessary to borrow from other languages. Some writers, however, from sheer exuberance or a desire to show off, sprinkle their work literally with foreign expressions, with no regard for the reader's comfort. It is a bad habit. Write in English."

4. With this calculation in mind, how should we interpret a news broadcast that tells us that "in a raid today, 20 suspected terrorists were killed?" If the ratio of terrorists to innocents is anything like the 1 to 1,000 we have assumed here, we can safely believe that most, if not all, of the dead were innocent.

5. If this newest way to build tests is called "evidence-based," how should we characterize the previous approach? "Faith-based" seems apropos although my distinguished colleague Don Melnick suggested "Intelligent design," which on reflection has the right feel. More seriously, Linda Steinberg proposes the analogy that traditional test design is to evidence-based design as description is to argument. That is as good a summary as I have heard.

6. The discussion that follows is a fairly close paraphrasing of a description provided to me by Linda Steinberg (October 10, 2007). My gratitude for her help is immense. Any unattributed quotations are directly from her.

7. Example of extrapolation: if you had "A bird in the hand . . . " as the stimulus to be comprehended, the explanation might contain a description of a situation settling for what is sure as opposed to taking a risk. So "extrapolation to a new context" is an evidentiary requirement that can be operationalized in many ways, depending on what you mean by "can read" and the use of various stimuli that flow from that—again, determined by purpose and audience.

8. For more information on this act, consult Bourque (2005). The expectation is that essentially all students will achieve proficiency by the 2013–2014 academic year.

9. To take an extreme example, if we observe children's stature every year, some schools may have students who grow much more than the students at other schools, but few would argue that schools had much to do with the outcomes.

10. Note that end-of-year testing means that student gains during the academic year are confounded with gains (or losses) during the summer.

11. An obvious example from physics is string theory. It was launched with great fanfare as a theory of everything and soon fully half of all graduate students in theoretical physics at top departments were doing string theory to the neglect of other topics. Consensus has it that 90% of those currently doing string theory could more profitably spend their time elsewhere.

References

Almond, R. G. (1991). Building blocks for graphical belief models. *Journal of Applied Statistics*, *18*, 63–76.

Almond, R. G. (1995). *Graphical belief modeling.* New York: Chapman & Hall.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1*. Retreived from http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml. Also available as *CSE Technical Report 543*. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved October 26, 2005, from http://www.cse.ucla.edu/CRESST/Reports/TECH543.pdf

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2003). A framework for reusing assessment components. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 28–288). Tokyo: Springer.

Bayes, T. (1764). An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London* for 1763, *53*, 370–418.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289–300.

Bertin, J. (1983). *Semiology of graphics* (W. Berg & H. Wainer, Trans.). Madison, Wisconsin: University of Wisconsin Press. (Original work published 1973)

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità [Statistical theory of classes and calculating probability]. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8, 3-62.

Bourque, M. L. (2005). The history of No Child Left Behind. In Richard P. Phelps (Ed.), *Defending standardized testing* (pp. 227–254). Hillsdale, NJ: Lawrence Erlbaum.

Braun, H., & Wainer, H. (2007). Value-added assessment. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (volume 27) psychometrics* (pp. 867–892). Amsterdam: Elsevier Science.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: John Wiley.

Dunn, M. C., Kadane, J. B., & Garrow, J. R. (2003). Comparing harm done by mobility and class absence: Missing students and missing data. *Journal of Educational and Behavioral Statistics*, *28*, 269–288.

Fisher, R. A. (1935). *The design of experiments.* New York: Hafner.

Galsworthy, J. (1924). *The White Monkey and a silent wooing* (vol. IV of *The Forsyte Saga*; p. 52). New York: Charles Scribner.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). New York: Chapman & Hall.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice.* New York: Chapman & Hall.

Gulliksen, H. O. (1950). *Theory of mental tests.* New York: John Wiley. (Reprinted in 1987, Hillsdale, NJ: Lawrence Erlbaum).

Halmos, P. R. (1970). How to write mathematics. *L'Enseignement mathématique*, *16*, 122–152.

Hilbert, D. (1902). Mathematical problems. *Bulletin of the American Mathematical Society*, *8*, 437–479.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–970.

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement* (pp. 3–25). Hillsdale, NJ: Lawrence Erlbaum.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–146). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*, 41–54.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 62-A*, Part I, 74–82.

Lewin, K. (1951). *Field theory in social science; selected theoretical papers* (p. 169). D. Cartwright (Ed.). New York: Harper & Row.

Linn, R. L. (2004). *Rethinking the No Child Left Behind accountability system*. Paper presented at the Center for Education Policy Forum, Washington, DC. Retrieved October 26, 2005, from http://www.ctredpol.org

Linn, R. L. (2005, June 28). Conflicting demands of "No Child Left Behind" and state systems: Mixed messages about school performance. *Education Policy Analysis Archives, 13*. Retrieved May 15, 2006, from http://epaa.asu.edu/epaa/v13n33/

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: John Wiley.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*, 304–305.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison Wesley.

Meehl, P. E., & Rosen, A. (1953). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194–216.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). Westport, CT: Praeger.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–67.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education*, *15*, 363–378.

Neyman, J. (1923). On the application of probability theory to agricultural experiments (D. Dabrowska & T. Speed, Trans.). *Statistical Science*, *5*, 462–472.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*, 1–13.

Playfair, W. (1801a). *The statistical breviary.* London: T. Bensley. (Facsimile reprint edited and annotated by Howard Wainer and Ian Spence, 2006, New York: Cambridge University Press).

Playfair, W. (1801b). *The commercial and political atlas, representing, by means of stained copper-plate charts, the progress of the commerce, revenues, expenditure, and debs of England, during the whole of the eighteenth century*. (Facsimile reprint edited and annotated by Howard Wainer and Ian Spence, 2006, New York: Cambridge University Press).

Rogosa, D. (1988, June 28). The Ballad of the Casual Modeler. A song performed at the Annual Meeting of the Psychometric Society by *Statboy & the Casuals*, Los Angeles, California.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer-Verlag.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29*, 103–116.

Savage, L. J. (1954). *The foundations of statistics.* New York: John Wiley.

Savage, S., & Wainer, H. (2008). Until proven guilty: False positives and the war on terror. *Chance*, *21*, 55-58.

Strunk, W. Jr., & White, E. B. (1959). *The elements of style.* New York: Macmillan.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (chap. 4, pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum.

Tufte, E. R. (1983/2000). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

Tufte, E. R. (1990). *Envisioning information.* Cheshire, CT: Graphics Press.

Tufte, E. R. (1997). *Visual explanations.* Cheshire, CT: Graphics Press.

Tufte, E. R. (2006). *Beautiful evidence.* Cheshire, CT: Graphics Press.

Tukey, J. W. (1953). *The problem of multiple comparisons.* . Mimeographed manuscript. Princeton, NJ: Princeton University. (Reprinted in full in volume VIII of the *Collected Works of John W. Tukey* by H. Braun, Ed., 1994, New York: Chapman & Hall).

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement*, *30*, 1–21.

Wainer, H. (1995). Measurement problems. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories in measurement: Problems and issues* (pp. 375–407). Ontario, Canada: University of Ottawa Press.

Wainer, H. (1997). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot.* New York: Copernicus Books. (Reprinted in 2000, Hillsdale, NJ: Lawrence Erlbaum)

Wainer, H. (2005). *Graphic discovery: A trout in the milk and other visual adventures.* Princeton, NJ: Princeton University Press.

Wainer, H. (2009). *Picturing the uncertain world: How to understand, communicate and control uncertainity through graphical display*. Princeton, New Jersey: Princeton University Press.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications.* New York: Cambridge University Press.

Wainer, H., Bridgeman, B., Najarian, M., & Trapani, C. (2004). How much does extra time on the SAT help? *Chance*, *17*, 19–24.

Wang, X., Bradlow, E. T., Wainer, H., & Muller, E. (2008). A Bayesian method for studying DIF: A cautionary tale filled with surprises and delights. *Journal of Educational and Behavioral Statistics*, *33*, 363–384.

Wilkinson, L. (1999). *The grammar of graphics.* New York: Springer-Verlag.

Wilkinson, L. (2005). *The grammer of graphics* (2nd ed.). New York: Springer-Verlag.

Author

HOWARD WAINER is Distinguished Research Scientist at the National Board of Medical Examiners and adjunct Professor of Statistics at the Wharton School of the University of Pennsylvania, 3750 Market Street, Philadelphia, PA 19104; HWainer@NBME.org. He is a fellow of the American Statistical Association and the American Educational Research Association. He received the Educational Testing Services Senior Scientist Award, the 2007 NCME Career Contribution Award, the 2006 NCME Award for Scientific Contribution to a Field of Educational Measurement (jointly with Xiaohui Wang and Eric Bradlow), and the 2009 *Samuel J. Messick Award for Distinguished Scientific Contributions* from Division 5 of the American Psychological Association. His most recent book is *Picturing the Uncertain World: How to Understand, Communicate and Control Uncertainity Through Graphical Display* (Princeton University Press, 2009).