Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 6

Revolutions and Evolutions in Current Educational Testing*

Robert L. Brennan†

May 2004

---

*A keynote address presented at the Seventh Wallace National Research Symposium on Talent Development, The University of Iowa, Iowa City, IA, May, 2004. The author thanks Michael T. Kane, Robert L. Linn, and Ernest T. Pascarella for helpful comments on a previous draft. Date of last revision: June 2, 2004.

†Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Director, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: robert-brennan@uiowa.edu).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Contents

Abstract

Educational testing is currently undergoing both revolutions and evolutions that will have far-reaching and long-term consequences. Perhaps the most salient revolution is the unprecedented movement towards using testing for high-stakes accountability decisions in the K–12 arena, primarily through the "No Child Left Behind" Act. Also, in various educational arenas computerization is influencing virtually all aspects of testing, the effects of which may be revolutionary in some cases and evolutionary in others. In addition, litigation (even merely the threat thereof) is playing an increasingly important role in testing—a role that has serious actual and potential implications. These revolutions and evolutions are abetted and/or impeded by various factors including, for example, the availability of widely endorsed *Standards for Educational and Psychological Testing*, a failure to recognize the difficulty and expense of developing and validating good tests, a notable shortage of professionals who have extensive training in testing, and other capacity problems.

Most educational historians would agree, I think, that Sputnik was the impetus for an evolutionary, and perhaps revolutionary, change in American education. That occurred nearly 50 years ago. Years from now, I suspect historians will look back on the current times and declare that this too was a time of historic change in American education. What distinguishes the current revolution from previous ones, however, is the tremendous emphasis given to testing from pre-kindergarten all the way through professional licensure and certification. I certainly believe that testing can be, and usually is, a positive force throughout our educational system, but I have some reservations about the directions being taken in testing, its expanded use, and the unbridled enthusiasm for testing that seems so widespread. Both optimism and pessimism are evident in this paper's treatment of current revolutions and evolutions in educational testing.[1]

I often view educational measurement in three different testing contexts: K–12, admissions to college and professional schools, and licensure and certification. Most important technical issues pervade all contexts, but other issues tend to vary by context, at least in emphasis. In this paper, I give only passing attention to technical issues; my main focus is on trends, and even megatrends, that are emerging in educational testing in the 21st century.

This paper is not a typical research article. There is a factual and research component to most of the topics covered (see, especially, the footnotes), but there are also historical and philosophical perspectives presented here that are surely subject to debate. In fact, this paper is intended, in part, to advance a healthy debate among educators, measurement specialists, politicians, and others who have a stake in testing—which means virtually everyone! Testing is not, and never has been, the sole prerogative or responsibility of measurement experts. Indeed, testing is never an end unto itself—it serves some other purpose(s). Many if not most debates about testing are essentially debates about the value of testing for a particular purpose in a particular context.

In this paper I focus on three principal topics:

- accountability in the K–12 arena;
- computerization; and
- litigation and notions of fairness.

I believe that the current attention being given to each of these topics is dramatic enough for them to qualify as revolutions or at least evolutions in testing. I end this paper with a consideration of some issues that are abetting and/or impeding progress in these areas and others.

Accountability in the K–12 Arena

I readily admit that if I had been asked 15 years ago to predict the status of K–12 educational testing in 2004, I would have been very wrong on several counts.

[1] Slightly rewritten paragraph from Brennan (2001, p. 6).

Most importantly, I never would have predicted that the public and politicians on both sides of the aisle would be so enthusiastic about using testing as a high-stakes instrument of public policy and accountability. Rather, I would have guessed that there would be widespread skepticism about testing, with frequent references to it being overused and misused. I was wrong! I failed to recognize that a testing revolution was underway in this country that was based on the nearly unchallenged belief (with almost no supporting evidence) that high-stakes testing can and will lead to improved education.

The single most defining event of this revolution was the passage of the revised Elementary and Secondary Education Act of 2001, with the rhetorically brilliant name "No Child Left Behind" (NCLB, 2002), which was signed into law by President Bush on January 8, 2002. Very few Washington politicians or bureaucrats could or would argue against an act with a name that seemed to promise a quality education for every child. The current conventional wisdom espoused by most politicians, many business persons, and a large number of educators seems to be that NCLB may need "tweaking" and more funding, but otherwise the Act is on target. My contention, however, is that, although the branding encapsulated in the NCLB name is extraordinarily clever, NCLB in anything like its current form is not likely to advance reasonable use of tests in advancing sound educational policy. The accountability provisions of the Act and its regulations are outrageously unrealistic and poorly conceived from a measurement perspective. The belief that a few mid-course corrections and additional funding are all that are needed is misguided, at best. There are some very worthwhile aspects of NCLB, but for the most part, I would argue that a more accurate title for the Act might be "Most Children Left Behind." Before addressing this claim and related problems with the Act, I first consider some of the history that led to NCLB and then some of its provisions/regulations.

Some History

In 1965 congress passed the Elementary and Secondary Education Act (ESEA) which became the cornerstone of the federal government's efforts to help the educationally disadvantaged.[2] At nearly the same time, the National Assessment of Educational Progress (NAEP) began as a largely invisible (at least to the public) federal testing program that reported how the nation's students were performing at three age/grade levels (roughly fourth, eighth, and twelfth grade) on selected items (not tests).[3] The ESEA and NAEP were not "linked" in any serious measurement or policy sense. In fact, there was considerable fear among some that NAEP might become a high-stakes federal testing program like those in some European countries. To help preclude that possibility, it was written into law that NAEP could not report scores for individual students.

As time went by, both the ESEA and NAEP evolved and became much more

[2]ESEA has numerous provisions. Those of concern here fall mainly within Title I or Chapter I, depending on which version/reauthorization of the Act is under consideration.

[3]See Pellegrino, Jones, & Mitchell (1999, pp. 12–20) for a brief history of NAEP; see also Jones & Olkin, 2004.

prominent and influential.[4] In particular, although the provisions of ESEA under consideration here seemed to focus only on educationally disadvantaged students in the various states, in fact ESEA had tremendous influence on other students and many aspects of K–12 education for two reasons. First, there were strings attached to the receipt of ESEA funds. Second, the amounts of money distributed to the states under ESEA were large enough that they leveraged a great deal of educational policy and practice. Fundamentally, most school districts badly needed the money that ESEA provided, but fulfilling the requirements incurred by receipt of the funds had consequences (perhaps unintended) for the delivery of instruction to many students—disadvantaged or not.

By the late 1980s and early 1990s NAEP had evolved dramatically. For example, item scores were replaced by test scores, many new tests were introduced, reports were beginning to be provided to states on a "trial" basis, and the primary reporting mechanism was beginning to change from scale scores to achievement levels (below basic, basic, proficient, and advanced.) Also, there were tentative efforts made during the first Bush administration that could have led to NAEP reporting at the student level; subsequently, Clinton considered much the same matter when he proposed the so-called "Voluntary National Tests" (VNTs) in his 1997 State of the Union address.[5]

Although NAEP and NCLB are legislatively distinct, there are at least two ways that they are related. First, for both NAEP and NCLB, achievement levels (particularly "proficient") loom large as reporting categories. Second, NAEP will be playing some kind of confirmatory or monitoring role with respect to states' reports of their NCLB status, as discussed later.

Since the 1980s, there has been a series of reports, including "A Nation at Risk" (National Commission on Excellence in Education, 1983) and the "Goals 2000: Educate America Act" of 1994, suggesting that the United States' educational system is in serious trouble. At the same time, various international studies have suggested that the United States' educational system is at best "average" compared to other developed countries. These reports and studies were particularly influential among business people who basically claimed that the United States was not providing students with a good enough education to perform adequately in the workplace. It is not my intent to argue these points, although I think many of them are at least debatable if not grossly misleading (see, for example, Berliner, 2004). Rather, I mention these matters as one contributing factor in the history that led to NCLB.

In short, in my view, the ever increasing influence of the ESEA, the evolving nature of NAEP, and the high visibility of negative reports about schools in the US have all contributed to a national movement towards the use of high-stakes

[4]During the Clinton administration, the reauthorization of the ESEA was called the Improving America's School ACT (IASA) of 1994. As noted by Linn (2003), it "charted a new direction for testing and reporting for purposes of Title I by the states" (p. 7). According to Cohen (2002), IASA "placed considerable trust in states to work out the details for themselves" (p. 43). That trust largely evaporated with the next reauthorization of ESEA, namely NCLB.

[5]Both Presidents' initiatives failed, but we will return to this matter later.

testing in K–12 education—a movement that culminated in the No Child Left Behind Act of 2001. At the same time, there has been an almost unchallenged assumption, with very little supporting evidence, that if the testing stakes are high enough, the educational system can and will improve to a dramatic degree.

Some NCLB Provisions/Regulations

The provisions of NCLB are many and complicated. No attempt will be made to list all of them here; nor will they be considered in the order or format in the legislation or the regulations. Rather, the intent here is to capture the highlights. Essentially NCLB requires that every state receiving ESEA funding do the following:[6]

- develop (or adopt) challenging academic content standards;
- administer annual tests in reading/language arts and mathematics to every student in grades 3–8 now and in one high-school grade beginning in 2005–2006;
- administer annual tests in science to every student in one of the grades 3–5, 6–9, and 10–12 beginning in 2007–2008;
- identify challenging student academic achievement standards (i.e., achievement levels) that provide (among other things) a state's definition of what it means to be "proficient" relative to the *state's* standards;
- ensure that tests are aligned to the state's content standards;
- ensure that tests meet accepted professional measurement standards;
- provide a plan for Adequate Yearly Progress (AYP) that leads to *all* students being proficient (or above) by 2014;
- ensure that AYP is satisfied *each* year for *all* subgroups, where subgroups include economically disadvantaged students, major racial/ethnic groups, disabled students, and students with limited English proficiency; and
- participate in biennial administrations of state NAEP in reading and mathematics in grades 4 and 8.

The AYP provisions have two additional features:

- Ninety-five percent of each subgroup must take the assessments on which the state's AYP is based. This is a requirement over and beyond the proficiency requirement. That is, for a subgroup to achieve AYP, 95% of the subgroup must be tested *and* the percent proficient standard for that year must be achieved.

[6] It is acknowledged that some of these requirements have been softened recently, and some states have obtained limited exemptions to some aspects of these requirements. However, these provisions are generally still in place and central to the thrust of NCLB.

- The so-called "safe harbor" provision permits a subgroup exception if: (a) the number of students in the subgroup scoring below proficient is reduced by at least 10% from the previous year; and (b) the subgroup made progress on at least one other state indicator.[7] The safe harbor provision, however, fails to benefit many schools.[8]

It is sometimes not recognized that the requirements of NCLB are conjunctive, not compensatory. That is, all of the requirements must be met individually; high performance in one area does not offset low performance in another. Conjunctive requirements are known to be particularly difficult to achieve. All of these conditions for a particular state apply to every school and district within that state. For any school, failure to meet these conditions in any one year results in a warning; failure two years in a row leads to a "needs improvement" label. Repeated failure results in progressively more onerous sanctions.

NCLB Promises, Pitfalls, and Contradictions

Many aspects of these provisions seem laudable, at least at first blush. In particular, very few people would argue against the imposition of challenging academic content and achievement standards for students. However, there is no real consensus among educators, politicians, or the public about what "challenging" means or should mean. Furthermore, there is more than ample evidence now available that the standards of "proficiency" vary dramatically by state (see, for example, Linn, 2003, and McLaughlin & Bandeira de Mello, 2002). In essence this implies that proficiency in reading/language arts and in mathematics means very different things in different states. Therefore, it is not only possible, it is virtually certain, that the sanctions meted out by the federal government will be inequitable in many cases. For example, consider two schools in different states whose students are equivalent in reading/language arts skills. It is entirely possible that one of these schools will be sanctioned and the other will not solely because their respective states have different definitions of "proficient." If the public were aware of this, would they believe that such differential treatment is reasonable? I doubt it—especially for the public in the state with the sanctioned school!

NCLB also has the laudable requirement, in my opinion, that every student be tested in grades 3–8 (and once in high school) in reading/language arts and mathematics (with slightly less stringent requirements in science). However, as discussed below, the every-student testing provisions of NCLB are not likely to be nearly as educationally beneficial as they could be, and these provisions are not nearly as focused on every child as the phrase "No Child Left Behind" would suggest.

Cohort-to-cohort vs. longitudinal change. NCLB does not mandate or even focus on monitoring the progress of *individual* students over time—i.e., individ-

[7] See Linn (2003, Winter) for a hypothetical example.
[8] Linn, personal communication, May 12, 2004.

ual longitudinal analyses.[9] Rather, the focus of AYP in NCLB is on the progress of successive groups of third graders, successive groups of fourth graders, and so forth—i.e., cohort-to-cohort analyses.[10] These are not simply two different ways of addressing the same issue; they address fundamentally different concerns. Well-conducted longitudinal analyses permit defensible conclusions about student progress and instructional effectiveness.[11] Cohort-to-cohort analyses are essentially evaluations of changes in teacher/school performance, without any direct evidence about the progress of *individual* students. Furthermore, in cohort-to-cohort analyses teachers and schools are essentially evaluated against a moving target of different cohorts of students, which makes year-to-year comparisons both ambiguous and highly suspect. This is particularly problematic for small schools, because their student populations can differ dramatically from year to year.

Every-student focus? One of the government-stated motivations of NCLB is, "All children in America must have the chance to learn and succeed."[12] NCLB and its regulations, however, do not really focus on "all" children, or even most children. Many students who might be labeled "average" or "above average" and virtually all talented and gifted students don't count in satisfying AYP.[13] Rather, the focus is on children who are not proficient, according to very fluid definitions of proficient. Furthermore, as depicted in Figure 1, attempting to meet the provisions of NCLB could well lead to extraordinary attention being given to students who are just below the basic/proficient cutpoint (sometimes called "bubble" students), with perhaps considerably less attention given to other students. From these perspectives, I would argue that the current reauthorization of ESEA might be more aptly titled "Most Children Left Behind."

I am not arguing that states and districts purposely neglect other students. I am simply pointing out that under NCLB there are few rewards (and hence little motivation) for improving the educational attainment of students who are already proficient—which, in states with relatively modest definitions of "proficient," is most students.[14]

Also, I emphatically do not wish to make any claim that one group of students is more or less worthy of educational attention than another. *Indeed, I would argue that the overarching goal of our educational system should be to maximize the potential of every student, no matter what that student's status may be.*[15] The rhetoric surrounding NCLB seems to support this position, but

[9] The safe harbor provision might be viewed as an exception, but, as noted earlier, it seldom results in a subgroup achieving AYP.

[10] Cohort-to-cohort analyses are sometimes called cross-sectional analyses.

[11] Linn, Baker, and Herman (2003, Winter) discuss how various types of longitudinal analyses might be folded into NCLB.

[12] http://www.ed.gov/nclb/overview/welcome/closing/edlite-slide047.html

[13] An individual state might adopt a definition of AYP that pays attention to the progress of all students, but the regulations per se do not require doing so.

[14] The only exception known to me is that NCLB provides a modest amount of funding to reward schools that substantially close the achievement gap between the lowest- and highest-performing students.

[15] That is one reason why I believe that analyses of longitudinal data at the individual-student level are much better than cohort-to-cohort analyses for judging the effectiveness of

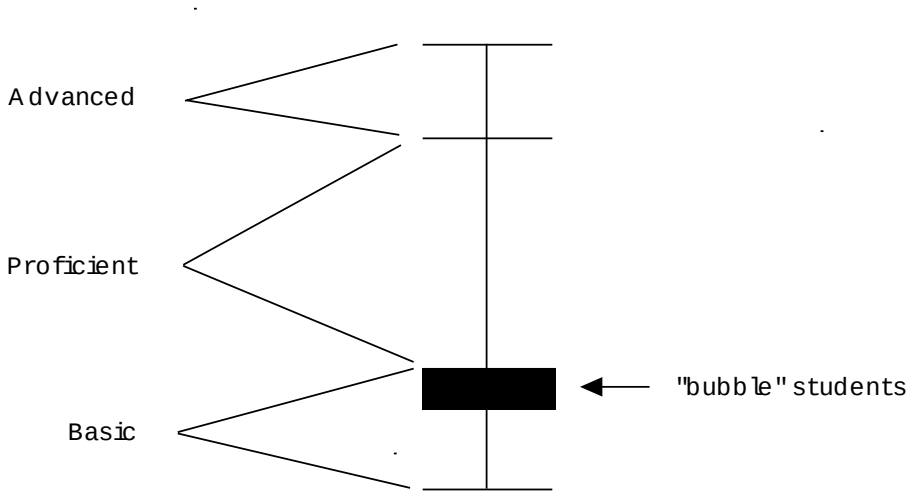

Figure 1: Is it really "NO child left behind"?

the provisions and regulations of the Act do not.

As many others have pointed out, there are two fundamental principles that appear to guide educational policy in the United States—*equity* and *excellence* (see, for example, Gallagher, 2004). These principles are almost always in some state of tension because it is very difficult for both of them to be achieved simultaneously. Still, it seems blatantly obvious that NCLB neglects average, above average, and gifted students to a dramatic degree. Can our educational system be "first class" in any meaningful sense if disproportionate attention is given to students who fall in only one particular range of achievement, at the expense of other students? Such a question clearly raises philosophical and practical concerns that beg to be addressed.[16]

Extreme difficulty of achieving AYP. Since most students are already classified as "proficient" in many states, a natural conclusion that might be drawn is that achieving AYP should not be too much of a challenge in those states. Unfortunately, nothing could be further from the truth. Analyses by knowledgeable, competent researchers consistently lead to the conclusion that achieving AYP almost certainly will get progressively more difficult year after year for most schools,[17] with increasingly large numbers of schools failing the AYP hurdle in virtually every state—even schools that already have relatively large proportions of proficient students. The problem is not so much achieving some degree of consistent progress over time, but rather achieving *sufficient* progress so that *all* students are proficient by the 2013–2014 academic year. No one has been able to demonstrate, or even provide a reasonable basis for believing, that such

our educational system.

[16]Kaplan (2004) has recently raised such questions in the context of talented and gifted students, but the questions apply even more broadly.

[17]See, for example, Linn (2003) and Linn (2003, Winter).

a goal is attainable. As Linn (2003) states,

> (a)t the very least, there should be what I call an existence proof. That is, we should not set a goal for all schools that is so high that no school has yet achieved it. For example, if no school has 100% of its students scoring at the proficient level or higher, we should not expect all schools to reach that level . . . (p. 4).

A particularly salient and, in my opinion, inconsistent aspect of NCLB is the requirement of 100% proficiency for all schools by the 2013–2014 school year[18] with no common definition of proficiency across states. Presumably, the federal government wants to avoid demanding that each state achieve a *federally*-defined standard, but in failing to do so, the federal government is demanding 100% adherence to an incommensurably defined set of standards.

Using NAEP to confirm states' results. It certainly appears that the federal government recognizes the inconsistencies in states' definitions of "proficiency." I presume that is one of the reasons for the requirement of biennial testing with NAEP to confirm state results. The following is a sampling of some potential problems and issues involved in this use of NAEP.

- The proficiency standards for different states do not usually correspond with NAEP proficiency standards (see, for example, Linn's, 2003, pp. 8–9 discussion of the Colorado and Massachusetts standards). Furthermore, states' proficient levels often seem closer to the NAEP basic level than the NAEP proficient level.

- The National Assessment Governing Board (NAGB), which is the independent agency responsible for developing NAEP, is proposing a very weak notion of "confirmation." Specifically, NAGB states, "Any amount of growth on the National Assessment should be sufficient to 'confirm' growth on state tests" (NAGB, 2002, p. 9). It remains to be seen whether the U.S. Department of Education, the states, and the public will accept such confirmatory evidence as sufficient.

- When NAEP scores and states' own AYP data suggest different conclusions, which conclusions will be believed and acted upon? For example, it is a virtual certainty that a rank ordering of states according to NAEP scores will differ (and probably considerably) from a rank ordering of states based on states' AYP results. The public will almost certainly ask which results are credible.[19] Some states may appear to be performing better based on NAEP than on their own AYP results; for other states the reverse may be true.

[18] Linn, R. L. (2003, Winter) provides a very readable discussion of this issue and an illustration of some of its likely consequences.

[19] The mere fact that NAEP and the various state testing programs are different virtually guarantees this inconsistency. See, for example, the report entitled *Uncommon Measures: Equivalence and Linkage among Educational Tests* issued by the the National Research Council (Feuer et al., 1999).

- Given the high stakes of NCLB for schools and states, one likely effect of using NAEP to confirm states' AYP results is the *de facto* elevation of NAEP to a federally-mandated *high*-stakes testing program. The only substantive sense in which this is not quite true is that NAEP is not an *every*-student testing program—at least not yet! I am not arguing whether or not NAEP should be an every-student testing program, but I am asserting that its potential use in confirming states' AYP could move NAEP (or some clone of it, such as reincarnated VNTs) one step closer to becoming an every-student testing program sponsored by the federal government.[20]

"Scientifically-based" research. A particularly blatant contradiction, in my opinion, involves the role of research in the Act. The reauthorization of ESEA makes repeated reference to the need to make educational decisions based on "scientifically-based" research, which I would argue is a laudable goal. However, there is no provision in NCLB for a scientifically-based evaluation of NCLB! At a minimum, it would seem sensible that some small fraction of the funding for NCLB be set aside for an independent evaluation of the Act by qualified researchers.[21] Without such a provision, there is a real risk that "evaluations" of NCLB are likely to be highly subjective and influenced by political considerations at the expense of objectivity. If educational accountability is an NCLB goal, then I contend that goal should be extended to the Act itself.

Excess meaning in labels. There are other aspects of NCLB that seem to be internally contradictory, or nearly so. For example, learning disabled students are required to make AYP and achieve 100% proficiency by the 2013–2014 school year, just like all other students. Whether or not this goal is realistic, suppose for the moment that it were attained in a particular state. In what sense would it then be meaningful to characterize any of the state's students as learning disabled if all of them are proficient according to the same definition of proficiency used with other students in the state? This very confusing matter gets greatly compounded when we recognize that the various states have different definitions of proficiency.

Testing for accountability vs. instruction. Testing for accountability purposes is essentially verifying the extent to which learning has occurred, which is sometimes called summative evaluation. In formative evaluation, testing is used to promote learning. These two forms of evaluation might be called assessment *of* learning and assessment *for* learning, respectively.[22] They are not the same thing. It is particularly important to note that under NCLB, assessment *of* learning is a once a year summative activity. By contrast, assessment *for* learning is virtually continuous, or should be. The current high-stakes account-

[20] There is a peculiar irony in using a testing program that prohibits every-student score reporting (NAEP) to confirm results for an Act that requires every-student testing (NCLB).

[21] This is not a particularly novel idea; such independent evaluations have been part of other legislation in the past (e.g., Head Start).

[22] Richard Stiggins has made this point eloquently on numerous occasions including in an address at the CASMA-ACT Conference on "Current Challenges in Educational Testing" in Iowa City on November 8, 2003.

ability movement in K–12 education gives very little, if any, consideration to assessment *for* learning.

State vs. national considerations. Historically, in the United States the responsibility for K–12 education has been vested in the states and, accordingly, the lion's share of the funding for education comes from the states. However, the states' control over education has eroded over time, primarily because of the strings attached to ESEA funding (most recently NCLB). The proportional amount of funding for education coming from NCLB (approximately 5-7%) may not appear substantial, but the absolute dollar amounts are so large that no state can do without the money. Also, this funding comes with costly requirements that are difficult to implement and that impact virtually all students, not just the educationally disadvantaged for whom NCLB is presumably intended. In effect, NCLB funding leverages the entire educational system in the United States. In that sense, important aspects of education are no longer really under the control of the states.

I would argue that, in the context of NCLB, the principal way the federal government acknowledges states' rights with respect to education is by assigning to each state the responsibility for defining and implementing its own definition of proficient. That is a slippery slope, however. The net result is that proficiency means different things in each state. Does it really make any educational or accountability sense to say, for example, that proficiency in reading in grade four in Iowa is—and should be—different from proficiency in reading in grade four in Massachusetts? NCLB encourages this type of cacophony.[23]

Computerization

Nearly since the advent of computers, it has been predicted that they would revolutionize education and testing. Some of these predictions have come true, others are beginning to be implemented, and still others have yet to be realized. Some knowledgeable persons might argue that computers have not yet revolutionized testing, but even those persons would agree, I think, that eventually computers will have a major impact on measurement. My own belief is that the role of computers in testing is partly evolutionary and partly revolutionary.[24]

In the 20th century perhaps the single most important technological development in testing was E. F. Lindquist's invention of the optical scanner.[25] Without Lindquist's invention, it would have been impossible for testing to advance at the rate that it did in the second half of the twentieth century. The optical scanner, however, primarily impacts only one aspect of testing—namely, the conversion of bubbled responses on an answer sheet to item and examinee raw scores.[26] Furthermore, the positive impact is largely with respect to speed

[23] Such glaring inconsistencies actually predate NCLB (see, for example, Musick, 1996), but NCLB exacerbates the problem because it elevates the stakes.

[24] See Roorda (2004) for an expansive perspective on the role of technology in testing.

[25] See Petersen (1983) for an historical treatment.

[26] Scanners are now being used to scan essays and other types of constructed responses prior

and cost.[27]

Indirectly, it can be argued that the invention of the optical scanner had consequences that the inventor perhaps did not fully anticipate. It made scoring multiple-choice items so easy, fast, and inexpensive that, in most contexts, no other testing modality could compete. In this sense, it might be argued that the optical scanner effectively impeded the growth of what are called these days "alternative" assessments. It is interesting to speculate about how Lindquist would perceive the influence of his invention. My own guess is that he would be encouraging us to make innovative use of computers, even if doing so resulted in reduced use of his invention. After all, Lindquist was almost always ahead of his time!

Whereas the optical scanner primarily impacts only one aspect of testing, computers have the potential to impact virtually all aspects. At the risk of oversimplification, consider the following tasks that are part of just about every testing program:

1. registration
2. item development
3. test assembly
4. test administration
5. item/test scoring
6. score reporting

Many testing programs already make considerable use of computerized graphical-user interfaces and complex databases to register examinees over the web. In some testing programs, the development of items is partially automated through the use of algorithmically-generated items (e.g., item forms, item clones, etc.).[28] However, in most instances item development is still an art practiced by highly experienced professionals. In particular, no one has yet been able to generate other-than-trivial items to test passage-related reading comprehension. Test assembly is still largely an art, too, but there are testing programs (most recently the CPA exams) that use sophisticated linear programming software (see van der Linden, in press) to assign items to tests (actually testlets in the case of the CPA exams).

to scoring, but that is not my focus at this point.

[27]It might be claimed, as well, that scanners are more accurate than human scorers, but that claim is by no means universally acknowledged. Even today, examinees who challenge their scores often request a "hand-scoring" of their answer sheets; seldom do they ask that their answer sheets be scanned again.

[28]Such item generation procedures were discussed decades ago (see, for example, Hively, Paterson, & Page, 1968) before computers were widely available. See Embretson (1998) for a more recent, sophistocated discussion in the area of abstract reasoning tests.

Test Administration

It seems likely that most of the public and many educators think that the primary role of computers in testing is with respect to test administration, which is now generally referred to as computer-based testing (CBT). As far back as the 1960s many researchers predicted that computerized test administration would become common practice in the "near" future. Clearly, such predictions were premature, although they probably will be correct eventually. The principal barriers to widespread use of CBT have not been in the areas of measurement theory or practice; rather they have been cost and/or test volume. To the best of my knowledge, nearly every testing program that has adopted CBT has experienced a dramatic increase in costs, and, except for business environments and the military, almost always these costs are passed on to examinees. Examinees seeking licensure, certification, and occasionally admission to graduate programs have been willing to pay such costs (e.g., examinees taking medical licensure, nursing, architecture, CPA, and GRE tests), but examinees in lower-stakes contexts have been less willing or unwilling to do so. In college admissions testing, CBT is not a "major player," in part because the per-year testing volume (over three million) is so large. There simply are not enough test centers with enough computers to accommodate the volume—at least not yet.

In the K-12 market, it seems to me that every-student CBT will not be viable for years to come, except perhaps in a few districts or small states. Costs, the number of functioning computers needed, the space to accommodate them, and the technical expertise required to maintain them are not going to be available in the near future without a massive increase in school funding, which does not seem likely. Rather, it seems much more likely to me that in the K-12 arena CBT will be used for "niche" testing with carefully selected subgroups of students.

In short, currently CBT is used mainly in licensure and certification, used somewhat in admissions testing, and used relatively little in K-12 (except for occasional "niche" testing). Usage of CBT tends to be positively associated with examinees' ability/willingness to pay increased testing costs and with testing companies' actual or perceived need for security. It is noteworthy that evidence of improved measurement under CBT is relatively rare; many studies focus simply on the extent to which scores are comparable for paper-and-pencil tests and computerized tests. Also, in my opinion, there is more "hype" than reality in much of the enthusiasm surrounding CBT currently, but there are definitely reasons to believe that CBT will become much more prevalent in the future, as discussed next.

First, some of the costs associated with the delivery of major paper-and-pencil testing programs may seem mundane, such as shipping, creation and delivery of score reports, etc. However, these and other costs have been growing and likely will continue to do so. By contrast, computers are becoming cheaper and more ubiquitous. At some point, it seems likely that many paper-and-pencil tests will become "economically-challenged" alternatives, at least in some contexts, regardless of their measurement merits.

Second, it is undeniable that CBT offers the promise of substantial, positive changes in what is tested and how it is tested. How fast this potential will be realized is subject to considerable debate, but even now there are testing programs that are exploring alternative assessment formats for use in CBT. At a bare minimum, scoring is quicker with CBT, as discussed in the next section.

Third, students who now take tests are intimately familiar with computers in many areas of their lives. For them, a computer is a rather natural modality for testing. In the future, not only students but also those responsible for evaluating student performance (teachers, parents, school officials, and the public in general) will view CBT as a natural way to test, I think. Furthermore, they will likely view paper-and-pencil testing as outdated and perhaps "second-rate" no matter what the measurement arguments may be. This rather superficial reason for adopting CBT may be more compelling than some measurement experts would like to believe.

If these speculations are even close to correct, there is reason to believe that the use of CBT will increase substantially in the future.

Item and Test Scoring

As noted at the beginning of this section, in the second half of the last century, the optical scanner had a huge impact on item and test scoring, particularly with respect to speed. Obviously, however, for multiple-choice tests delivered via computer, there is no need for an optical scanner to be a "middle man" between the examinee's responses and scoring—the computer can do that, too. In this sense the future of the optical scanner is to some degree tied to paper-and-pencil multiple-choice testing. Stated differently, as CBT becomes more pervasive, the use of optical scanners in testing is likely to decline, and perhaps begin a trip to extinction, but I doubt that will happen soon.[29] For the near future, economic factors, if nothing else, virtually guarantee that optically-scanned multiple-choice tests will survive.

In the last 20 years there has been an increasing use of various types of performance assessments instead of, or in addition to, traditional paper-and-pencil testing. Essay testing is probably the most prevalent example, but there is a vast range of different types of performance assessments that have been studied and even used operationally. For the most part, these assessments have been delivered in some non-computerized manner, and a persistent and costly problem has been the scoring of such assessments by human raters. Recently, computerized scoring of essays (and other performance assessments) has become a topic of considerable research interest, and there are even testing programs that use computerized scoring operationally at least to some extent. Furthermore, some testing programs (e.g., in architecture and accounting) are already delivering simulations via computer. There are a number of reasons for believing that many future testing programs may be characterized, at least in part,

[29] Even if paper-and-pencil testing were to disappear completely—a very unlikely scenario, scanners would still be needed to scan essays and other forms of constructed responses prior to scoring.

by sophisticated simulations that are scored in real time. This seems to me to be a natural evolutionary result of the intersection of computers and testing. These trends will be costly to support, however.

Score Reporting

Of the six testing tasks noted above, it might be argued that, at least at the present time, computers have the greatest potential for improving testing in the area of score reporting. At the risk of offending many talented testing professionals who work on well-respected testing programs, score reporting has not advanced that much in the last 50 years. For example, it is still common practice for scores on major testing programs to be reported many weeks (or even months) after test administration. This is particularly problematic for grade level testing programs for which test performance is intended to guide instruction. When scores are reported so late, the opportunity for instruction is at best delayed substantially. The crux of the reason that score reporting is so frequently delayed is that it is driven by a paper-laden process that involves scanning, creating and printing of score reports, and delivery of these reports in traditional ways (e.g., mail in one form or another). By contrast, even for paper-and-pencil testing programs, score reports could be delivered much faster over the internet, with appropriate security precautions, of course. This is being done for some programs, but it is still relatively rare.

The principal change that I have seen in score reports in the past several decades is the inclusion of more scores, more details about them, some diagnostic score reporting, and occasionally some rather crude graphical profiles. However, score reports are generally static in the sense that, for the most part, they are dominated by a "one size fits all" approach. There are exceptions. For example, some testing programs provide both a narrative and traditional score reports. However, it is relatively rare for a testing company to tailor the information in its reports to the needs of particular types of examinees and users of scores. There is no technological reason why such tailored score reporting could not be done right now.

Indeed, computer-delivered score reports could be interactive to some extent, allowing the user to "drill down" to a much deeper level of detail to get information such as definitions of types of scores (e.g., percentile ranks) and characteristics of scores (e.g., standard errors of measurement). Such interactive reports could be provided not only at the student level but also at group levels (e.g., classrooms, schools, districts, etc.). My strong suspicion is that computer-delivered, tailored score reports with interactive features would give much more "bang for the buck" than just about any other use of computers in testing. After all, the best test in the world is worthless if the scores are not understood and used properly.

Another frustrating feature of score reports in the K–12 market is the lack of longitudinal information provided to users. Most large K–12 testing programs scale the various levels of the tests so that student-level change can be measured from grade to grade (see Kolen & Brennan, in press). Indeed, the capability of

doing so is a highly touted claim of those who market such tests. Yet, score reports are often merely a snapshot of a student's performance in a particular grade without reference to that student's prior performance. The same type of statement applies to group-level reports. In past decades, when data had to be stored on "flat files," it was difficult and costly to incorporate longitudinal data into score reports, but those days are long past. The use of relational databases makes the provision of longitudinal data an attainable goal. Unfortunately, however, there are still testing programs that are using technology and software that are decades out of date.

Unintended Consequences

The inexorable advancement of computers into testing is not without its perils, however. First, there is always the danger that this technology will overly influence the nature of what is measured.[30] For example, computerized item generation is extraordinarily attractive largely because it substantially reduces test development costs. However, not all knowledge and constructs are amenable to being tested using item clones, at least not yet. There is a danger, therefore, that tested knowledge and constructs could be twisted to accommodate the capabilities of item generation. It is reasonable to reflect on Marshall McLuhan's dictum that "the medium is the message." If that happens to measurement, it will be a step backwards.

Second, it is probably inevitable that the public will believe that tests developed and/or delivered by computer will be "state-of-the-art" and, therefore, less fallible, in both a lay sense and a more technical measurement sense. By no means is that necessarily certain. "Garbage in, garbage out" is still an applicable aphorism. Furthermore, I am especially concerned that some of my colleagues are willing to assume that new procedures for test assembly will be so successful that equating of test forms (see Kolen & Brennan, 1995) will no longer be necessary. Such an assumption seems to me to be highly questionable; at a minimum, it needs to be challenged with real data in real testing programs. In *2001: A Space Odyssey*, Hal (the computer) comments about his role in the space mission by stating with confidence, "We are, by all practical definition of the words, foolproof and incapable of error." The ending of the movie demonstrates Hal's fallacy and fallibility. Unfortunately, however, in the area of measurement Hal's confident statement is sometimes accepted without enough healthy skepticism and critical appraisal.

Litigation and Notions of Fairness

The courts have been playing an ever increasing role in the use of tests, especially in areas that the measurement community would consider primarily issues of fairness. Not surprisingly, the increased involvement of the courts seems to

[30] A case can be made, I think, that in past decades (and perhaps even now) the optical scanner's capabilities reinforced the use of objectively-scored multiple-choice items.

coincide to a considerable degree with high-stakes use of tests. In commenting on these matters, I note that I am not a legal scholar. Consequently, my observations and views may be imperfect (or worse!) from a legal perspective. Still, it seems obvious to me that any serious overview of evolutions in testing must recognize the direct and indirect role of the courts.

Since the early 1970s actual or threatened litigation involving testing has become relatively common. It is not so much that the number of court cases has been huge; rather, the bases for litigation have been many and the impact has been considerable. Directly or indirectly, a number of aspects of testing have been subjected to some form of legal scrutiny. Among the federal laws that have been used as a basis for legal arguments are the U.S. Constitution (particularly the 14th Amendment), the Civil Rights Act of 1964, Section 504 of the Rehabilitation Act of 1973, the Individuals with Disabilities Education Act of 1991, and the Americans with Disabilities Act of 1994.[31]

A Sampling of Some Cases/Issues

Among the issues that have been addressed in cases that have been decided by the courts are: the use of content validity evidence in defending teacher licensure tests; allegations of reverse discrimination based in part on test scores; the use of IQ tests as a basis for placing students in classes for the educable mentally retarded; the use of a basic skills test to award/deny high school diplomas to African-American students who had attended segregated schools; and allegations of racial discrimination in a high school graduation test.

In addition, there are many matters that began as legal challenges (or the threat of such) but were ultimately settled out of court, one way or the other. The following is but a partial list.

- There have been numerous challenges involving cheating, other types of violation of test security, and copyright infringement.
- In the 1970s and 1980s there were efforts by various states and organizations to force companies involved in admissions testing to release test items used to determine an examinee's score. If these efforts had succeeded to their fullest extent, it would have been virtually impossible to equate test forms, which would have meant that testing companies could not have given assurances that scores earned on different forms were comparable. To protect the integrity of their testing programs, while still bowing to the spirit of their critics' demands, most admissions testing companies decided to release many (but not all) test forms shortly after they were used. This strategy was not optimal from a measurement viewpoint, and it necessitated a substantial and costly increase in test development, but this self-imposed settlement seemed to satisfy most of the critics.
- In Breimhorst v. Educational Testing Service (ETS) (2001) the plaintiff challenged the use of a "flag" on his score report for an ETS test

[31] State laws have also provided bases for legal arguments.

taken with extended-time. (A "flag" is simply some designator, such as an asterisk, indicating that an examinee took a test under one or more unspecified, non-standard conditions.) Before the matter went to court, ETS decided to stop flagging extended-time scores for any of *its* testing programs. This did not entirely resolve the matter, however, because the best-known test administered by ETS—namely the SAT—is owned by the College Board, which was not immediately willing to endorse ETS' shift in flagging policy. Therefore, as part of the settlement the College Board (in conjunction with the Disabilities Rights Advocate group) convened a blue-ribbon panel to advise them on the matter. The majority of the committee recommended dropping the flag, although a minority disagreed.[32] The College Board eventually decided to drop flagging examinee scores obtained under extended time, and almost immediately ACT (the College Board's competitor in college admissions testing) followed suit.

This last example is particularly illustrative of the extent to which merely raising a legal challenge relative to a seemingly narrow issue can have far-reaching consequences. First, even without the force of law (there never was a legal ruling, only a threat of one), three of the largest and most visible testing organizations in the world adopted a dramatic change in policy that each of them had vigorously defended in the past. Second, in my opinion, failing to flag examinees who are granted extended time effectively (and perhaps substantially) weakens one leg of the "standardization table" used to support score interpretations, unless, of course, it can be shown that extended-time and standard-time generate comparable scores.[33]

Role of *The Standards* in Litigation

In legal cases and in settlements involving educational tests, considerable weight is generally given to the *Standards for Educational and Psychological Testing* (abbreviated as *The Standards*) developed by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), with the most recent version published in 1999.[34] Still, the legal arena is not bound by *The Standards*, does not accord *The Standards* the same consideration as case law, and does not always concur with the emphases that are implicit or explicit in *The Standards*. This is particularly evident with respect to validity. The courts have not shown themselves to be terribly impressed with complex perspectives on validity; rather, a recurrent theme seems to be the primacy of content validity and predictive validity in the 50 year-old senses of those terms. This one example is illustrative of the fact that there are occasional, serious

[32] This author was part of the minority.

[33] To the best of my knowledge, in most contexts, there is no substantial body of evidence to support this notion of comparability.

[34] The Uniform Guidelines on Employee Selection Procedures (1985) are also sometimes considered, although they are used primarily in the employment arena. Teacher testing in one highly visible area in which the Uniform Guidelines might be considered relevant.

disconnects between the courts and *The Standards*. These disconnects may be exacerbated by the lack of a professional mechanism for enforcing *The Standards*, although the *The Standards* do have a kind of ethical imperative (see, for example, the NCME, 1995, *Code of Professional Responsibilities in Education*).

The measurement community sometimes forgets that the laws that are typically used in legal challenges to tests and testing practices are typically not laws that were created primarily to address testing issues. Rather, they tend to be laws deeply embedded in the American legal system that address what are viewed to be fundamental rights of citizens—rights that were often achieved only after intense political debate. For this reason, we should not be too surprised when legislatures and courts do not accord the degree of primacy to measurement standards and principles that measurement experts might prefer. In short, the tension between legislatures and courts vis-a-vis the measurement community has increased in recent decades and is likely to become even more pronounced in the future, I think. One example of this tension centers around matters of "standardization," as discussed next.

Tensions Surrounding Standardization

One trend in testing litigation seems to center around arguments that involve tailoring the testing experience, or the decision about a tested examinee, to personal characteristics of the examinee. This trend is understandable in the context of various laws and legal precedents, but it is often at variance with the measurement practice of standardization, which has a two-pronged goal: (a) keep the conditions of measurement the same for all examinees; and (b) use the same standards for making decisions about all examinees. Standardization has been a hallmark of testing for decades, largely because it creates a "level playing field," and in this sense contributes to fairness in testing.[35]

The usual argument against standardization (sometimes made in legal and other forums) is that treating everyone the same is not always equivalent to treating everyone fairly. One frequently cited example is the silly scenario of administering a paper-and-pencil test to a blind person. Measurement concerns about standardization are not blind to the need for such exceptions, but they should not be overgeneralized.

Some Reasons for, and Consequences of, Increased Litigation

Even a cursory review of testing in the legal arena in the last several decades quickly reveals that as the testing stakes increase so does the likelihood of litigation. In our country, this is perhaps inevitable, but it also has a number of

[35] Contrary to many statements in the popular press, standardization is not synonymous with multiple-choice testing. For example, an essay test typically has many standardization conditions such as one or more common essay prompts, a fixed testing time for each prompt, detailed scoring rubrics, etc.

possibly unintended, but usually negative, consequences. For example, litigation almost always has the effect of increasing the cost of testing, although the public may not realize it.

Also, contrary to what might be expected, fear of litigation can be a motivation for *avoiding* good validity studies, because such studies inevitably come to conclusions that have a shade of grey, with a healthy amount of reference to the impact of errors of measurement and alternative explanations. As Cronbach (1980) has stated:

> The job of validation is not to support an interpretation, but to find out what might be wrong with it. A proposition deserves some degree of trust only when it has survived serious attempts to falsify it. (p. 13)

Appropriately qualified conclusions often give ammunition to testing critics. This, I think, is one potential explanation for what I perceive to be a dearth of good, thoughtful, published validity studies in high-stakes testing programs.

The involvement of the courts in testing matters is a trend that seems likely to increase. For example, the extraordinarily high stakes associated with NCLB might well lead to legal challenges, and perhaps unprecedented ones. As noted previously, I am not a legal scholar. Still, it seems to me that NCLB is particularly vulnerable to potential legal challenges that focus on "opportunity to learn." Under NCLB, the stakes are very high for states, districts, schools, and students. Yet, it seems almost self-evident that the resources necessary to attain these extraordinarily high goals are often lacking. If so, do all students truly have an "opportunity to learn" the knowledge and skills that constitute each state's definition of proficient?

Also, the steady progression to computer administration of tests (at least in licensure and certification) may well lead to legal challenges unlike any seen before. For example, one form of CBT is computerized adaptive testing (CAT) in which examinees with different ability or proficiency are administered different sets of items. I think it may be difficult to convince some segments of the public and the courts that CAT gives scores that are equitable for all examinees, no matter how sophisticated the measurement arguments may be. In addition, computerized grading of essays may not be acceptable to the public and the courts for high-stakes tests, no matter how compelling the measurement and cost-savings arguments may be.

Ultimately, in our society issues of fairness are probably bound to involve the joint consideration of various measurement standards along with legal precedents and arguments. It does not appear to me that there is a "gold standard" that applies universally. I suspect our perspectives on fairness in testing will evolve continually, and arguments about the merits of particular testing conditions and practices will continue as well.

Change Catalysts and Impediments

The current revolutions and evolutions in testing are being abetted or impeded by various societal goals, economic considerations, and measurement capabilities or lack thereof. Some of these are discussed in this section.

Testing Stakes

To some extent, the revolutions underway in testing are being driven by increases in the testing stakes (real or imagined). For licensure and certification the stakes have always been very real and very high, but, in a sense, the stakes are even higher now for the simple reason that there are an increasingly large number of professions that are using certification to achieve an enhanced status that generally leads to economic benefits for their members.

The stakes for admission testing vary quite a bit. The stakes are generally very high for admission to professional schools. However, the stakes for college admissions are not uniformly high. The public tends to believe they are high based largely, I think, on the inordinate amount of media attention given to the SAT and particularly its use at elite institutions. This impression is misguided in two senses. First, the ACT is used nearly as frequently as the SAT.[36] Second, in many institutions, the principal uses of the ACT and SAT are to create a data base of information about students and to assist in placement; the SAT or ACT may be used as a gatekeeper to a limited extent, but not to the degree that the public tends to believe. Still, the belief itself, even though it is often unjustified, is a powerful force in our society. It is particularly unfortunate that ACT and SAT scores are so widely used as unqualified measures of institutional quality (see, for example, Pascarella et al., 2004).

Without question, the K–12 accountability movement has engendered the most striking change in testing stakes in the past decade. Prior to that, K–12 testing was regarded as a low-stakes or medium-stakes activity in most cases. By contrast, the testing required by NCLB is definitely high stakes because the consequences are so serious. Apparently, most policy-makers assume that accountability in education can be accomplished only through the imposition of high-stakes testing, although there is no body of evidence known to me to support that assumption. One sometimes-voiced defense for this assumption is the claim that testing works in a business environment. This argument seems to me to be particularly flawed. In a business environment, job applicants who do not possess the necessary qualifications are not hired—i.e., they are not admitted into the workplace. In our society, universal education precludes such an option. All students, no matter what their backgrounds and abilities, are entitled to an education. The business argument would be sensible only if every business hired every applicant no matter how (un)qualified and, furthermore, every business kept all employees no matter how well they performed.

[36] Roughly speaking, the SAT dominates on the coasts and the ACT dominates in the middle of the country.

In my opinion, we need to seriously reconsider the role that testing should play in K–12 educational accountability. When testing becomes high stakes, it is almost inevitable that it will drive instructional decisions, usually by narrowing the curriculum in the direction of emphasizing the content and skills tested. This may be an unintended outcome, but it has real consequences that may not be desirable. To the extent that tests drive instruction, teachers who are closest to students tend to have less influence over what is taught, how it is taught, and how it is assessed.

The Standards

One of the requirements of NCLB is that tests meet accepted professional measurement standards. This is a laudable goal, and the field has an excellent set of standards, the *Standard for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). However, developing tests that meet these standards takes considerable time, talent, and money. It seems unlikely to me that the federal government and the states are willing/able to absorb the costs involved in the dramatic increase in testing required under NCLB. However, even if the money is made available, this does little to shorten the amount of time it takes to develop a good testing program that meets professional standards, and money per se certainly does not immediately solve the shortage of measurement professionals that is discussed more fully later.

It is easy to write test items; it is very difficult to write *good* items, and relatively few of us are really good at it! Indeed, even the best item writers spend much of their time *re*writing and editing test items. This is one reason why the process of developing a good test is an iterative one that takes considerable time—rarely less than three years in most practical environments.[37] The process involves numerous steps: constructing test specifications, writing items, pretesting them, revising and editing items, conducting various types of bias reviews (statistical and judgmental), layout, printing, etc. Further, once a test is created, data must be collected for scaling, norming, equating (if multiple forms are involved), and documenting technical characteristics. Insufficient money can retard the process, but large amounts of money cannot speed it up very much because the steps are primarily linear or iterative—they cannot be done simultaneously.

Shortage of Measurement Professionals

It is well known within the measurement community that virtually all graduate measurement programs are having difficulty attracting U.S. students (see, for example, Sireci, 2000). It is widely acknowledged that the numbers of graduate students are not going to increase rapidly soon—not primarily because of lack of

[37] Development may be accelerated for a single test in a single grade in a single district, for example, but generally non-professionals grossly underestimate the length of the test development cycle.

money, but rather for other reasons, such as competition among academic specialties (e.g., statistics, computer science, etc.) for talented graduate students with quantitative skills, and particularly a lack of recognition among undergraduates that the field of measurement even exists as an academic discipline.

I believe the visibility of measurement as a profession would be substantially enhanced if a course in testing were required for licensing teachers and other K–12 professionals. Doing so would have the added (and perhaps even more important) advantage of educating teachers in assessment procedures that are playing an increasingly important role in their jobs. However, there is great resistance to making licensure conditional on a testing course. Some argue that testing is already treated in various content-area courses; others argue that there are already too many requirements for licensure of teachers. I recognize that both arguments have merit, but not enough, in my opinion, to offset the need for considerably greater knowledge of testing among members of the teaching profession.

Given the relatively small numbers of newly trained measurement professionals coming into the field, and given the increased amounts of testing, it is not surprising that there are many more measurement jobs available than there are persons to fill them. There are two aspects of this problem, however, that I think are sometimes overlooked. First, it seems to me that many high-level jobs in measurement get filled by relatively well-qualified persons who rise through the ranks or move from one job to another, with the latter being quite common these days.[38] It is the entry and middle-level positions that are the hardest to fill, as anyone responsible for recruiting knows all too well.

Second, most measurement positions actually do get filled one way or the other, but the persons filling the positions do not always have the qualifications and experience required to do the job well, at least initially. It is quite rare, I think, for a testing initiative to be abandoned because of a shortage of measurement professionals. The problem, of course, is that the quality of testing programs is likely to be negatively affected if the persons responsible for them are not well trained in measurement.

Score Reporting and Data Management

Another "capacity" problem in the measurement system is the seemingly simple matter of scoring tests and assessments, and generating the associated reports, in a *timely* manner—i.e., quick enough so that the results are available soon enough to be of optimal use.[39] The problem is often particularly obvious when constructed responses must be scored by human readers. However, even for multiple-choice testing programs the time between test-taking and score-

[38] Over a decade ago, Brennan and Plake (1990) did a study for the National Council on Measurement in Education that gave results consistent with this observation. For a more recent study see Patelis, Kolen, and Parshall (1997).

[39] There are a tremendous number of tests and assessments scored each year with very high accuracy. Errors are widely publicized, but they are still extraordinarily rare as a proportion of the total amount of scoring conducted. Indeed accurate scoring is a hallmark of the system!

reporting is often longer than most of us would like, especially when the principal purpose of testing is instructional improvement. As noted in a previous section, computerization holds the prospect of ameliorating this problem, but, in my opinion, we are quite far from the goal of providing scores and associated reports as rapidly as many wish.

Another non-trivial capacity problem, at least for K–12 under NCLB, is the seemingly simple matter of data management. Gathering, storing, relating, processing, and verifying all the data elements needed to meet the requirements of NCLB is a far more complicated and expensive task than the public tends to realize. Furthermore, since the data elements, AYP criteria, and other factors differ by state, each state must be prepared to face this task (or at least parts of it) alone. The technology for dealing with data management is available, but it is costly both in terms of money and time.

Technical Advances

In the last 50 years there have been tremendous advances in technical measurement areas such as validity, reliability, generalizability theory, item response theory, equating, and scaling to name but a few. Many of these advances were documented 15 years ago in the third edition of *Educational Measurement* edited by Linn (1989); the fourth edition of this "bible" for the field is currently under development (Brennan, in preparation).

I am quite confident that, in principle, the field of measurement has advanced enough to support the tremendous changes in testing that are now underway. New developments will be required, of course, but I believe the required technical measurement theories and procedures are largely in place. I am somewhat less convinced that, as a society, we will make as good use of these theories and procedures as we could. Too frequently, in my opinion, political pressures and/or economic considerations "trump" measurement concerns and standards. This is particularly problematic, I think, when the net effect is the appearance of a measurement imprimatur on a particular set of political values or business exigencies.

Concluding Comments

This paper has discussed the author's views on current, important trends in testing that are likely to have far-reaching and long-term consequences. The discussion necessarily involves subjective judgments that may be wrong. First, of course, the importance of the three trends discussed (accountability, computerization, and litigation) may be overstated; or, other equally important or more important trends may have been overlooked. Second, even if the trends have been correctly identified, important aspects of the trends may have been missed or misstated. I would contend, however, that there is definitely an ongoing revolution in educational testing that is consequential now and for the future.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Breimhorst v. Educational Testing Service (ETS), *Settlement Agreement*, Case No. 99-3387 (N.D. Cal. 2001).

Berliner, D. C. (2004, March/April). *If the underlying premise for No Child Left Behind is false, how can that act solve our problems?* (The Iowa Academy of Education Occasional Paper #6). Des Moines, IA: FINE Foundation.

Brennan, R. L. (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice, 20*(4), 6–18.

Brennan, R. L. (Ed.) (in preparation). *Educational measurement* (4th ed.). Washington, DC: American Council on Education and Macmillan.

Brennan, R. L. & Plake, B. S. (1990). Surveys of programs and employment in educational measurement. *Educational Measurement: Issues and Practice*, *10*(2), 32.

Cohen, M. (2002). Unruly crew: Accountability lessons from the Clinton administration. *Education Next, 2*(3), 42–47.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? *New directions in testing and measurement: Measuring achievement over a decade. Proceedings of the 1979 ETS Invitational Conference*, 99-108. San Francisco: Jossey-Bass.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380–396.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.) (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Research Council.

Gallagher, J. J. (2004). No Child Left Behind and gifted education. *Roeper Review, 26*(3), 121–123.

Hively, W., Patterson, H. L., & Page, S. (1968). A universe-defined system of arithmetic achievement tests. *Journal of Educational Measurement, 5*,

275–290.

Goals 2000: Educate America ACT (1994). H.R. 1804.

Jones, L. V., & Olkin, I. (2004). *The nation's report card.* Washington, DC: American Educational Research Association.

Kaplan, S. N. (2004). Where we stand determines the answers to the question: Can the No Child Left Behind legislation be beneficial to gifted students? *Roeper Review, 26*(3), 124–125.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

Kolen, M. J., & Brennan, R. L. (in press). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Linn, R. L. (Ed.) (1989). *Educational measurement* (3rd ed.). New York: American Council on Education and Macmillan. (Currently published by Greenwood).

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher, 32*(7), 3–13.

Linn, R. L. (2003, Winter). Requirements for measuring adequate yearly progress. *CRESST Policy Brief 6.* (National Center for Research on Evaluation, Standards, and Student Testing [CRESST], University of California, Los Angeles.)

Linn, R. L., Baker, E. L., & Herman, J. L. (2003, Winter). Alternative approaches to measuring adequate yearly progress. *The CRESST Line*, 4–6. (Newsletter of the National Center for Research on Evaluation, Standards, and Student Testing [CRESST], University of California, Los Angeles.)

McLaughlin, D., & Bandeira de Mello, V. (2002, April). *Comparison of state elementary school mathematics achievement standards using NAEP 2000.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Musick, M. D. (1996, July). *Setting education standards high enough.* Atlanta, GA: Southern Regional Education Board.

National Assessment Governing Board (NAGB) (2002, March). *Using the National Assessment of Educational Progress to confirm state test results.* Retrieved from `http://www.nagb.org/pubs/color_document.pdf`

National Commission on Excellence in Education (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office.

National Council on Measurement in Education (1995). *Code of Professional Responsibilities in Education*. Washington, DC: Author.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Pascarella, E. T., Cruce, T., Wolniak, G. C., Kuh, G. D., Umbach, P. D., Hayek, J. C., et. al. (2004, April). *Institutional selectivity and good practices in undergraduate education: How strong is the link?* (CASMA Research Report No. 2). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Patelis, T., Kolen, M. J., & Parshall, C. (1997). Surveys of programs and employment in educational measurement. *Educational Measurement: Issues and Practice*, *16*(3), 25–27.

Pellegrino, J. W., Jones, L. R.,& Mitchell, K. J. (Eds.) (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. Committee on the Evaluation of National and State Assessments of Educational Progress, Board of Testing and Assessment. Washington, DC: National Academy Press.

Peterson, J. J. (1983). *The Iowa Testing Programs: The first fifty years*. Iowa City, IA: University of Iowa Press.

Roorda, M. (2004, February). *Mega trends: Technology in testing.* Keynote address at the Annual Meeting of the Association of Test Publishers, Palm Springs, CA.

Sireci, S. G. (2000). Recruiting the next generation of measurement professionals. *Educational Measurement: Issues and Practice*, *19*(5), 5–9.

van der Linden, W. J. (in press). *Linear models for optimal test design*. New York: Springer-Verlag.

Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607 et seq. (1985).