Extended Comments on the Draft *Standards for Educational & Psychological Testing*
(But, in particular, Draft Chapters 9, 12, & 13)

to the Management Committee, American Psychological Association, National Council on Measurement in Education, and American Educational Research Association

April 10, 2011
New Orleans, LA

Richard P. Phelps

The *Standards for Educational and Psychological Testing*[1] are a set of guidelines for developing and administering tests. In the absence of any good alternative they have been used by the courts as semi-official codes of conduct. Thus, they have profound impact beyond the boundaries of our relatively tiny community of testing experts. Their content and character are not just our business, they are everyone's business. The call to participate in the 2011 revision of the *Standards* includes this phrase:

> "...[the version of the *Standards*] published in 1999, has been remarkably successful in terms of distribution and impact. Nearly 100,000 copies have been purchased, and the book has become recognized as the final arbiter of legal, ethical, and substantive issues in educational and psychological testing."

The 1999 version runs about 200 pages. If the draft version of the 2011 revision survives intact, it will be substantially larger. Most of the draft *Standards* are in terrific shape, in my opinion; but a few sections are in execrable condition. **The critical comments below apply predominantly to the education policy aspects of chapters 9[2] and 12, and to all of chapter 13.**

---

[1]American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME).

[2]Chapter 9 seems to have been written with only psychological testing in mind and needs to be adjusted to include consideration of the implications for educational testing.

My main issues overall with these three chapters are: (1) what I perceive to be a condescending attitude toward test users (e.g., politicians and the public); (2) a suggested expansion of work requirements so large that most school systems will not be able to afford them; and (3) an implicit or coincidental trade between test developers and the anti-testing professorate whereby test developers turn a blind eye toward the institutionalization of flawed research in return for a large increase in work (and revenue).

It is important for us—the members of the three associations responsible for the *Standards*—to remember that tests are not our private property. It is our place to comment on psychometric issues and provide guidance to make test development and administration reliable, valid, and fair. It is not our place to impose our social norms on the rest of the world, nor is it our place to insist that the rest of the world must interpret terms and conditions as we would prefer they interpret them. The *Standards* should be written so that a wide audience of citizens can clearly understand their meaning. They should be written for them; much of these three draft chapters, unfortunately, is written for us or, rather, a certain segment among us.

Furthermore, some parts of these three draft chapters insinuate (allegedly limited) knowledge and (allegedly misguided) motives for others (particularly politicians), yet seem to assume that all of us are inerrant saints. They have biases; we don't. They often believe in falsehoods; but we never do. We insist that those who "impose" or use tests meet high quality, equity, and ethical standards. But, we do not always demand the same of ourselves.

Whether or not we still deserve it, for the time being many in society consider us to be honest and competent scientists. It would be better for all if that trust were cherished rather than exploited to our advantage.

I characterize these three chapters of the draft *Standards* in five ways:

1) Lack of humility.
Some of the psychometricians I most respect tell me that they no longer bother to comment on draft documents because their experience tells them that, because they are outside the community that tends to control these types of activities, their suggestions are routinely ignored. They are smart and clever people, and the draft would have benefitted from their contributions, but they have been socialized not to bother anymore, just as, I assume, I will be after my precious six minutes.

Nonetheless, despite the unrepresentative character of the author group, two of the three draft chapters in question aggressively demand time, expenditure, and behavior from others, and in particular from our poor, forlorn state and local government agencies. To borrow a Jessie Jackson phrase, these two particular draft chapters might be described as "reverse Robin

Hood" manifestos, with prosperous Norman psychometricians demanding ever higher taxes from their already-impoverished state and local education agency Saxon serfs.

2) Conflicts of interest.
Two of the three chapters recommend an expansion of work requirements in education so vast that most school systems will not be able to afford them (and little of the expansion would produce information directly useful to educators but, rather, more background technical studies and reports of interest to psychometricians). It so happens that the authors of these two chapters are in positions where they stand to gain professionally and financially from their own recommendations.

3) North American focus.
The U.S. and Canadian education systems differ greatly from those in the rest of the world in both their governance and systems of assessment. No one from outside the U.S. and Canada is involved in the creation of the *Standards*. Yet, our colleagues overseas sometimes must defend themselves and their programs from attacks that cite the *Standards*. At least some of them would like our three organizations to express far more deliberately that these *Standards* apply ONLY to the United States and, at most, also Canada, and should NOT be presumed by anyone to apply in any meaningful way to other jurisdictions.

We all do believe that, don't we? Any plan for this draft to be considered an international standard would be inappropriate and scary.

4) The extra large influence of special people in our community.
The lifeblood of a healthy scholarly body is free and open inquiry and debate. The best scholarship is not withheld from scrutiny but, instead, embraced by it. In fully-functioning scholarly communities, no research conclusions are considered valid unless and until they respond to challenges.

But, within our community are special people whose research is not open to challenge. These special people simply ignore contrary evidence and any who disagree with them; sometimes they even declare enormous research literatures nonexistent. Some leaders in our community have enabled these behaviors for over two decades, usually by turning a blind eye, but sometimes by themselves directly suppressing opposing evidence and points of view.

In the 1980s and 1990s these special people published studies purporting to prove that "stakes" and stakes alone cause artificial test score gains (i.e., test-score inflation). They argue that absent stakes, a test's scores and score trends are "natural", valid, trustworthy, and "uncorrupted." "All high-stakes tests are corruptible," they say.

Astonishingly, the high-stakes-and-only-high-stakes-cause-test-score-inflation theory has been

adopted as dogma. Efforts to dispute it have been met with insults, ridicule, or worse. Even the naive (about testing) economics and political science researchers who have taken to studying the effect of testing this past decade have incorporated it into their work.[3]

One rhetorical device employed by these special people in their research might be called the "floating definition." As they write up a study they ascribe an unfamiliar definition to a key term, in the fine print as it were. Then, when the study is read, most readers assume that they have defined the term the standard way, and interpret the study conclusions accordingly.

Few readers seem to realize that these special people used then and still use today an antiquated definition of the term "high-stakes" when categorizing tests and, thus, when ascribing the effects of tests. By their definition, one ascribes "high stakes" to any test for which: (1) "teachers feel judged by the results"; (2) "parents receive reports of their child's test scores"; or (3) "test scores are widely reported in the newspapers." These days, this definition subsumes virtually any large-scale test, including most of those commonly considered to be "no stakes." There was a time before most of us were born when educators could withhold test results from parents and the public whenever they pleased, but that day has long since passed.

All but one of those tests producing test-score inflation back in the 80s and 90s were no- or

---

[3]The National Assessment of Educational Progress (NAEP) is a federally funded "no-stakes" monitoring assessment. Its scores and score trends have become a favorite of researchers over the past decade to "audit" or benchmark the score trends on state standards-based tests. The rising scores on state standards-based tests are compared with the relatively flat scores on the NAEP and declared to be "inflated." The NAEP has been lauded by these researchers as a terrific candidate for "auditing" due to its "low stakes."

Ironically, according to the definition of "stakes" employed by our special people, the NAEP is actually a high-stakes test--teachers do feel judged by its results and its results are widely reported in the newspapers. Our special people wish to have it both ways -- call a no-stakes test a high-stakes test when it suits their argument and call a high-stakes test a no-stakes tests when that suits their argument.

Regardless, if one were determined to use an "audit test" despite the labyrinthine obstacles to the validity of the exercise, the NAEP would serve as a better candidate than most tests. Why? The researchers using the NAEP as an audit test cite its "low-stakes." But, the NAEP serves relatively well as a steadfast benchmark of score trends because it is externally administered, all the way down to the classroom level. Independently hired administrators not affiliated with the local school district manage all aspects of the test administration leaving not a single opportunity for anyone with a vested interest in the results to manipulate the process. It is the external control of the test administration that is important, not the low stakes. Besides, as time goes by, the stakes of the NAEP are rising.

low-stakes tests by any accepted current definition and, indeed, by the *Standards'* own past, current, and proposed definitions (see draft Glossary, lines 284–285, 359–360). Nonetheless, the antiquated, overly-broad definition of "high stakes" is resurrected on the first page of draft chapter 13 (lines 31–42) and then assumed throughout that chapter.

Anyone genuinely familiar with large-scale test administration must realize that the high-stakes-and-only-high-stakes-cause-test-score-inflation theory makes little sense. They know how casually many no-stakes tests are administered, and how slack their test security can be. After all, when tests have no stakes they "don't count"—so why bother to pay for high levels of security, or for frequent test form and item rotation? Indeed, the "Lake Wobegon" scandal in the 1980s occurred with no-stakes tests, with administrators using identical test forms year-after-year, as they could given the absence of test security protocols.

Sure, high-stakes tests incentivize dishonest educators to cheat (when the stakes apply to them). But, no-stakes test administrations typically do nothing to discourage cheating (and little to maintain the integrity of test materials). Moreover, the higher the stakes, the more likely cheaters will be caught (because security is more likely to be high). With most no-stakes tests, the likelihood of catching cheaters is close to nil.

Low-stakes tests and score trends tend to be less stable, less secure, less aligned to content standards, and locally administered, with less frequently rotated test forms and items (if forms and items are rotated at all).[4] Draft chapter 13 recommends relying on them to "audit" high-stakes tests and score trends. With any high-stakes test subject to audit by any low-stakes test, its perceived quality is determined entirely by the low-stakes test. Indeed, those who oppose high-stakes testing could add an easily manipulated and unmonitored low-stakes test and tailor it to discredit score gains on their jurisdiction's externally-mandated and monitored high-stakes test.

Score gains on the highest-quality high-stakes tests with the tightest security and most frequent test form and item rotation will be labeled "artificial." Indeed, the perceived quality of any high-stakes test will be no better than the actual quality of the worst low-stakes test used to audit it.

Imagine this scenario. A jurisdiction has a standards-based high-stakes test with tight security

---

[4]The draft standards themselves—including, ironically, draft chapter 13—frequently acknowledge the problems of no/low-stakes tests, such as the volatility of their score trends, unpredictably varying degrees of test-taker motivation, low or no integrity of test materials, and overly-light sampling schemes. See, for example, chapter 9, lines 350–352 and chapter 13, lines 75–85, 196–202, and 400–404.

and 50% item rotation, and gradually rising average scores. Critics then claim that the score rise is simply "inflation" and "artificial." So, the jurisdiction implements a marginally aligned, low-cost, no-stakes "audit test." What will happen? If the audit test is administered with tight security and ample item rotation, its scores on the unaligned content may not rise, allegedly proving that the high-stakes score rise is "inflated." Or, if the audit test is administered with lax security and no item rotation, its scores could rise at a rate higher than the high-stakes test scores, suggesting that student achievement is actually declining. It is difficult to imagine a scenario in which this "auditing" will produce unambiguously accurate information.

The perspective of draft chapter 13 is rooted in a lost world. In the 1960s and before, state tests were uncommon and many school districts decided on their own to administer large-scale exams themselves and, if they did administer, to release or not release the results, as befit their interests. For the authors of draft chapter 13, a "valid and uncorrupted" exam is one for which the results are not necessarily made public and for which local educators exert total control, from test preparation to possible press release interpretations. Any other type of exam, including virtually any large-scale test known today, is corrupted and unnatural.

This describes an outdated mind set. Granted, many criticisms of educators are unfair and, it seems, even more so lately. But, so, too, are many of the criticisms of any other occupational group.

Some of our community's leaders continue to expect that we educators deserve the right to control taxpayer access to, and even their interpretation of, any education information and, in particular, any information that could possibly be used to judge our performance. I cannot imagine another public service occupation in the year 2011 with leaders who that still think like this—that would consider it an "unnatural" "imposition" to provide parents information about their own children's progress or would characterize public release of a public agency's aggregate statistics a "corruption" of those measures. Nonetheless, these expectations of a return to a lost age when education administrators had far more arbitrary power have been written into draft chapter 13.

5) Abnegation of responsibility.

Educators need a lot of help with test security issues and we give them almost none. Indeed, our predominant answer is to not "impose" tests with stakes at all, because they are "unnatural" and encourage cheating.

Recall this old joke:

> A doctor receives a patient in an examining room and inquires as to the problem. The visitor gesticulates with one arm and says, "Doctor, it hurts when I do this." The doctor replies, "Then, don't do that."

Indeed, most of us have doctoral degrees. But, despite all of our many accomplishments and scientific advances, this is essentially all we have to say to classroom educators on the topic of test security.

Arguably, standards and guidance on test security protocols have been for over a quarter century and remain today the single most profound need of education agencies. And, what to we have to tell them? "Then, don't do that." Don't administer tests with stakes because people will cheat and score gains will be meaningless.

I offer a rough outline of some useful test security standards that could be written, borrowing heavily from a list composed by John J. Cannell in 1989:

enact *and enforce* formal, written, and detailed test security and test procedures policies;

formally investigate all allegations of cheating, *and prosecute* the offenders;

ensure that educators cannot see test questions either before or after the actual test administration and *enforce consequences* for those who try;

prohibit test administrators from looking at the tests even during test administration;

prohibit any adult with a self interest in the results from a testing room (i.e., teachers should not administer tests to their own students, if those results may directly affect the teacher);

use outside test proctors;

hold and seal test booklets in a secure environment until test time;

keep test booklets away from the schools until test day;

rotate items and forms annually;

spiral different forms of the same test (i.e., having different students in the same room getting tests with different question ordering) to discourage student answer copying;

employ technologies that reduce cheating (e.g., optical scanning, computerized variance analysis) and *use them* to identify cheaters; and

provide to the authorities responsible for investigating allegations of cheating the

> documentary evidence (e.g., original answer sheets) they need to prosecute their investigations, without cost and without hesitation.

The italicizing above expresses my own belief that we are prevented from doing more to discourage cheating often by our (testing expert) community's own proclivity to avoid controversy at any cost (even, sometimes, at the expense of truth and accuracy).

The bullets above outline a straightforward policy response to test security concerns. Special people in our community, however, would recommend diverting attention and resources in the opposite direction—toward developing a separate, unrelated audit test and other weaker measures of achievement or performance. Resources and attention that could be employed to strengthen test security and confidence in the primary, most salient, most relevant measure would, instead, be diluted among one to several secondary, less salient, less relevant, and less valid measures.

And, all of them internally generated, by the way. In education, results from externally-managed tests are popular with some members of the public precisely because, in many cases, they are the only performance indicators not produced by the education agencies themselves. The prescription of our special people seems to suggest that the public not be allowed access to results from externally-managed tests unless those results are first muddled with information that is less independent and interpreted by the very people with an incentive to interpret the results in a self-serving way.

Policy Implications

When the most far-reaching federal intervention in U.S. assessment policy was being considered, from 2000 to 2002, special people in our community managed to convince policy makers that no relevant research on the effects of tests used for accountability existed to help guide them in their program design. That casually, a century's worth of relevant research on the optimal design of assessment systems and the effects of testing—thousands of studies—that could have helped policy makers design a reasonable, rational, and efficient program was declared nonexistent. The result? The research-uninformed No Child Left Behind Act.

Draft chapter 13 threatens to continue the trend—public policy will be starved of a great mass of useful and relevant information, in favor of the pet theories of a special few and the contrived policy recommendations they ensue. And, just in time for Congressional consideration of the revision of the No Child Left Behind Act.

What is most wrong with this draft chapter is what is most wrong with our community as a whole.

Before we “impose” standards on everyone else, perhaps we should adopt some standards for ourselves. Behavior that in other scholarly groups would be considered unethical is unfortunately quite acceptable in ours. To my observation, the only ones in our community who pay any price for bad behavior are those who dare to suggest that it exists. Voices of protest and dissension are quickly shushed and sometimes doused with character assassination. Communities intolerant of disagreement, such as ours, have little legitimacy in lecturing others; communities that suppress dissension, such as ours, are also incapable of any real progress.

The comments of the humanist Wilfred McClay on the failure of expert economists to anticipate the recent financial crisis are also appropriate to us testing experts:

> “Can we pinpoint the reasons why experts fail, and find means of guarding against such failures in the future? The example of the financial crisis offers clues.... First of all, it reinforces the crucial importance of keeping diversity of opinion and perspective alive in the disciplinary communities — or, given the ideological monocultures that dominate so many of these communities, of introducing such diversity for the first time in memory. It shows us why the dominant paradigms and "normal science" of the day must be kept from exerting a tyrannical control over the community of experts. The lazy and dishonest argument so often heard — that [one] who has original views is "not in the mainstream" of the field — ought to be regarded as a canard unworthy of a genuinely serious community of expert knowledge, and a criterion that does not have the health of that community truly at heart.”

McClay thinks economists have an intolerant ideological monoculture. But, theirs might be considered a sky of bright transparent clarity compared to ours.

For the sake of our integrity, draft chapter 13 must be deleted in its entirety. For the sake of the financial solvency of our public schools, draft chapter 12 should be cut by a third. For the sake of our character, the language in draft chapter 9 should be revised. Finally, if we want the *Standards* to be more useful to educators, we should thoroughly confront test security issues; at this point, we seem unwilling to admit that they exist.

Thank you for your time and attention.

References

Cannell, J. J. (1989). *How public educators cheat on standardized achievement tests*. Albuquerque, NM: Friends for Education.

McClay, W. M. (2009, Fall). What do experts know? *National Affairs, 1*, 145–159. Available at http://www.nationalaffairs.com/publications/detail/what-do-experts-know

Phelps, R. P. (2008/2009). Educational achievement testing: Critiques and rebuttals. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing*, Washington, DC: American Psychological Association.

Phelps, R. P. (2008/2009). The rocky score-line of Lake Wobegon. Appendix C in R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing*, Washington, DC: American Psychological Association. Available at: http://supp.apa.org/books/Correcting-Fallacies/appendix-c.pdf

Phelps, R. P. (2009, November). Worse than plagiarism? Firstness claims and dismissive reviews. (slide show). *Nonpartisan Education Review / Resources.* Available at: http://nonpartisaneducation.org/Review/Resources/WorseThanPlagiarism.htm

Phelps, R. P. (2010, July). The source of Lake Wobegon [updated]. *Nonpartisan Education Review / Articles*, 1(2). Available at: http://nonpartisaneducation.org/Review/Articles/v1n2.htm

Phelps, R. P. (n.d.). Censorship has Many Fathers: A progeny of excuses for censoring information that one dislikes. *Nonpartisan Education Review / Foundation / CensorshipHasManyFathers*. Available at: http://www.nonpartisaneducation/Foundation/CensorshipHasManyFathers.htm

Staradamskis, P. (2008, Fall). Measuring up: What educational testing really tells us. Book review, *Educational Horizons, 87*(1). Available at: http://nonpartisaneducation.org/Foundation/KoretzReview.htm