8

SUMMARY AND DISCUSSION

RICHARD P. PHELPS AND LINDA S. GOTTFREDSON

There are so many criticisms of standardized testing that some of them must be true.

The foregoing statement is fallacious, but there is one truth within it: Criticisms of standardized testing are myriad. However, most are simply false, and many turn the truth upside down. The chapters in this book have not shied away from pointing to testing's limitations, but they document how most criticism today avoids the evidence on testing altogether and instead conjures inconsistent but mutually reinforcing falsehoods meant to discredit the entire enterprise, from intent to impact. Exhibit 8.1 lists the fallacious criticisms described in this book, by chapter.

When one surveys the seeming multitude of fallacies across the various types of test use, however, many begin to look alike. Indeed, some of the apparent differences may reflect only variations in terminology. For example, the idea that psychological diagnostic (chap. 2) and employment (chap. 5) tests are easily faked differs only slightly from the idea that college admissions tests (chap. 4) are easily gamed, say, through test-preparation coaching. In each case, the tests are accused of invalidity because they are allegedly subject to manipulation.

For convenience's sake, we classify the fallacious criticisms of standardized tests into three groups:

The views expressed here are the authors' own and not necessarily those of ACT, Inc.

EXHIBIT 8.1
Fallacious Criticisms of Testing, by Chapter

CHAPTER 1: INTELLIGENCE TESTS

Test-Design Illogic

1. Yardstick mirrors construct.
2. Intelligence is a marble collection.

Test-Score-Differences Illogic

3. Nonfixedness proves malleability.
4. Improvability proves equalizability.
5. Interactionism (gene–environment codependence) nullifies heritability.
6. Similarity (99.9%) negates differences.

Test-Validation Illogic

7. Contending definitions negate evidence.

Causality Illogic

8. *Phenotype* equals *genotype.*
9. *Biological* equals *genetic.*
10. *Environmental* equals *nongenetic.*

Standards-of-Evidence Illogic

11. Tests are useful only if perfect.
12. Reject so-called dangerous conclusions until proved beyond all possible doubt.
13. Accept happy speculation until conclusively disproved.

CHAPTER 2: PSYCHOLOGICAL DIAGNOSTIC TESTS

1. Tests are too expensive.
2. Tests are not valid.
3. Tests lack cultural generalizability.
4. Tests are peculiar and irrelevant.
5. Tests may be (easily) faked.

CHAPTER 3: EDUCATIONAL ACHIEVEMENT TESTS

1. You can't stop progress.
2. Tests cost too much.
3. High stakes induce artificial test score increases.
4. There is little evidence of the effects of testing and no evidence of its benefits.
5. Mischaracterized because tests are difficult to understand.

CHAPTER 4: COLLEGE ADMISSION TESTS

1. Results are misused.
2. Original history of the SAT discredits current use.

3. Test items are biased.
4. Tests are neither valid nor useful.
5. Tests are just a proxy for socioeconomic status.
6. Tests can be coached.
7. An increasing number of colleges are making admission tests optional.

CHAPTER 5: EMPLOYMENT TESTS

Fairness Issues

1. Tests are unfair to ethnic and racial minorities.
2. Tests are illegal when ethnic and minority groups obtain lower scores on average
3. Tests are unfair because they have not been validated for every type of job and context.
4. Tests invade privacy.

Accuracy Issues

5. Tests result in some bad hiring decisions.
6. Tests tell nothing one would not learn by talking with the applicant.
7. Tests often assess the wrong content.
8. Tests can be easily faked.

Administrative Efficiency Issues

9. Tests are unnecessary because one can train anyone to do any job.
10. Tests take too long.
11. Tests are too expensive.

CHAPTER 6: LICENSURE AND CERTIFICATION TESTS

1. Tests do not measure important skills for success in a profession.
2. Tests are predominantly "multiple guess" and basically measure test-taking skills.
3. Passing scores are designed to limit access to a profession.
4. Tests are used mainly in the United States.

CHAPTER 7: LARGE-SCALE COGNITIVE DIAGNOSTIC TESTS (LS-CDTs)

1. Tests cannot be cognitively diagnostic because they measure lower level, basic knowledge and skills.
2. LS-CDTs will narrow the curriculum because they focus on a narrow set of processing skills.
3. LS-CDTs inform common problem solving and not less well-known problem solving.
4. Verbal reports are untrustworthy.
5. Abilities tend to be correlated and so will essentially produce uninformative profiles.
6. *Large-scale cognitive diagnostic test* is simply a new name for *formative assessment.*
7. LS-CDTs essentially involve the reporting of subscores.
8. LS-CDTs are the cure for high-stakes testing.

- Tests are not valid (e.g., they are not "natural" or "authentic"; they only measure some simple and low-level mental processes; they narrow curricula; their scores do not represent anything real).
- Tests do not work well (e.g., they are too long and expensive; they can be easily gamed, faked, and coached; they produce inaccurate results; they are inferior to readily available alternatives; parameters such as passing scores are set arbitrarily).
- Tests are not fair (e.g., they are biased against certain groups of people; they have shady historical origins; they only measure socio economic status).

A cross-classification of fallacious criticisms is shown in Table 8.1. The numbers in the cells correspond to the fallacies listed in Exhibit 8.1.

Censorship and suppression of evidence, which abet the widespread promotion of these false notions by self-interested parties, complete the picture of the antitesting epidemic. Yet what allows this phenomenon to persist, year after year, seemingly unchecked by contrary evidence?

SCIENCE AND ADVOCACY

In his foreword to this volume, Thomas Oakland reminds readers that authors were selected to participate in this volume on the basis of their demonstrated commitment to scientific principles, procedures, and goals. One of those goals is to find probable causes and effects. Especially in the social sciences, the attainment of absolute certainty is rare, if not impossible. However, one can—through a systematic review of a research literature, careful observation and measurement, consideration of rival hypotheses, and honest analysis—sometimes determine the most probable cause of an outcome or effect of an action.

Advocacy is different from science, however—as different as advertising is from good journalism or aggressive politicking is from consensual governing. For the zealous advocate, cause and effect are predetermined to serve one's interests. An advocate need not even believe a cause or effect that she claims; her goal is to persuade others to believe it.

An advocate searches not for probable causes and effects but, rather, for merely plausible ones—ones that others are willing to believe. This is not as easy as it may sound, as any advertising executive can attest; successful market research can be tedious and time consuming. Regardless, the desired outcome is neither truth nor understanding, but conversion—getting others to view a situation in a manner that serves one's own interests.

One senses our chapter authors' frustration with the need to respond to the nonscientific nature of much testing criticism. However, that criticism may well have originated as advocacy and been expressed in advocacy fo-

TABLE 8.1
Cross-Classification of Fallacious Criticisms of Testing, by Chapter

	Chapter						
Criticism	1	2	3	4	5	6	7
Tests are not valid, they . . .							
are not natural or authentic and measure the wrong things.	2, 3	4			7	1	3
only measure simple and low-level mental processes.						2	1
narrow curricula.			3				2
produce artificial scores.	4		3		3		7
are not supported by research.	7		1, 4				
Tests are not useful, they . . .							
can be gamed, faked, coached.		5		6	8		
are too long and expensive.		1	2		10, 11		
produce inaccurate, unreliable results.	11		3		5		5
are inferior to the alternatives.	5, 13				6, 9		6
are constructed arbitrarily.	1						4
Tests are not fair, they . . .							
are biased against [fill in the blank].	9	3		3	1, 2		
have shady historical origins.	8, 12			2			
are misused; are arbitrary.	6			1	4	3	7
only measure socioeconomic status.	10			5			

Note. Numbers in cells refer to the numbered lists beneath the corresponding chapter in Exhibit 8.1, this volume, pp. 248–249.

rums (rather than in scientific journals or at scientific meetings). In this context, advocates, and those who believe them, can make any cause-and-effect claims they wish, their persuasion limited only by plausibility—by what audiences unfamiliar with the subject, and the already converted, are willing to believe.

Plausible arguments are then reinforced through repetition. From our observation, this repetition is supported not so much by other professionals or idealists as by self-interested groups, and chief among these may be education researchers and administrators opposed to the use of standardized tests. Research articles on testing can differ dramatically in their conclusions on the basis of their venue, with those in psychology and technical measurement journals more willing to acknowledge positive results in studies of test use than those in education journals and practitioner magazines. Indeed, articles in the latter venues can be unrestrained in their conviction and criticism. Furthermore, as shown in chapter 3 of this volume and Appendix D (see the accompanying Web site to this volume: http://www.apa.org/books/resources/Phelps/), some education researchers have ventured outside their own domain of expertise to condemn the use of testing in other contexts (in this case, employment testing). The allegiance of so many education professionals and, in particular, education professors to the antitesting cause socializes a multitude of new critics and provides numerous venues in

which only those critical of testing are heard and read, and supportive evidence is either ignored or declared not to exist.

Often these critics invoke the welfare of the public, parents, students, teachers, or any group but their own when leveling their criticisms. However, hundreds of polls conducted over the past 4 decades in North America verify solid and unwavering public support for standardized testing in the schools, in the workplace, and for psychological diagnosis (Phelps, 2005). In the area of certification and licensure, pollsters sometimes have difficulty finding more than a negligible proportion of the public opposed. Student support for testing has likewise been solid and unwavering and, until recently, so has teachers'.[1]

EDUCATION'S CENTRAL ROLE IN ANTITESTING CRITIQUES

Why, then, is the education profession and its allied professoriate such a wellspring of antitesting hostility? One answer lies in the democratic dilemma arising from cognitive diversity (see chap. 1, this volume). In no public arena is the dilemma more conspicuous yet more hotly contested than in the public schools, so in no other arena is evidence of cognitive diversity so unwelcome. Anyone or anything that provides unambiguous evidence of this diversity invites passionate rebuttal.

Many social scientists and others have assumed that equal educational chances for children from different social backgrounds would yield equal educational outcomes as well, apparently on the mistaken belief that intellectual talent and academic efforts are equally distributed across all individuals. They therefore wrongly conclude that schools have failed to equalize opportunity when they observe that schools have failed to equalize achievement for all students. Schools are fated not only to fail this latter, social leveling mission but also to underscore the very differences in learning ability and effort that sustain achievement differences.

Specifically, public education puts all students through graduated series of cognitive tasks (in reading, math, science, etc.) that increase in complexity from Grades 1 through 12. This steady escalation of cognitive demands is akin to administering a lengthy and highly public aptitude test battery to each cohort of a nation's children. School performance is influenced by many factors, of course, but none more powerful overall than the combination of ability and effort. Differences in academic achievement track IQ differences between demographic groups, between families, and even between siblings growing up in the same home.

[1]The decline in the proportion of teachers who favor the use of standardized testing in the past several years may be related to the provision of the No Child Left Behind Act that holds schools, but not students, accountable for student performance.

Moreover, American schools are now expected to educate a great diversity of children within the same classrooms through at least middle school, despite some children learning multiple times faster than other children. Although inclusion practices are meant to reduce differences in achievement and the stigma of separation, they make differences in ability and effort all the more conspicuous in the classroom by having students of markedly different academic dispositions work side by side, hour after hour, day after day. Achievement differences can be narrowed in inclusive classrooms only by restricting opportunities for ability and effort to affect observed performance, for example, by restricting how much material is taught or assessed (as do some forms of cooperative learning) or by lowering performance standards or making them less academic (as is the trend in gifted education). Organizations can relax standards only so far, however, before they cross the "point of organizational embarrassment" (Gordon, 1988, p. 84) and trigger outcries for higher academic standards.

Nonetheless, the schooling-related professions generally hold that equality and quality—EQuality—could be achieved simultaneously if they were given sufficient resources. They capitalize on reports of low test scores to justify calls for greater funding but argue that testing otherwise imperils the quality and equality of education: quality, when instruction is distorted by teaching to the test, and equality, when children are labeled or sorted by ability. In fact, they continue, the very notion of an intellectual hierarchy threatens EQuality, and students are better served by a belief in multiple, coequal intelligences whereby they can all be smart in some way—as if reality would follow belief.

As noted in chapter 1 of this volume, it is perhaps ironic that the federal No Child Left Behind Act of 2001 now holds schools accountable for the EQuality that educationists have said it is within their power to produce. Yearly progress toward meeting the Act's mandate to raise all student populations to the same high level of academic proficiency is gauged with state-developed tests in specified subjects and grade levels. Schools that fail to level-up performance on schedule face escalating sanctions, including state takeover. Massive failure looms,[2] as would be predicted by the democratic dilemma, and has prompted some states to create the illusion of progress by lowering the threshold for what counts as "proficiency" on their tests. These illusions are periodically punctured by results from the National Assessment of Educational Progress (NAEP). The U.S. government has administered the NAEP to national samples of students since the 1970s to provide "report cards" on public education. Recent NAEP test results in different academic

[2]Knowledge levels can be raised without increasing *g*, to be sure, but rising averages are usually accompanied by wider gaps in achievement when there is no artificial ceiling on performance (Ceci & Papierno, 2005).

subjects for different grades and demographic groups suggest only scant, spotty, and inconsistent progress toward higher proficiency levels and smaller gaps (Fuller, Wright, Gesilci, & Kang, 2007).[3]

PUBLIC DEBATES ABOUND, BUT ONLY ONE SIDE IS INVITED

The democratic dilemma helps address the main theme of this chapter: Why are there so many criticisms of testing? Furthermore, how have fallacious claims come to rule policy debates about testing? The answer to the second question lies partly in the way that any advertising campaign might succeed. That is, we believe that many antitesting advocates will try any argument that works—any argument they can persuade others to believe. They keep trying until their efforts are successful. The result is as many criticisms as audiences informed by only one side of the debate will accept.

Why, however, do antitesting advocates seem to succeed more easily than advertisers who seek to capture the market for their products? We suggest this answer: The critics play by different rules than do their competitors. Specifically, debate tactics differ between scientists and advocates. Scientists seek the scrutiny of their peers to confirm (or deny) the value of their work. Advocates may wish to avoid scrutiny, especially when selling happy falsehoods. Scientists do not circumvent the research literature but engage it. They must respond to rival hypotheses with counterevidence, not innuendo. Scientists confront conflicting scientific results, whereas advocates may simply ignore them or, as described in some chapters, repackage advocacy to look like superior science.

Indeed, as several chapter authors in this volume confirm, it has become common for testing opponents to declare nonexistent an enormous research literature that contradicts their claims. Moreover, with the help of the Fourth Estate, they seem to be fairly successful in eradicating from collective memory thousands of studies conducted by earnest researchers over the course of a century.

The easiest way to win a debate is by not inviting the opponent. The critics rightly fear an open, fair scientific contest.

REFERENCES

Ceci, S. J., & Papierno, P. B. (2005). The rhetoric and reality of gap closing—When the "have-nots" gain but the "haves" gain even more. *American Psychologist, 60*(2), 149–160.

[3]Trends in NAEP scores provide an imperfect benchmark for trends in the scores on state tests because NAEP's content standards are not aligned with those in most states. Indeed, the degree of alignment can vary quite a lot from state to state, with some state tests far less aligned to NAEP than others. Comparisons between NAEP and state test scores become more valid when adjustments are made for differences in content standards. For more information, see Zenisky, Hambleton, and Sireci (2008a, 2008b).

Fuller, B., Wright, J., Gesilci, K., & Kang, E. (2007). Gauging growth: How to judge No Child Left Behind? *Educational Researcher, 36*(5), 268–278.

Gordon, R. A. (1988). Thunder from the left [Review of *Storm over biology: Essays on science, sentiment, and public policy*]. *Academic Questions, 1*, 74–92.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 107th Congress, January 8, 2002.

Phelps, R. P. (2005). Persistently positive: Forty years of public opinion on standardized testing. In R. P. Phelps (Ed.), *Defending standardized testing* (pp. 1–22). Mahwah, NJ: Erlbaum.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2008a, July). *Communicating the utility of NAEP score reports*. Paper presented at the Sixth International Test Commission Conference, Liverpool, England.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2008b, July). *Customizing the view: Evaluating the communication of national assessment results*. Paper presented at the Sixth International Test Commission Conference, Liverpool, England.

Correcting Fallacies About Educational and Psychological Testing

Edited by

Richard P. Phelps

2009

American Psychological Association • Washington, DC