Gottfredson, L. S. (2009). Logical fallacies used to dismiss the evidence on intelligence testing. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 11--65), American Psychological Association. https://doi.org/10.1037/11861-001

1

LOGICAL FALLACIES USED TO DISMISS THE EVIDENCE ON INTELLIGENCE TESTING

LINDA S. GOTTFREDSON

> Human intelligence is one of the most important yet controversial topics in the whole field of the human sciences. It is not even agreed whether it can be measured or, if it can, whether it should be measured. The literature is enormous and much of it is highly partisan and, often, far from accurate. (Bartholomew, 2004, p. xi)

Intelligence testing may be psychology's greatest single achievement, but it is also among its most publicly reviled activities. Measurement technology is far more sophisticated than in decades past, but antitesting sentiment has not waned. The ever denser, proliferating network of interlocking evidence concerning intelligence is paralleled by ever thicker knots of confusion in public debate over it. Why these seeming contradictions?

Mental measurement, or *psychometrics*, is a highly technical mathematical field, but so are many others. Its instruments have severe limitations, but so do the tools of all scientific trades. Some of its practitioners have been wrongheaded and some of its products misused, but this does not distinguish mental measurement from any other expert endeavor. The problem with intelligence testing, one suspects, is that it succeeds too well at its intended job.

HUMAN VARIATION AND THE DEMOCRATIC DILEMMA

IQ tests, like all standardized tests, are structured, objective tools for doing what individuals and organizations otherwise tend to do haphazardly,

informally, and less effectively: assess human variation in an important psychological trait—in this case, general proficiency at learning, reasoning, and abstract thinking. The intended aims of testing are both theoretical and practical, as is the case for most measurement technologies in the sciences. The first intelligence test was designed for practical ends—specifically, to identify children unlikely to prosper in a standard school curriculum, and indeed, school psychologists remain the major users of individually administered IQ test batteries today. Vocational counselors, neuropsychologists, and other service providers also use individually administered mental tests, including IQ tests, for diagnostic purposes.

Group-administered aptitude batteries (e.g., Armed Services Vocational Aptitude Battery [ASVAB], General Aptitude Test Battery [GATB], SAT) have long been used in applied research and practice by employers; the military; universities; and other mass institutions seeking more effective, efficient, and fair ways to screen, select, and place large numbers of individuals. Although not designed or labeled as "intelligence tests," these batteries often function as good surrogates for them. In fact, all widely used cognitive ability tests measure general intelligence (the general mental ability factor, *g*) to an important degree (Carroll, 1993; Jensen, 1998; Sattler, 2001).

Psychological testing is governed by detailed professional codes (e.g., American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Society of Industrial and Organizational Psychology, 2003). Developers and users of intelligence tests also have special legal incentives to adhere to published test standards because among mental tests, those that measure intelligence best (are most *g loaded*) generally have the greatest disparate impact on Blacks and Hispanics (Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997). That is, such tests yield lower average scores for these populations than for Asians and Whites. In employment settings, differing average results by race or ethnicity constitute prima facie evidence of illegal discrimination against the lower scoring groups, a charge that the accused party must then disprove, partly by showing adherence to professional standards (see chap. 5, this volume).

Tests of intelligence are also widely used in basic research in diverse fields, from genetics to sociology. They are useful, in particular, for studying human variation in cognitive ability and the ramifying implications of that variation for societies and their individual members. Current intelligence tests gauge relative, not absolute, levels of mental ability (their severest limitation, as described later). Other socially important sociopsychological measures are likewise *norm-referenced indicators*, not *criterion-referenced indicators*. Oft-used examples include neuroticism, grade point average, and occupational prestige.

Many of the pressing questions in the social sciences and public policy are likewise norm referenced, that is, they concern how far the various mem-

bers of a group fall above or below the group's average on some social indicator (e.g., academic achievement, health) or hierarchy (e.g., occupation, income), regardless of what the group average may be: Which person in the applicant pool is most qualified for the job to be filled? Which sorts of workers are likely to climb highest on the corporate ladder or earn the most and why? Which elementary school students will likely perform below grade level (a group average) in reading achievement, or which college applicants will fail to maintain a grade point average of at least C, if admitted?

Such questions about the relative competence and well-being of a society's members engage the core concern of democratic societies—social equality. Democratic nations insist that individuals should get ahead on their own merits, not through their social connections. Democracies also object to some individuals or groups getting too far ahead of or behind the pack. They favor not only equal opportunities for individuals to deploy their talents but also reasonably equal outcomes. Yet when individuals differ substantially in merit, however it is defined, societies cannot simultaneously and fully satisfy both of these goals. Mandating strictly meritocratic advancement will guarantee much inequality of outcomes, and, conversely, mandating equal outcomes will require that talent be restrained or its fruits redistributed (J. W. Gardner, 1984). This is the democratic dilemma, which is created by differences in human talent. In many applications, the chief source today of the democratic dilemma is the wide dispersion in human intelligence because higher intelligence has been well documented as providing individuals with more practical advantages in modern life than any other single attribute, including social class background (Ceci, 1996a; Herrnstein & Murray, 1994).

Democratic societies are reluctant, by their egalitarian nature, to acknowledge either the wide dispersion in intelligence or the conflicts among core values that this creates for them. Human societies have always had to negotiate such trade-offs, often institutionalizing their choices through legal, religious, and social norms (e.g., meat-sharing norms in hunter-gatherer societies).

One effect of research on intelligence tests has been to make such choices and their societal consequences clearer and more public. A sizeable literature now exists in personnel selection psychology, for example, that estimates the costs and benefits of sacrificing various levels of test validity to improve racial balance by varying degrees when selecting workers for different kinds of jobs (e.g., Schmitt et al., 1997). This literature also shows that the more accurately a test identifies who is most and least intellectually apt within a population, the more accurately it predicts which segments of society will gain or lose from social policies that attempt to capitalize on ability differences, to ignore them, or to compensate for them.

Such scientific knowledge about the distribution and functional importance of general mental ability can influence prevailing notions of what constitutes a just social order. Its potential influence on public policy and prac-

tice (e.g., require racial preferences? ban them?) is just what some applaud and others fear. It is no wonder that different stakeholders often disagree vehemently about whether test use is fair. Test use, misuse, and nonuse all provide decision makers tools for tilting trade-offs among conflicting goals in their preferred direction.

In short, the enduring, emotionally charged, public controversy over intelligence tests reflects mostly the enduring, politically charged, implicit struggle over how a society should accommodate its members' differences in intelligence. Continuing to dispute the scientific merits of well-validated tests and the integrity of persons who develop or use them is a substitute for, or a way to forestall, confronting the vexing realities that the tests expose.

That the testing controversy is today mostly a proxy battle over fundamental political goals explains why no amount of scientific evidence for the validity of intelligence tests will ever satisfy the tests' critics. Criticizing the yardstick rather than confronting the real differences it measures has sometimes led even testing experts to promulgate supposed technical improvements that actually reduce a test's validity but provide a seemingly scientific pretext for implementing a purely political preference, such as racial quotas (Blits & Gottfredson, 1990a, 1990b; Gottfredson, 1994, 1996). Tests may be legitimately criticized, but they deserve criticism for their defects, not for doing their job.

GULF BETWEEN SCIENTIFIC DEBATE AND PUBLIC PERCEPTIONS

Many test critics would reject the foregoing analysis and argue that evidence for the validity of the tests and their results is ambiguous, unsettled, shoddy, or dishonest. Although mistaken, this view may be the reigning public perception. Testing experts do not deny that tests have limits or can be misused. Nor do they claim, as critics sometimes assert (Fischer et al., 1996; Gould, 1996), that IQ is fixed, all important, the sum total of mental abilities, or a measure of human worth. Even the most cursory look at the professional literature shows how false such caricatures are.

In "Mainstream Science on Intelligence" (1994; Gottfredson, 1997), 52 experts summarized 25 of the most elementary and firmly established conclusions about intelligence and intelligence testing. In brief, professionally developed IQ tests are reliable, valid, unbiased measures of a general proficiency in learning, reasoning, and abstract thinking (the exception being verbal tests given to nonnative speakers). IQ differences among individuals are stable and highly heritable by adolescence, and they correlate genetically with many brain structures and processes. IQ level is the best single predictor of many important life outcomes, but its predictive validity varies from low to high depending on kind of outcome (e.g., .2 for law abidingness; .6 for

years of education; and .2–.8 for job performance, the correlations rising with job complexity). Average racial–ethnic differences in IQ are the rule worldwide, typically reflect average differences in phenotypic intelligence, predict average differences in life outcomes, and are perhaps both genetic and nongenetic in origin. Received wisdom outside the field is often quite the opposite (Snyderman & Rothman, 1987, 1988), in large part because of the fallacies I describe here.

Table 1.1 illustrates how the scientific debates involving intelligence testing have advanced during the past half century. The list is hardly exhaustive and no doubt reflects the particular issues I have followed in my career, but it makes the point that public controversies over testing bear little relation to what experts in the field actually debate today. For example, researchers directly involved in intelligence-related research no longer debate whether IQ tests measure a "general intelligence," are biased against American Blacks, or predict anything more than academic performance.

Those questions were answered several decades ago (answers: yes, no, and yes; e.g., see Bartholomew, 2004; Brody, 1992; Carroll, 1993; Deary, 2000; Deary et al., 2004; Gottfredson, 1997b, 2004; Hartigan & Wigdor, 1989; Hunt, 1996; Jensen, 1980, 1998; "Mainstream Science on Intelligence," 1994; Murphy & Davidshofer, 2005; Neisser et al., 1996; Plomin, DeFries, McClearn, & McGuffin, 2001; Sackett, Schmitt, Ellingson, & Kabin, 2001; Schmidt & Hunter, 1998; Wigdor & Garner, 1982).

The new debates can be observed in special journal issues (e.g., Ceci, 1996b; Frisby, 1999; Gottfredson, 1986, 1997a; Lubinski, 2004; Williams, 2000), handbooks (e.g., Colangelo & Davis, 2003; Frisby & Reynolds, 2005), edited volumes (e.g., Detterman, 1994; Flanagan, Genshaft, & Harrison, 1997; Jencks & Phillips, 1998; Neisser, 1998; Plomin & McClearn, 1993; Sternberg & Grigorenko, 2001, 2002; Vernon, 1993), reports from the National Academy of Sciences (e.g., Hartigan & Wigdor, 1989; Wigdor & Garner, 1982; Wigdor & Green, 1991; see also Yerkes, 1921), and the pages of professional journals such as *American Psychologist*; *Exceptional Children*; *Intelligence*; *Journal of Applied Psychology*; *Journal of Psychoeducational Assessment*; *Journal of School Psychology*; *Personnel Psychology*; and *Psychology, Public Policy, and Law*.

Scientific inquiry on intelligence and its measurement has therefore moved to new questions. To take an example, yes, all IQ tests measure a highly general intelligence, albeit imperfectly (more specifically, they all measure a general intelligence factor, *g*), but do all yield exactly the same *g* continuum? Technically speaking, do they converge on the same *g* when factor analyzed? This illustrates how the questions debated today are more tightly focused, more technically demanding, and more theoretical than those of decades past.

In contrast, public controversy seems stuck in the scientific controversies of the 1960s and 1970s, as if those basic questions remained open or had not been answered to the critics' liking. The clearest recent example is the

TABLE 1.1
Examples Illustrating How Scientific Debate on Intelligence and IQ Tests Has Advanced Over the Past Half Century

Early debates	More recent debates
What fundamental distinctions (constructs) do intelligence tests measure, and how well?	
Do IQ tests measure a general intelligence or just a narrow academic ability?	Do different IQ test batteries yield the same general intelligence factor (converge on the same true *g*) when factor analyzed?
Which specific mental abilities add up to create overall intelligence?	To what extent does *g* constitute the common core of different specific mental abilities?
Do IQ tests yield statistically reliable (consistent) results?	Do different methods of factor analysis yield the same *g* factor?
Do test items and formats that more closely resemble the criterion (i.e., have higher face validity, or fidelity) have higher predictive validity? If so, do they simultaneously reduce disparate impact against Blacks and Hispanics?	Raw scores on IQ tests have risen over time (the *Flynn Effect*), so do IQ tests measure different things in different epochs, or has general intelligence (*g*) increased over time, or both?
Are people's IQ levels stable over the life course?	To what extent is stability (and change) in IQ/*g* relative to agemates traceable to genetic influences? To nongenetic ones?
Can early interventions raise low IQs?	Can the fade-out of IQ gains be prevented if early interventions are continued into adolescence?
Is IQ level heritable (do differences in IQ phenotype partly reflect differences in genotype)?	How does the heritability of IQ/*g* differ by chronological age, epoch, and social circumstance?
Can broad abilities (verbal, spatial ability, etc.) be measured independently of IQ?	What is the joint heritability (and environmentality) of *g* with the group factors measured by IQ tests (verbal ability, memory, etc.) and with outcomes such as academic achievement and occupational status?
Are IQ tests biased against (systematically mismeasuring) members of minority groups (i.e., is there measurement bias)?	Does a given IQ test battery measure exactly the same construct(s) in different races, sexes, and age groups?
Do IQ tests predict important life outcomes, and how well (including relative to other predictors)?	
Do IQ levels above some low threshold (e.g., not mentally retarded) predict differences in job or school performance?	Do IQ levels above some high threshold (e.g., giftedness) predict differences in job or school performance?
Does a whole battery of different ability tests (verbal, spatial, etc.) predict outcomes (e.g., educational or occupational) substantially better than just an overall IQ score?	Which classes of cognitive and noncognitive tests provide incremental validity, when used with *g*, in predicting performance on different classes of tasks (instrumental, socioemotional)?

Do IQ tests predict performance of nonacademic tasks in everyday life?	Why does IQ predict performance to some extent in most domains of daily life, but better in some than others?
Do IQ tests predict job performance equally well for all races (i.e., is there prediction bias)?	Do IQ scores predict adult outcomes (e.g., job level, health, law abidingness) better than does socioeconomic background?
Proper test use and test utility	
Should schools stop using IQ scores for placing students into special education, gifted education, or ability groups?	Should schools stop using IQ tests (i.e., IQ-achievement gaps) to help diagnose learning disabilities?
How can clinicians make best use of subtest profiles?	When evaluating individual students, should school psychologists stop analyzing a child's profile of subtest scores (factor discrepancies) and focus just on the (more reliable) overall IQ and composite scores?
Should IQ tests be used to identify students who are intellectually gifted?	Should *giftedness* include noncognitive talents, and should selection into gifted programs rely on teacher, parent, and self ratings?
Should employers give less weight to technical expertise and more to organizational citizenship when hiring employees in order to improve racial balance?	Should colleges give less weight to cognitive abilities and more to noncognitive strengths when admitting students in order to improve racial balance?
Should the federal government race-norm its employment tests in order to equalize, by race, the scores it reports to potential employers?	Should courts allow colleges to use different SAT and ACT requirements for different races?
Which noncognitive tests should employers use instead of cognitive tests when selecting employees?	Which noncognitive tests should employers use in addition to cognitive tests when selecting employees?
Should IQ testing be banned in deciding whether an underperforming Black student is eligible for special education?	Should IQ testing be required in deciding whether a convicted killer is ineligible for the death penalty?

cacophony of public denunciation that greeted publication of *The Bell Curve* in 1994 (Herrnstein & Murray, 1994). Many journalists, social scientists, and public intellectuals derided the book's six foundational premises about intelligence as long-discredited pseudoscience when, in fact, they represent some of the most elemental scientific conclusions about intelligence and tests. Briefly, Herrnstein and Murray (1994) stated that six conclusions are "by now beyond serious technical dispute": individuals differ in general intelligence level (i.e., intelligence exists), IQ tests measure those differences well, IQ level matches what people generally mean when they refer to some individuals as being more intelligent or smarter than others, individuals' IQ scores (i.e., rank within age group) are relatively stable throughout their lives, properly administered IQ tests are not demonstrably culturally biased, and

individual differences in intelligence are substantially heritable. The cautious John B. Carroll (1997) detailed how all these conclusions are "reasonably well supported" (p. 25).

Statements by the American Psychological Association (Neisser et al., 1996) and the previously mentioned group of experts (see Gottfredson, 1997a; "Mainstream Science on Intelligence," 1994), both of whom were attempting to set the scientific record straight in both public and scientific venues, did little if anything to stem the tide of misrepresentation. Reactions to *The Bell Curve*'s analyses illustrate not just that today's received wisdom seems impervious to scientific evidence but also that the guardians of this wisdom may only be inflamed further by additional evidence contradicting it.

Mere ignorance of the facts cannot explain why accepted opinion tends to be opposite the experts' judgments (Snyderman & Rothman, 1987, 1988). Such opinion reflects systematic misinformation, not lack of information. The puzzle, then, is to understand how the empirical truths about testing are made to seem false, and false criticisms made to seem true. In the millennia-old field of *rhetoric* (verbal persuasion), this question falls under the broad rubric of *sophistry*.

SOPHISTRIES ABOUT THE NATURE AND MEASUREMENT OF INTELLIGENCE

In this chapter, I describe major logical confusions and fallacies that in popular discourse seem to discredit intelligence testing on scientific grounds but actually do not. My aim here is not to review the evidence on intelligence testing or the many misstatements about it but to focus on particularly seductive forms of illogic. As noted earlier, many aptitude and achievement tests are de facto measures of *g* and reveal the same democratic dilemma as do IQ tests, so they are beset by the same fallacies. I am therefore referring to all highly *g*-loaded tests when I speak here of intelligence testing.

Public opinion is always riddled with error, of course, no matter what the issue. However, fallacies are not simply mistaken claims or intentional lies, which could be answered effectively with facts contradicting them. Instead, fallacies tend to corrupt public understanding systematically. They not only present falsehoods as truths but also reason falsely about the facts, thus making those persons they persuade largely insensible to correction. Effectively rebutting a fallacy's false conclusion therefore requires exposing how its reasoning turns the truth on its head. For example, a fallacy might start with an obviously true premise about Topic A (within-individual growth in mental ability), then switch attention to Topic B (between-individuals differences in mental ability) but obscure the switch by using the same words to describe both ("change in"), and then use the uncontested fact about A (change) to seem to disprove well-established but unwelcome facts about B

(lack of change). Contesting the fallacy's conclusion by simply reasserting the proper conclusion leaves untouched the false reasoning's power to persuade—in this case, its surreptitious substitution of the phenomenon being explained.

The individual antitesting fallacies that I describe in this chapter rest on diverse sorts of illogic and misleading argument, including non sequiturs, false premises, conflation of unlikes, and appeals to emotion. Collectively they provide a grab bag of complaints for critics to throw at intelligence testing and allied research. The broader the barrage, the more it appears to discredit anything and everyone associated with intelligence testing.

The targets of fallacious reasoning are likewise diverse. Figure 1.1 helps to distinguish the usual targets by grouping them into three arenas of research and debate: Can intelligence be measured, and if so, how? What are the causes and consequences of human variation in intelligence? Finally, what are the social aims and effects of using intelligence tests—or not using them—as tools in making decisions about individuals and organizations? These are labeled in Figure 1.1, respectively, as the measurement model, the causal network, and the politics of test use. Key phenomena (actually, fields of inquiry) within each arena are distinguished by numbered entries to illustrate more easily which fact or field each fallacy works to discredit. The arrows (→) represent the relations among the phenomena at issue, such as the causal impact of genetic differences on brain structure (Entry 1 → Entry 4 in Figure 1.1), or the temporal ordering of advances in mental measurement (Entries 8 → 9 → 10 → 11 in Figure 1.1). As we shall see, some fallacies work by conflating different phenomena (e.g., Entry 1 with 4, 2 with 3, 8 with 11 in Figure 1.1), others by confusing a causal relation between two phenomena (e.g., 1 → 5) with individual differences in one of them (5), yet others by confusing the social criteria (6 and 7) for evaluating test *utility* (the costs and benefits of using a valid test) with the scientific criteria for evaluating its validity for measuring what is claimed (11), and so on.

MEASUREMENT MODEL

Psychological tests and inventories aim to measure enduring, underlying personal traits, such as extraversion, conscientiousness, or intelligence. The term *trait* refers to notable and relatively stable differences among individuals in how they tend to respond to the same circumstances and opportunities: For example, Jane is sociable, and Janet is shy among strangers. A psychological trait cannot be seen directly, as can height or hair color, but is inferred from striking regularities in behavior across a wide variety of situations—as if different individuals follow different internal compasses as they engage the world around them. Because they are inferred, traits are called *theoretical constructs*. They therefore represent causal hypotheses about why

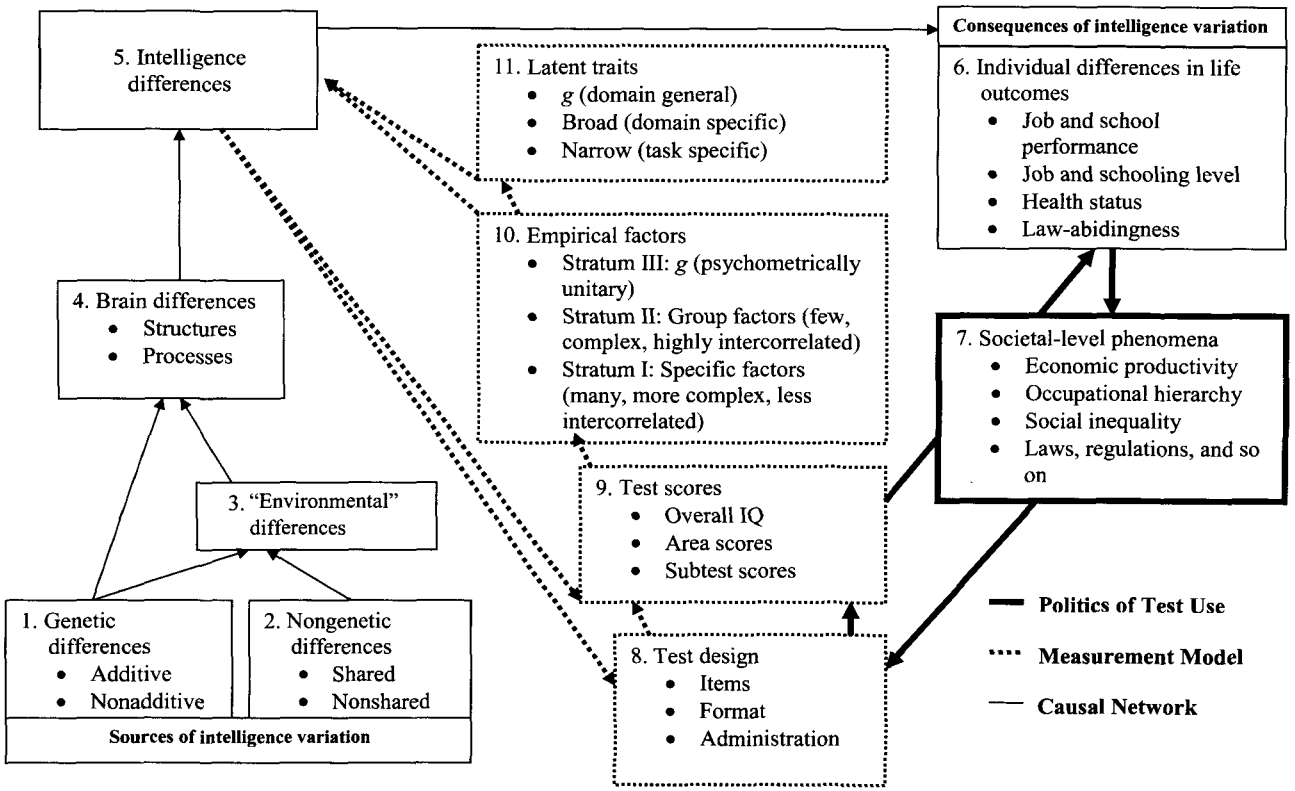

Figure 1.1. Three foci of fallacious reasoning: Measurement of intelligence, causes and consequences of intelligence differences, and the politics of test use.

individuals differ in patterned ways. Many other disciplines also posit influences that are not visible to the naked eye (e.g., gravity, electrons, black holes, genes, natural selection, self-esteem) and that must be detected through their effects on something that is observable. Intelligence tests consist of a set of tasks that reliably instigates performances requiring mental aptness and of procedures to record quality of task performance.

The measurement process thus begins with a hypothesized causal force and ideas about how it manifests itself in observable behavior. This nascent theory provides clues to what sort of task might activate it. Designing those stimuli and ways to collect responses to them in a consistent manner is the first step in creating a test. It is but the first step, however, in a long forensic process in which many parties collect evidence to determine whether the test does indeed measure the intended construct and whether initial hypotheses about the construct might have been mistaken. Conceptions of the phenomenon in question and how best to capture it in action evolve during this collective, iterative process of evaluating and revising tests. General intelligence is by far the most studied psychological trait, so its measurement technology is the most developed and thoroughly scrutinized of all psychological assessments.

As techniques in the measurement of intelligence have advanced, so, too, have the fallacies about it multiplied and mutated. Figure 1.1 delineates the broad stages (Entries 8 through 11 in Figure 1.1) in this coevolution of intelligence measurement and the fallacies related to it. In this section, I describe the basic logic guiding the design, the scoring, and the validation of intelligence tests and then, for each in turn, several fallacies associated with them. Later sections describe fallacies associated with the causal network for intelligence and with the politics of test use. Table 1.2 lists several examples of each fallacy. The examples illustrate that many important opinion makers use these fallacies, some use them frequently, and even rigorous scholars (Examples xx, xxi, and xxix) may inadvertently promulgate them. Each example is quoted at length and dissected more fully in Appendix A (see http://www.apa.org/books/resources/Phelps/).

Test-Design Fallacies

There were no intelligence tests in 1900 but only the perception that individuals consistently differ in mental prowess and that such differences have practical importance. Binet and Simon (1916), who produced the progenitor of today's IQ tests, hypothesized that such differences might forecast which students have extreme difficulty with schoolwork. So they set out to invent a measuring device (Entry 8 in Figure 1.1) to reveal and quantify differences among schoolchildren in that hypothetical trait (Entry 5 in Figure 1.1), as Binet's observations had led him to conceive it. The French Ministry of Education had asked Binet to develop an objective way to identify

TABLE 1.2
Thirteen Especially Influential Logical Fallacies About Intelligence Testing and Where to Find Examples of Each

Example no.	Reference	Context
	Test-design fallacy 1. Yardstick mirrors construct: Portrays the superficial appearance of a test as if it mimicked the inner essence of the phenomenon it measures.	
i	Fischer et al. (1996, pp. 42–43, 56–59)	The authors argued that the Armed Forces Qualification Test (AFQT) does not measure IQ or "intelligence broadly understood" but only learning in school.
ii	Flynn (2007, p. 55)	The author proposed a skills-based definition of intelligence that is "narrow enough to offer good advice to those who want to make intelligence measurable and specific."
iii	Sternberg, Wagner, Williams, and Horvath (1995, p. 913)	The authors argued that different item formats ("academic" vs. "practical") necessarily call forth different intelligences. They claimed that IQ tests use only the former and thus can measure only an "academic intelligence" (*g*).
	Test-design fallacy 2. Intelligence is marble collection: Portrays general intelligence (*g*) as if it were just an aggregation of many separate specific abilities or skills, not a singular phenomenon in itself, because IQ batteries calculate IQs by adding up scores on different subtests.	
iv	Flynn (2007, p. 55), from example ii above	The author proposed a skills-based definition of intelligence that is "narrow enough to offer good advice to those who want to make intelligence measurable and specific."
v	Flynn (2007, pp. 4, 9–10, 18)	The author explained how secular increases in IQ test scores might represent a rise in overall intelligence but not in *g*, the issue at hand being that scores on some highly *g*-loaded IQ subtests (e.g., Similarities) have risen a lot but others (e.g., Vocabulary) hardly all—"or, how can IQ gains be so contemptuous of *g* loadings?"
vi	Howe (1997, pp. 161–162)	The author listed what he viewed as "Twelve Well-Known 'Facts' about Intelligence Which Are Not True."

	Test-score-differences fallacy 1. Nonfixedness proves malleability: Uses evidence of any fluctuation or growth in the mental functioning of individuals as if it were proof that their rates of growth can be changed intentionally.	
vii	Blakemore (1994)	This ABC newscast contested *The Bell Curve's* claim that intelligence is a stable, measurable trait.
viii	Howe (1997, p. 138)	The author discussed what he considered better alternatives to "traditional intelligence theory."
	Test-score-differences fallacy 2. Improvability proves equalizability: Uses evidence that intellectual skills and achievements can be improved within a population as if it were proof that they can be equalized in that population.	
ix	Howe (1997, pp. 62–63)	The author argued for interventions to raise the IQs of individuals in disadvantaged groups.
x	The White House (2001)	The Executive Summary of the No Child Left Behind Act of 2001, which appeared on the White House Web site, highlighted the Act's intent to close achievement gaps by bringing all students up to the same high level of achievement.
xi	Dionne (1994, p. A17)	This *Washington Post* columnist argued that *The Bell Curve* "is not a 'scientific' book at all but a political argument offered by skilled polemicists aimed at defeating egalitarians."
	Test-score-differences fallacy 3. Interactionism (gene–environment codependence) nullifies heritability: Portrays the gene–environment partnership in creating a phenotype as if conjoint action within the individual precluded teasing apart the roots of phenotypic differences among individuals.	
xii	Sternberg (1997, p. 48)	The author distinguished what he described as the "conventional IQ-based view" of intelligence from his proposed notion of "successful intelligence."
xiii	Andrews and Nelkin (1996, p. 13)	This letter to *Science* disputed key conclusions in *The Bell Curve.*

continues

TABLE 1.2
Continued

Example no.	Reference	Context
Test-score-differences fallacy 4. 99.9% similarity negates differences: Portrays the study of human genetic variation as irrelevant or wrong-headed because humans are 99.9% (or 99.5%) alike genetically, on average.		
xiv	Park (2002, pp. 395–398)	This anthropology textbook discussed "why [in its author's view] there are no biological races within the human species."
xv	Holt (1994, p. A23)	This *New York Times* opinion–editorial piece disputed the idea that racial differences in intelligence could have any genetic basis.
xvi	Marks (1995, pp. 273–275)	The author summed up his argument by saying that genetic differences by race are minor but are exaggerated in order to justify and perpetuate social inequality.
Test-validation fallacy 1. Contending definitions negate evidence: Portrays lack of consensus in verbal definitions of intelligence as if this negated evidence for the construct validity of IQ tests.		
xvii	Singham (1995, pp. 272, 278)	The author advised educators that *The Bell Curve* is, in his view, unscientific and ideological.
xviii	"The 'Bell Curve' agenda" (1994, p. A16)	This *New York Times* editorial argued that "what is new about [*The Bell Curve* book]—the fixation on genes as destiny—is surely unproved and almost surely wrong" and that therefore IQ level actually is manipulable.
Causal-network fallacy 1: *Phenotype* equals *genotype*: Portrays phenotypic differences in intelligence as if they were necessarily genotypic.		
xix	Duster (1995, p. 1)	The author argued that "there has always been a tendency to link existing social orders with so-called innate physical, intellectual and spiritual qualities."
xx	Bartholomew (2004, pp. 122–123)	The author described the difficulty of determining whether the Black–White IQ difference originates in whole or part in the genes or whether it can be wholly accounted for by environmental factors.

	Causal-network fallacy 2. *Biological* equals *genetic*: Portrays biological differences (such as brain phenotypes) as if they were necessarily genetic.	
xxi	Bartholomew (2004, p. 138)	The author discussed possible sources of the "Flynn effect" (the secular rise in IQ).
xxii	"Race and Intelligence" (2007)	National Public Radio's *News & Notes* program followed up an interview with J. P. Rushton, who reported correlations between race, brain size, and intelligence, by interviewing a critic of intelligence research.
	Causal-network fallacy 3. *Environmental* equals *nongenetic*: Portrays external environments as if they were necessarily nongenetic, that is, unaffected by and unrelated to the genotypes of individuals in them.	
xxiii	Monastersky (2008, ¶ 1)	This news article reported research on "how poverty alters the brain."
xxiv	Fischer et al. (1996, p. 68)	The authors argued that the AFQT measures differences in opportunity to learn, not "raw intelligence."
	Standards-of-evidence fallacy 1. Imperfect measurement pretext: Maintains that valid, unbiased intelligence tests should not be used for making decisions about individuals until the tests are made error free.	
xxv	FairTest (2007, ¶¶ 12, 14, and 17)	One of FairTest's University Testing Fact Sheets on its Web site argues that the ACT, SAT, and SAT Subject Tests are not accurate enough to be used in evaluating applicants for college admissions and scholarships.
xxvi	Miller (2001, p. A14)	This news article in *Chronicle of Higher Education* reported complaints in the education profession about large-scale testing.
xxvii	Hartigan and Wigdor (1989, pp. 7–8)	The authors of this National Academy of Sciences report explained why they recommended that the U.S. Employment Service continue to race-norm job applicants' employment test scores.

continues

TABLE 1.2
Continued

Example no.	Reference	Context
	Standards-of-evidence fallacy 2. Dangerous-thoughts trigger: Maintains that scientific conclusions purported to be divisive or dangerous should not be entertained until proved beyond all possible doubt.	
xxviii	C. Kiesler (January 17, 1980, personal communication to A. R. Jensen)	The editor of the *American Psychologist* explained why he rejected Arthur Jensen's manuscript, "The Nature of the Average Difference between Whites and Blacks on Psychometric Tests: Spearman's Hypothesis" (which was later published in another journal as a target article; see Jensen, 1985).
xxix	Hunt and Carlson (2007, p. 210)	The authors proposed standards for conducting and evaluating research on group differences in intelligence.
	Standards-of-evidence fallacy 3. Happy-thoughts leniency: Maintains that mere theoretical possibility elevates the scientific credibility of a politically popular idea above that of an empirically plausible but unpopular conclusion.	
xxx	Diamond (1999, pp. 16, 19)	The author argued that "biological differences" cannot account for "why . . . human development proceed[ed] at such different rates on different continents" over human history, despite seemingly compelling arguments that they do.
xxxi	"The 'Bell Curve' agenda" (1994, p. A16)	This *New York Times* editorial argued that "what is new about [*The Bell Curve* book]—the fixation on genes as destiny—is surely unproved and almost surely wrong" and that therefore IQ level actually is manipulable.

Note. See Appendix A (on the following Web site: http://www.apa.org/books/resources/Phelps/) for extensive, annotated excerpts from the cited sources.

students who would not succeed academically without special attention. He began with the observation that students who had great difficulty with their schoolwork also had difficulty doing many other things that children their age usually can do. Intellectually, they were more like the average child a year or two younger—hence the term *retarded* development. According to Binet and Simon (1916), the construct to be measured is manifested most clearly in quality of reasoning and judgment in the course of daily life:

> It seems to us that in intelligence there is a fundamental faculty, the alteration or lack of which is of the utmost importance for practical life. This faculty is judgment, otherwise called good sense, practical sense, initiative, the faculty of adapting one's self to circumstances. To judge well, to reason well, these are the essential activities of intelligence. A person may be a moron or an imbecile if he is lacking in judgment: but with good judgment he can never be either. Indeed the rest of the intellectual faculties seem of little importance in comparison with judgment. (pp. 42–43)

This conception provided a good starting point for designing tasks that might effectively activate intelligence and cause it to leave its footprints in observable behavior. Binet and Simon's (1916) strategy was to develop a series of short, objective questions that sampled specific mental skills and bits of knowledge that the average child accrues in everyday life by certain ages, such as asking the child to point to nose, eyes, and mouth (age 3); count 13 pennies (age 6); note omissions from pictures of familiar objects (age 8); arrange five blocks in order of weight (age 10); and discover the sense of a disarranged sentence (age 12). In light of having postulated a highly general mental ability or broad set of intellectual skills, it made sense to assess performance on a wide variety of mental tasks children are routinely exposed to outside of schools and are expected to master in the normal course of development. For the same reason, it was essential not to focus on any specific domain of knowledge or expertise, as would a test of knowledge in a particular job or school subject.

The logic is that mastering fewer such everyday tasks than is typical for one's age signals a lag in the child's overall mental development; that a short series of items that are strategically selected, carefully administered, and appropriately scored (a *standardized test*) can make this lag manifest; and that poorer performance on such a test will forecast greater difficulty in mastering the regular school curriculum (i.e., the increasingly difficult series of cognitive tasks that schools pose for pupils at successively higher grade levels). For a test to succeed, its items must range sufficiently in difficulty at each age in order to capture the range of variation at that age. Otherwise, it would be like having a weight scale that can register nothing below 50 pounds or above 100 pounds.

Most modern intelligence tests still follow the same basic principle—test items should sample a wide variety of cognitive performances at different

difficulty levels. Over time, individually administered intelligence test batteries have grown to include a dozen or more separate subtests (e.g., Wechsler Intelligence Scale for Children, 4th ed. [WISC–IV; Wechsler, 2003] subtests such as Vocabulary, Block Design, Digit Span, Symbol Search, Similarities) that systematically sample a range of cognitive processes. Subtests are usually aggregated into broader content categories (e.g., the WISC–IV's four index scores: Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed). The result is to provide at least three tiers of scores (see Entry 9 in Figure 1.1): individual subtests, clusters of subtests (area scores, indexes, composites, etc.), and overall IQ. The overall IQs from different IQ test batteries generally correlate at least at .8 among themselves (which is not far below the maximum possible in view of their reliabilities of .9 or more), so they are capturing the same phenomenon. Mere similarity of results among IQ tests is necessary, of course, but not sufficient to confirm that the tests measure the intended construct.

Today, item content, test format, and administration procedure (Entry 8 in Figure 1.1) are all tightly controlled to maximize accuracy in targeting the intended ability and to minimize contamination of scores by random error (e.g., too few items to get consistent measurement) or irrelevant factors (e.g., motivation, differential experience, or unequal testing circumstances). Test items therefore ideally include content that is either novel to all test takers or to which all test takers have been exposed previously. Reliable scoring is facilitated (measurement error is reduced) by using more numerous test items and by using questions with clearly right and wrong answers.

The major intelligence tests, such as the Stanford–Binet and the Wechsler series for preschoolers (Wechsler Preschool and Primary Scale of Intelligence; Wechsler, 2002), school-aged children (WISC; Wechsler, 2003), and adults (Wechsler Adult Intelligence Scale [WAIS; Wechsler, 1997]) are administered orally to test takers one on one, item by item for 1 hour or more, by highly trained professionals who follow written scripts governing what they must and must not say to the individual to ensure standard conditions for all test takers (Sattler, 2001). Within those constraints, test administrators seek to gain rapport and otherwise establish conditions to elicit maximal performance.

The foregoing test-design strategies increase the likelihood of creating a test that is reliable and valid—one that consistently measures the intended construct and nothing else. Such strategies cannot guarantee this happy result, of course. This is why tests and the results from all individual test items are required to jump various statistical hurdles after tryout and before publication and why, after publication, tests are subjected to continuing research and periodic revision. These guidelines for good measurement result, however, in tests with superficial appearances that make them highly vulnerable to fallacious reasoning of the following sorts.

Test-Design Fallacy 1: Yardstick Mirrors Construct

This fallacy involves portraying the superficial appearance of a test (Entry 8 in Figure 1.1) as if it mimicked the inner essence of the phenomenon it measures (Entry 5 in Figure 1.1). For example, it would be nonsensical to claim that a thermometer's outward appearance provides insight into the nature of heat or that differently constructed thermometers obviously measure different kinds of heat. Yet some critiques of intelligence testing rest precisely on such reasoning. For example, Fischer et al. (1996; see also Appendix A, Example i, and Table 1.2, this volume) decided on face value that the Armed Forces Qualification Test (AFQT) measures "mastery of school curricula" and nothing deeper, and Flynn (2007; Appendix A, Example ii) asserted that various WISC subtests measure "what they say." Sternberg, Wagner, Williams, and Horvath (1995; see also Appendix A, Example iii) argued that IQ tests measure only "academic" intelligence because they pose tasks that appear to their eye only academic: well-defined tasks with narrow, esoteric, or academic content of little practical value that always have right and wrong answers and do not give credit for experience.

All three examples reinforce the fallacy they deploy: that one can know what a test measures by just peering at its items. Like reading tea leaves, critics list various superficialities of test content and format to assert, variously, that IQ tests measure only an aptness with paper-and-pencil tasks, a narrow academic ability, familiarity with the tester's culture, facility with well-defined tasks with unambiguous answers, and so on. Not only are these inferences unwarranted, but their premises about content and format are often wrong. In actuality, most items on individually administered batteries require neither paper nor pencil, most are not timed, many do not use numbers or words or other academic-seeming content, and many require knowledge of only the most elementary concepts (up–down, large–small, etc.). Neither the mechanics nor superficial content of IQ tests reveals the essence of the construct they capture. Manifest item content—*content validity*—is critical for certain other types of tests, specifically, ones meant to gauge knowledge or achievement in some particular content domain, such as algebra, typing, or jet engine repair.

Figuring out what construct(s) a particular test actually measures requires extensive validation research, which involves collecting and analyzing test results in many circumstances and populations (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). As described later in this chapter, such research shows that ostensibly different tests can be used to measure the same latent ability. Spearman (1927) characterized this as the "indifference to the indicator" (pp. 197–198). The yardstick–mirrors–construct fallacy, by contending that a test measures only what it "looks like," allows critics to assert, a priori, that IQ tests cannot possibly measure a highly

general mental capability. It thereby precludes, on seemingly scientific grounds, the very success that tests have already demonstrated.

Test-Design Fallacy 2: Intelligence Is Marble Collection

This fallacy involves portraying general intelligence (*g*) as if it were just an aggregation of many separate specific abilities, not a singular phenomenon in itself (Entry 10 in Figure 1.1) because of the way IQ scores are typically calculated, which essentially is to add up a person's scores on the various subtests in a battery (Entry 9 in Figure 1.1). This fallacy is similar to the previous one in that it presumes that the manner of calculating scores from IQ tests (the *measure*) mirrors how general intelligence itself (the hypothetical entity or *construct*) is constituted. That is, the marble-collection fallacy holds that intelligence is made up of separable components, the sum total of which we label *intelligence*. It is not itself an identifiable entity but, like marbles in a bag, just a conglomeration or aggregate of many separate things we choose to add to the collection.

Flynn (2007) conceptualized intelligence in this manner to cast doubt on the psychological reality of *g*. He viewed IQ subtests as isolating various "components" of "intelligence broad" (p. 55; see Appendix A, Example iv). "Understanding intelligence is like understanding the atom." Its parts can be "split apart," "assert their functional autonomy," and "swim freely of *g*" (pp. 4, 10, 18; see Appendix A, Example v). For Howe (1997), the IQ is no more than a "range of mental tasks" (p. 162; see Appendix A, Example vi).

This conglomeration view holds IQ tests hostage to complaints that they cannot possibly measure intelligence because they do not include the complainant's preferred type or number of marbles. Williams (1996), for example, suggested that "a broader perspective on intelligence may enable us to assess . . . previously unmeasured aspects of intelligence" (pp. 529–530). She favored an expansive conception of intelligence that includes a "more ecologically relevant set of abilities" (p. 350), including motivation, Sternberg's proposed practical and creative intelligences, and Gardner's postulated seven-plus multiple intelligences.

The conglomeration conception may have been a viable hypothesis in Binet's time, but it has now been decisively disproved. As discussed later in the chapter, *g* (Entry 10 in Figure 1.1) is not the sum of separate, independent cognitive skills or abilities but is the common core of them all. In this sense, general intelligence is psychometrically unitary. Whether *g* is unitary at the physiological level is an altogether different question (Jensen, 1998, 2006), but most researchers think that is unlikely.

Test-Score-Differences Fallacies

Answers to items on a test must be scored in a way that allows for meaningful interpretation of test results. The number of items answered cor-

rectly, or *raw score*, has no intrinsic meaning. Nor does percentage correct, because the *denominator* (total number of test items) also has no substantive meaning. Percentage correct can be boosted simply by adding easier items to the test, and it can be decreased by using more difficult ones. Scores become interpretable only when placed within some meaningful frame of reference. For example, an individual's score may be *criterion referenced*, that is, compared with some absolute performance standard ("90% accuracy in multiplying two-digit numbers") or it may be *norm referenced*, that is, lined up against others in some carefully specified normative population ("60th percentile in arithmetic among American fourth graders taking the test last year"). The first intelligence tests allowed neither sort of interpretation, but virtually all psychological tests are norm referenced today.

Binet and Simon (1916) attempted to provide interpretable intelligence test results by assigning a *mental age* (MA; the age at which the average child answers a given item correctly) to each item on their test. Because mental capacity increases over childhood, a higher MA score can be interpreted as a sign of more advanced cognitive development. To illustrate, if 8-year-olds answer an average of 20 items correctly, then a raw score of 20 on that test can be said to represent a mental age of 8; if 12-year-olds correctly answer an average of 30 items, then a raw score of 30 represents MA = 12. Thus, if John scores at the average for children aged 10 years, 6 months, he has a mental age of 10.5. How his mental age is interpreted depends, of course, on how old John is. If he is 8 years old, then his MA of 10.5 indicates that he is brighter than the average 8 year old (whose MA = 8.0, by definition). If he is age 12, his mental development lags behind that of other 12-year-olds (whose MA = 12.0).

In today's terms, Binet and Simon (1916) derived an *age equivalent*, analogous to the *grade equivalent* (GE) that is frequently used in reporting academic achievement in elementary school: "Susie's GE score on the school district's math test is 4.3"; that is, she scored at the average for children in the 3rd month of Grade 4.

The 1916 version of the Stanford–Binet Intelligence Scale began factoring the child's actual age into the child's score by calculating an *intelligence quotient* (IQ), specifically by dividing mental age by chronological age (CA) and multiplying by 100, to eliminate decimals. By this new method, if John were aged 10 (or 8, or 12), his MA of 10.5 would give him an IQ of 105 (or 131, or 88). IQ thus came to represent relative standing within one's own age group (MA/CA), not among children of all ages (MA). One problem with this innovation was that because mental age usually begins leveling off in adolescence but chronological age continues to increase, the MA/CA quotient yields nonsensical scores beyond adolescence.

The 1972 version of the Stanford–Binet inaugurated the *deviation IQ*, which has become standard practice. It indexes how far above or below the average, in standard deviation units, a person scores relative to others of the

same age (by month for children, and by year for adults). Distance from an age group's average is quantified by normalizing test scores, that is, transforming raw scores into locations along the normal curve (z scores, which have a mean of 0 and standard deviation of 1). This transformation preserves the rank ordering of the raw scores. For convenience, the Stanford–Binet transformed the z scores to have a mean of 100 and a standard deviation of 16 (the Wechsler and many other IQ tests today set $SD = 15$). Fitting test scores onto the normal curve in this way means that 95% of each age group get scores within 2 standard deviations of the mean, that is, between IQs 68 and 132 (when SD is set to 16) or between IQs 70 and 130 (when SD is set to 15). Translating z scores into IQ points is similar to changing temperatures from Fahrenheit into centigrade. The resulting deviation IQs are more interpretable than the MA/CA IQ, especially in adulthood, and normalized scores are far more statistically tractable. The deviation IQ is not a quotient, but the acronym IQ was retained—not unreasonably because the two forms of scores remain highly correlated in children.

With deviation IQs, intelligence became fully norm referenced. Norm-referenced scores are extremely useful for many purposes, but they, too, have serious limitations. To see why, look at the example of temperature. Consider the centigrade scale: Zero degrees is assigned to the freezing point for water and 100 degrees to its boiling point (at sea level). This gives substantive meaning to thermometer readings. IQ scores have never been anchored in this way to any concrete daily reality that would give them additional meaning. Norm-referenced scores such as the IQ are valuable when the aim is to predict differences in performance within a given population, but they allow us to rank individuals only relative to each other and not against anything external to the test. One searches in vain, for instance, for a good accounting of the capabilities that 10-year-olds, 15-year-olds, or adults of IQ 110 usually possess but similarly aged individuals of IQ 90 do not, or which particular intellectual skills a Verbal score of 600 on the SAT usually reflects. Such accountings are possible but require special research. Lack of detailed criterion-related interpretation is also teachers' chief complaint about many standardized achievement tests: "I know Sarah ranked higher than Sammie in reading, but what exactly can either of them do, and on which sorts of reading tasks do they each need help?"

IQ tests are not intended to isolate and measure highly specific skills and knowledge. This is the job of suitably designed achievement tests. However, the fact that the IQ scale is not tethered at any point to anything concrete that people can recognize understandably invites suspicion and misrepresentation. It makes IQ tests black boxes into which people can project all sorts of unwarranted hopes and fears. Psychometricians speaking in statistical tongues may be perceived as psychomagicians practicing dark arts.

Thermometers illustrate another limitation of IQ tests. We cannot be sure that IQ tests provide interval-level measurement rather than just ordi-

nal-level (i.e., rank-order) measurement. Fahrenheit degrees are 1.8 times larger than centigrade degrees, but both scales count off from zero and in equal units (degrees). So the 40-degree difference between 80 degrees and 40 degrees measures off the same difference in heat as does the 40-degree difference between 40 degrees and zero, or zero and –40. Not so with IQ points. Treating IQ as an interval-level scale has been a reasonable and workable assumption for many purposes, but we really do not know whether a 10-point difference measures the same intellectual difference at all ranges of IQ.

There is a more serious technical limitation, shared by both IQ tests and thermometers, which criterion-referencing cannot eliminate—lack of ratio measurement. Ratio scales measure absolute amounts of something because they begin measuring, in equal-sized units, from zero (total absence of the phenomenon). Consider a pediatrician's scales for height and weight, both of which start at zero and have intervals of equal size (inches or pounds). In contrast, zero degrees centigrade does not represent total lack of heat (absolute zero), nor is 80 degrees twice the amount of heat as 40 degrees, in absolute terms. Likewise, IQ 120 does not represent twice as much intelligence as IQ 60. We can meaningfully say that Sally weighs 10% more today than she did 4 years ago, she grew taller at a rate of 1 inch per year, or she runs 1 mile per hour faster than her sister. We can also chart absolute changes in all three rates. We can do none of this with IQ test scores because they measure relative standing only, not absolute mental power. They can rank but not weigh.

This limitation is shared by all measures of ability, personality, attitude, social class, and probably most other scales in the social sciences. We cannot say, for example, that Bob's social class increased by 25% last year, that Mary is 15% more extroverted than her sister, or that Nathan's self-esteem has doubled since he learned to play baseball. Although lack of ratio measurement might seem an abstruse matter, it constitutes the biggest measurement challenge facing intelligence researchers today (Jensen, 2006). Imagine trying to study physical growth if scales set the average height at 4 feet for all ages and variability in height to be the same for 4-year-olds as for 40-year-olds. Norm-referenced height measures like these would greatly limit our ability to study normal patterns of growth and deviations around it. Yet better this "deviation height" scoring than assigning ages to height scores and dividing that "height age" by chronological age to get "height quotient" because a height quotient would seem to show adults getting shorter and shorter with age! Such has been the challenge in measuring and understanding general intelligence.

Lack of ratio measurement does not invalidate psychological tests by any means, but it does limit what can be learned from them. It also nourishes certain fallacies about intelligence testing because without the absolute results to contradict them, critics can falsely represent differences in IQ scores (relative standing in ability) as if they gauged absolute differences in ability

to ridicule and discredit the test results. The following four measurement fallacies are not used to dispute the construct validity of intelligence tests, as did the two test-design fallacies. Rather, they target well-established facts about intelligence that would, if accepted, require acknowledging social trade-offs that democratic societies would rather not ponder. All four work by confusing different components of variation: (a) how individuals typically grow or change over time versus differences among them in growth or change, (b) changes in a group's mean versus changes in the spread of scores within the group, (c) the basic inputs required for any individual to develop (hence, not concerning variation at all) versus differences in how individuals develop, and (d) differences within a species versus differences between species.

Test-Score-Differences Fallacy 1: Nonfixedness Proves Malleability

This fallacy uses evidence of any fluctuation or growth in the mental functioning of individuals as if it were proof that their rates of growth can be changed intentionally. IQ level is not made malleable by any means yet devised (Brody, 1996), but many a critic has sought to dismiss this fact by pointing to the obvious but irrelevant fact that individuals grow and learn. The nonfixedness-proves-malleability fallacy succeeds by using the word change for two entirely different phenomena as if they were the same phenomenon. It first points to developmental "change" within individuals to suggest, wrongly, that IQ levels (relative differences between age mates) can be readily "changed." Asserting that IQ is stable (unchanging) despite this obvious growth (change) therefore makes one appear foolish or doggedly ideological.

Consider, for instance, the November 22, 1994, "American Agenda" segment of *World News Tonight With Peter Jennings* (Blakemore, 1994), which was devoted to debunking several of *The Bell Curve*'s six foundational premises (Appendix A, Example vii). It reported that intelligence is "almost impossible to measure" and cannot be "largely genetic and fixed by age 16 or 17" because the brain is constantly changing owing to "hydration, nutrition, and stimulation"; "learning"; and "everything it experiences, from its first formation in utero." Howe (1997; Appendix A, Example viii) provided a more subtle but more typical example when he criticized "intelligence theory" for "ignor[ing] the fact human intelligence develops rather than being static" (p. 138). By thus confusing within-individual growth with the stability of between-individual differences, he can accuse the field of denying that development occurs simply because it focuses on a different question.

Figure 1.2 distinguishes the two phenomena being confused: absolute growth versus growth relative to age mates. The three curves represent in stylized form the typical course of cognitive growth and decline for individuals at three levels of relative ability: IQs 70, 100, and 130. All three sets of individuals develop along similar lines, their mental capabilities rising in childhood (in absolute terms), leveling off in adulthood, and then falling somewhat in old age. The mental growth trajectories for brighter individuals

are steeper, so they level off at a higher point. This typical pattern has been ascertained from various specialized tests whose results are not age normed. As noted earlier, current tests cannot gauge absolute level of intelligence ("raw mental power" in Figure 1.2), so the shape of the curves cannot be verified. Evidence is unambiguous, however, that they differ greatly across individuals.

Current IQ tests cannot chronicle amount of growth and decline over a lifetime because they are not ratio measures. They compare individuals only with others of the same age, say, other 20-year-olds. If an individual scores at the average for his age group every year, then that person's IQ score will always be 100. In technical terms, the IQ will be *stable* (i.e., rank in age group remains the same). IQ level is, in fact, fairly stable in this sense from the elementary grades to old age. The stability of IQ rank at different ages dovetails with the disappointing results of efforts to raise low IQ levels, that is, to accelerate the cognitive growth of less able children and thereby move them up in IQ rank relative to some control group.

Ratio measurement would make the nonfixedness fallacy as transparent for intelligence as it would be for height: Children change and grow, so their differences in height must be malleable. Absent this constraint, it is easy for critics to use the inevitability of within-person change to deny the observed stability of between-person differences. One is invited to conclude that cognitive inequality need not exist. The next fallacy builds on the current one to suggest that the means for eradicating it are already at hand and only ill will blocks their use.

Test-Score-Differences Fallacy 2: Improvability Proves Equalizability

This fallacy portrays evidence that intellectual skills and achievements can be improved within a population as if it were proof that they can be equalized in that population. Stated more statistically, this fallacy asserts that if social interventions succeed in raising mean levels of skill, they must necessarily be effective for eradicating its members' differences in skill level. This flouts the fact that interventions that raise a group's mean usually *increase* (not decrease) its standard deviation (cf. Ceci & Papierno, 2005), a phenomenon so regular that Jensen christened it the "second law of individual differences" (Sarich & Miele, 2004, p. 258). Howe (1997) appealed to the improvability-proves-equalizability fallacy when he argued that

> in a prosperous society, only a self-fulfilling prophecy resulting from widespread acceptance of the false visions expounded by those who refuse to see that intelligence is changeable would enable perpetuation of a permanent caste of people who are prevented from acquiring the capabilities evident in successful men and women and their rewards. (pp. 62–63; see Appendix A, Example ix)

The equalizability fallacy is a virtual article of faith in educational circles. Public education was meant to be the great equalizer by giving all children a chance to rise in society regardless of their social origins, and thus nowhere has the democratic dilemma been more hotly denied yet more conspicuous than in the schools. Spurning the constraints of human cognitive diversity, the schooling-related professions generally hold that schools can simultaneously achieve equality and excellence—hence the catchphrase "EQuality" or "E-Quality"—and that beliefs to the contrary threaten both goals (Smith & Lusthaus, 1995). They contend, further, that schools could achieve both simultaneously if only educators were provided sufficient resources. Perhaps ironically, policymakers now use highly *g*-loaded tests of achievement to hold schools accountable for achieving the EQuality educationists have said is within their power to produce. Most dramatically, the federal No Child Left Behind Act of 2001 (The White House, 2001) requires public schools not only to close the long-standing demographic gaps in student achievement but to do so by raising all groups of students to the same high level of academic proficiency by 2014: "Schools must be accountable for ensuring that all students, including disadvantaged students, meet high academic standards" (The White House, 2001; see Appendix A, Example x). Schools that fail to level up performance on schedule face escalating sanctions, including state takeover.

The converse of the equalizability fallacy is equally common but far more pernicious—namely, the fallacy that nonequalizability implies nonimprovability. Thus did *Washington Post* columnist Dionne (1994) speak of the "deep pessimism about the possibility of social reform" owing to "the revival of interest in genetic explanations for human inequality" (see Appendix A, Example xi):

> If genes are so important to [inequality of] intelligence and intelligence is so important to [differences in] success, then many of the efforts made over the past several decades to improve people's life chances were mostly a waste of time. (p. A17)

This is utterly false. One can improve lives without equalizing them.

Test-Score-Differences Fallacy 3: Interactionism (Gene-Environment Codependence) Nullifies Heritability

This fallacy portrays the gene–environment partnership in creating a phenotype as if conjoint action within the individual precluded teasing apart the roots of phenotypic differences among individuals. Although the nonfixedness and equalizability fallacies seem to discredit a phenotypic finding (stability of IQ rank within one's age group), the fallacy of so-called "interactionism" provides a scientific-sounding excuse to denigrate as self-evidently absurd all evidence for a genetic influence (Entry 1 in Figure 1.1) on intelligence (Entry 5 in Figure 1.1).

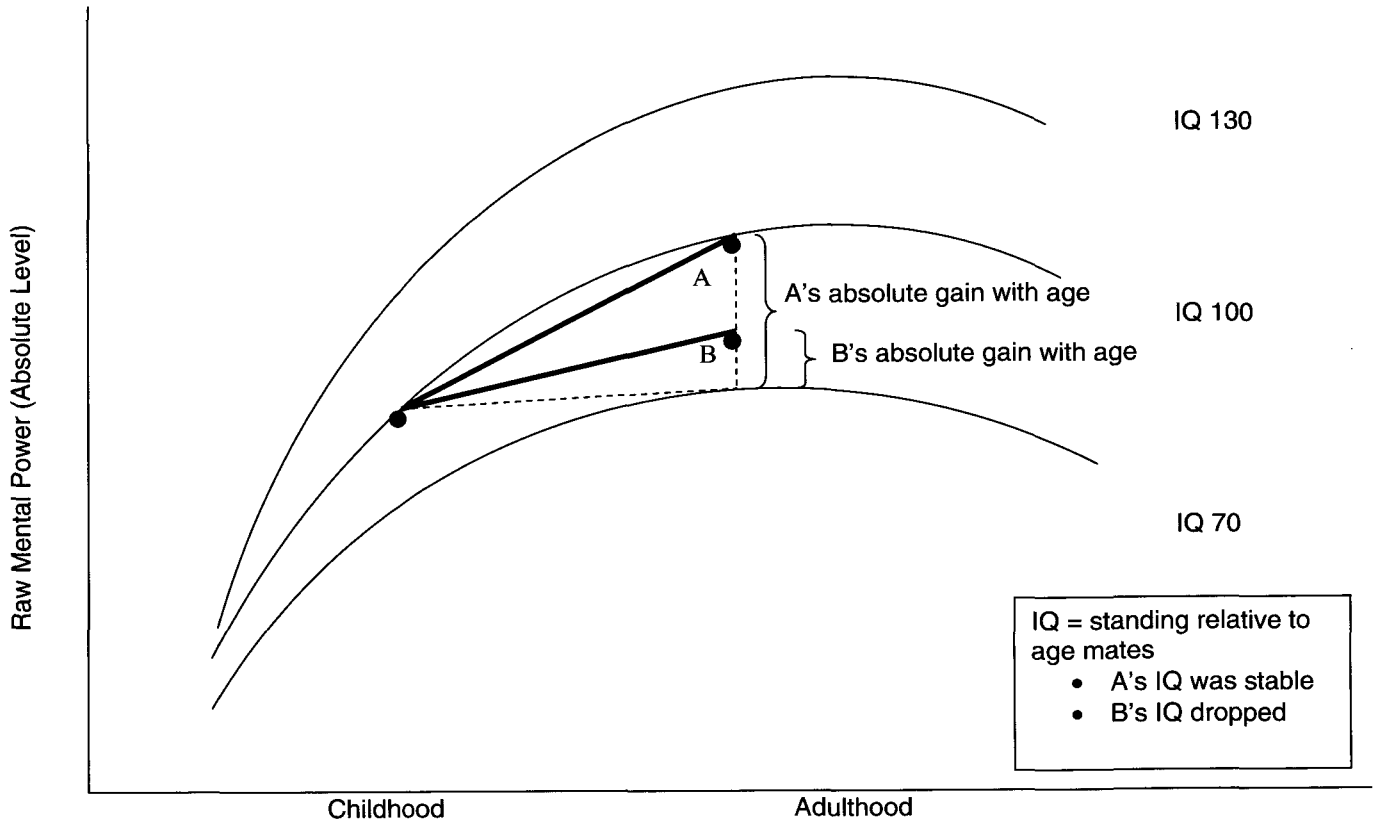

Figure 1.2. An oft-muddled distinction: Changes in relative versus absolute levels of intelligence.

To avoid confusion, I should clarify two concepts. First, *heritability* refers to the origins of observed—*phenotypic*—variation among individuals in a particular population. For example, if the heritability of height differences among White Canadian adult males were estimated to be 80%, this would mean that 80% of these men's differences in height are owing to their differences in genotype and 20% to their differences in environmental (nongenetic) circumstances and error of measurement. Second, the technical term *gene–environment interaction* refers to something altogether different than does the appeal to "interactionism." In behavior genetics, *gene-environment interaction* refers to a particular kind of nonadditive genetic effect in which environmental (nongenetic) effects are conditional on genotype, for example, when possessing a particular version (allele) of a gene renders the individual unusually susceptible to a particular pathogen.

The interactionism fallacy states an irrelevant truth to reach an irrelevant conclusion in order to dismiss peremptorily all estimates of heritability while appropriating a legitimate scientific term to connote scientific backing. The irrelevant truth is that an organism's development requires genes and environments to act in concert. The two forces are inextricable, mutually dependent, and constantly interacting. Development is their mutual product, like the dance of two partners. The irrelevant conclusion is that it is therefore impossible to apportion credit for the pair's joint product to each partner separately—say, 40% of the pair's steps to the man and 60% to the woman. The inappropriate generalization is that behavior geneticists cannot possibly do what they claim—namely, to decompose phenotypic variation among individuals within a particular population into its genetic and nongenetic sources of variation. This is analogous to saying that it would be impossible to estimate whether differences in quality of tango performances among American couples is owing more to skill variation among the male partners than to skill variation among the female partners (i.e., genetic vs. nongenetic variation)—or to what extent differences among couples in their quality of performance depend on the chemistry between two partners (i.e., gene–environment interaction).

To illustrate, Sternberg (1997) spoke of how it is "extremely difficult" to separate the genetic and nongenetic sources of variation in intelligence "because they interact in many different ways" (p. 48; see also Appendix A, Example xii). A letter to *Science* (Andrews & Nelkin, 1996) invoked the authority of geneticists and ethicists to dispute the claim that individual differences in intelligence are highly heritable "given the complex interplay between genes and environments" (p. 13; see also Appendix A, Example xiii). Both examples confuse the essentials for development (genes and environments must both be present and work together) with how the two requisites might differ from one person to another and thus head them down somewhat different developmental paths. Sternberg (1997; see also Appendix A, Example xii) implied that estimating heritabilities is absurd by further confusing the issue—specifically, when he likened calculating a heritability (the

ratio of genetic variance to phenotypic variance in a trait) to calculating the average temperature in Minnesota (a simple mean, but all means obscure variability, just as average quality of dancing has no bearing on why some couples dance better than others).

The interactionism fallacy creates its illusion by focusing attention on the preconditions for behavior (the dance requires two partners), as if that were equivalent to examining variation in the behavior itself (some couples dance better than others, perhaps mostly because the men differ in competence at leading). It confuses two important but quite different scientific questions (Jensen, 1981, p. 112): What is the typical course of human development versus to what extent can variations in development be traced to genetic variation in the population?

The field of behavior genetics seeks to explain not the common human theme but variations on it. It does so by measuring phenotypes for pairs of individuals who differ systematically in genetic and environmental relatedness. Such data allow decomposition of phenotypic variation in behavior within a population into its nongenetic (Entry 2 in Figure 1.1) and genetic (Entry 1 in Figure 1.1) sources. The field has actually gone far beyond estimating the heritabilities of traits such as intelligence. For instance, it can determine to what extent the phenotypic covariation between two outcomes, say, intelligence and occupational level in Sweden, represents a genetic correlation between them in that population (Plomin et al., 2001; Plomin & Petrill, 1997; Rowe, Vesterdal, & Rodgers, 1998).

Critics often activate the interactionism fallacy simply by caricaturing the unwanted evidence about heritability. When researchers speak of IQ's heritability, they are referring to the percentage of variation in IQ, the phenotype, which has been traced to genetic variation within a particular population. However, critics transmogrify this into the obviously false claim that an individual's intelligence is "predetermined" or "fixed at birth," as if it were preformed and emerged automatically according to some detailed blueprint, impervious to influence of any sort. No serious scientist believes that today. One's genome is fixed at birth, but its actions and effects on the phenotype are not fixed, predetermined, or predestined. The genome is less like a blueprint than a playbook for responding to contingencies, with some parts of the genome regulating the actions or expression of others depending on cellular conditions, themselves influenced by location in the body, age, temperature, nutrients available, and the like. Organisms would not survive without the ability to adapt to different circumstances. The behavior genetic question is, rather, whether different versions of the same genes (alleles) cause individuals to respond differently in the very same circumstances.

Test-Score-Differences Fallacy 4: Similarity of 99.9% Negates Differences

This fallacy portrays the study of human genetic variation as irrelevant or wrongheaded because humans are 99.9% (or 99.5%) alike genetically, on

average. Of recent vintage, the 99.9% fallacy impugns even investigating human genetic variation by implying, falsely, that a 0.1% average difference in genetic profiles (3 million base pairs) is trivial. (Comparably estimated, the human and chimpanzee genomes differ by about 1.3%.) The fallacy is frequently used to reinforce the claim, as one anthropology textbook explained (Park, 2002; see also Appendix A, Example xiv), that "there are no races" (p. 395). Its author reasoned that if most of that 0.1% genetic variation is among individuals of the same race, then "all the phenotypic variation that we try to assort into race is the result of a virtual handful of alleles" (pp. 397–398). Reasoning in like manner, Holt (1994) editorialized in the *New York Times* that "genetic diversity among the races is miniscule," a mere "residue" of human variation (p. A23; see also Appendix A, Example xv). The implication is that research into racial differences, even at the phenotypic level, is both scientifically and morally suspect. As spelled out by another anthropology text (Marks, 1995), "Providing explanations for social inequalities as being rooted in nature is a classic pseudoscientific occupation" (p. 273; see also Appendix A, Example xvi).

More recent estimates point to greater genetic variation among humans (only 99.5% alike; Hayden, 2007), but any big number will do. The fallacy works by having us look at human variation against the backdrop of evolutionary time and the vast array of species. By this reasoning, human genetic variation is inconsequential in human affairs because we humans are more similar to one another than to dogs, worms, and microbes. The fallacy focuses our attention on the 99.9% genetic similarity that makes us all human, *Homo sapiens*, to distract us from the 0.1% that makes us individuals. Moreover, as illustrated in diverse life arenas, "it is often the case that small differences in the input result in large differences in the final outcome" (Hart, 2007, p. 112).

The identical parts of the genome are called the *nonsegregating genes*, which are termed evolutionarily *fixed* in the species because they do not vary among its individual members. The remaining genes, for which humans possess different versions (alleles), are called *segregating genes* because they segregate (*reassort*) during the production of eggs and sperm. Only the segregating genes are technically termed *heritable* because only they create genetic differences that may be transmitted from parent to offspring generations. Intelligence tests are designed to capture individual differences in developed mental competence, so it is among the small percentage of segregating genes that scientists search for the genetic roots of those phenotypic differences. The 99.9% fallacy would put this search off limits.

Test-Validation Fallacies

Validating a test refers to determining which sorts of inferences may properly be drawn from the test's scores, most commonly whether it mea-

sures the intended construct (such as conscientiousness) or content domain (jet engine repair, matrix algebra) or whether it allows more accurate predictions about individuals when decisions are required (college admissions, hiring). A test may be valid for some uses but not others, and no single study can establish a test's validity for any particular purpose. For instance, Arthur may have successfully predicted which films would win an Oscar this year, but that gives us no reason to believe he can also predict who will win the World Series, the Kentucky Derby, or a Nobel Prize. Further, we certainly should hesitate to put our money behind his Oscar picks next year unless he has demonstrated a good track record in picking winners.

IQ tests are designed to measure a highly general intelligence, and they have been successful in predicting individual differences in just the sorts of academic, occupational, and other performances that a general intelligence theory would lead one to expect (Entry 6 in Figure 1.1). The tests also tend to predict these outcomes better than does any other single predictor, including family background (Ceci, 1996a; Herrnstein & Murray, 1994). This evidence makes it plausible that IQ tests measure differences in a very general intelligence, but it is not sufficient to prove that they do so or that intelligence actually causes those differences in life outcomes.

Test validation, like science in general, works by pitting alternative claims against one another to see which one best fits the totality of available evidence: Do IQ tests measure the same types of intelligence in different racial–ethnic groups? Do they measure intelligence at all, or just social privilege or familiarity with the culture? Advances in measurement have provided new ways to adjudicate such claims. Entries 10 and 11 in Figure 1.1 represent two advances in identifying, isolating, and contrasting the constructs that cognitive tests may be measuring—respectively, factor analysis and latent trait modeling. Both provide tools for scrutinizing tests and test items in action (Entry 9 in Figure 1.1) and asking whether they behave in accordance with one's claims about what is being measured. If not in accord, then the test, the theory it embodies, or both need to be revised and then reexamined. Successive rounds of such psychometric scrutiny reveal a great deal—not only about tests but also about the phenomena they poke and prod into expressing themselves.

Psychometricians have spent decades trying to sort out the phenomena that tests reveal. More precisely, they have been charting the structure, or relatedness, of cognitive abilities as assayed by tests purporting to measure intelligence or components of it. From the first days of mental testing, it was observed that people who do well on one mental test tend to perform well on all others, regardless of item type, test format, or mode of administration. All mental ability tests correlate positively with all others, suggesting that they all tap into the same underlying abilities.

Intelligence researchers developed the method of factor analysis to extract those common factors (Entry 10 in Figure 1.1) from any large, diverse

set of mental tests administered to representative samples of individuals. With this tool, the researchers can ask the following questions: How many common factors are there? Are those factors the same from battery to battery, population to population, age to age, and so on? What kinds of abilities do they seem to represent? Do tests with the same name measure the same construct? Do tests with different names measure different abilities? Intent is no guarantee.

These are not esoteric technical matters. They get to the heart of important questions such as whether there is a single broadly useful general ability versus many independent coequal ones specialized for different tasks and whether IQ batteries measure the same abilities equally well in all demographic groups (answers thus far: only one, and yes). For present purposes, the three most important findings from the decades of factor analytic research (Carroll, 1993) are that (a) the common factors running through mental ability tests differ primarily in level of generality, or breadth of content (from very narrow to widely applicable) for which that factor enhances performance; (b) only one factor, *g*, consistently emerges at the most general level (which Carroll [1993] labeled "Stratum III," the highest level in his model); and (c) the group factors in Carroll's Stratum II, such as verbal or spatial ability, correlate moderately highly with each other because all reflect mostly *g*—explaining why Carroll (1993) referred to them as different "flavors" of the same *g*.

Carroll (1993, p. 641) noted that some of the Stratum II abilities in his model probably coincide with four of H. Gardner's (1983) seven intelligences: linguistic, logical–mathematical, visuospatial, and musical. The remaining three appear to fall mostly outside the cognitive domain: bodily–kinesthetic, intrapersonal, and interpersonal. Carroll (1993, p. 639) also noted that although the Horn–Cattell model (Horn, 1988) claims there are two *g*s, fluid and crystallized, evidence usually locates both at the Stratum II level or finds fluid *g* isomorphic with *g* itself. In like manner, Sternberg's (1997) claim to have found three intelligences also rests, like Horn and Cattell's claim for two *g*s, on stopping the factoring process just below the most general level (Brody, 2003), thus precluding its discovery.

In short, there are many cognitive abilities, but all turn out to be suffused with or built around *g*. Their most important feature, overall, is how broadly applicable they are for performing different tasks, ranging from the all-purpose (*g*) to the narrow and specific (e.g., associative memory, reading decoding, pitch discrimination). The hierarchical structure of mental abilities discovered through factor analysis, represented in Carroll's three-stratum model, has integrated the welter of tested abilities into a theoretically unified whole. This unified system, in turn, allows one to predict the magnitude of correlations among tests and the size of group differences that will be found in new samples.

The *g* factor is highly correlated with the IQ (usually .8 or more), but the distinction between *g* (Entry 10 in Figure 1.1) and IQ (Entry 9 in Figure

1.1) cannot be overstated (Jensen, 1998). The IQ is nothing but a test score, albeit one with social portent and, for some purposes, considerable practical value. However, *g*, is a discovery—a replicable empirical phenomenon, not a definition. It is not yet fully understood, but it can be described and reliably measured. It is not a thing, but a highly regular pattern of individual differences in cognitive functioning across many content domains. Various scientific disciplines are tracing the phenomenon from its origins in nature and nurture (Entries 1 and 2 in Figure 1.1; Plomin et al., 2001) through the brain (Entry 4 in Figure 1.1; Deary, 2000; Jung & Haier, 2007), and into the currents of social life (Entries 6 and 7 in Figure 1.1; Ceci, 1996a; Gottfredson, 1997a; Herrnstein & Murray, 1994; Lubinski, 2004; Williams, 2000). It exists independently of all definitions and any particular kind of measurement.

The *g* factor has been found to correlate with a wide range of biological and social phenomena outside the realm of cognitive testing (Deary, 2000; Jensen, 1998; Jensen & Sinha, 1993), so it is not a statistical chimera. Its nature is not constructed or corralled by how one chooses to define it but is inferred from its patterns of influence, which wax and wane under different circumstances and from its co-occurrence with certain attributes (e.g., reasoning) but not others (e.g., sociability). It is reasonable to refer to *g* as general intelligence because the *g* factor captures empirically the general proficiency at learning, reasoning, problem solving, and abstract thinking—the construct—that researchers and laypersons alike usually associate with the term *intelligence* (Snyderman & Rothman, 1987, 1988). Because the word *intelligence* is used in so many ways and comes with so much political baggage, researchers usually prefer to stick with the more precise empirical referent, *g*.

Discovery of the *g* factor has revolutionized research on both intelligence (the construct) and intelligence testing (the measure) by allowing researchers to separate the two—the phenomenon being measured, *g*, from the devices used to measure it. Its discovery shows that the underlying phenomenon that IQ tests measure (Entry 10 in Figure 1.1) has nothing to do with the manifest content or format of the test (Entry 8 in Figure 1.1): It is not restricted to paper-and-pencil tests, timed tests, ones with numbers or words, or academic content, or any particular format. The active ingredient in intelligence tests is something deeper and less obvious—namely, the cognitive complexity of the various tasks to be performed (Gottfredson, 1997b). The same is true for tests of adult functional literacy—it is complexity and not content or readability per se that accounts for differences in item difficulty (Kirsch & Mosenthal, 1990).

This separation of phenomenon from measure also affords the possibility of examining how well different tests and tasks measure *g* or, stated another way, how heavily each draws on or taxes *g* (how *g* loaded each is). To illustrate, the WAIS Vocabulary subtest is far more *g* loaded than the Digit Span subtest (.83 vs. .57; Sattler, 2001, p. 389). The more *g* loaded a test or task, the greater the edge in performance it gives individuals of higher *g*. Just

as individuals can be characterized by *g* level, tests and tasks can be characterized by their *g* loadings and thereby show us which task attributes ratchet up their cognitive complexity (amount of distracting information, number of elements to integrate, inferences required, etc.). Such analyses would allow more criterion-related interpretations of intelligence test scores as well as provide practical guidance for how to reduce unnecessary complexity in school, work, home, and health, especially for lower *g* individuals. We may find that tasks are more malleable than people, *g* loadings more manipulable than *g* level.

All mental tests, not just IQ test batteries, can be examined for how well each measures not just *g* but something in addition to *g*. Using hierarchical factor analysis, psychometricians can strip the lower order factors and tests of their *g* components to reveal what each measures uniquely and independently of all other tests. This helps to isolate the contributions of narrower abilities to overall test performance because they tend to be swamped by *g*-related variance, which is usually greater than it is for all the other factors combined. Hierarchical factor analysis can also reveal which specialized ability tests are actually functioning mostly as surrogates for IQ tests and to what degree. Most tests intended to measure abilities other than *g* (verbal ability, spatial perception, mathematical reasoning, and even seemingly noncognitive abilities such as pitch discrimination) actually measure mostly *g*, not the specialized abilities that their names suggest. This is important because people often wrongly assume that if there are many kinds of tests, each intended to measure a different ability, then there must actually be many independent abilities—like different marbles. This is not true.

All the factor analyses mentioned thus far use *exploratory factor analysis*, which extracts a parsimonious set of factors to explain the commonalities running through tests and causes them to intercorrelate. It posits no constructs but waits to see which dimensions emerge from the process (Entry 10 in Figure 1.1). It is a data reduction technique, which means that it provides fewer factors than tests to organize test results in a simpler, clearer, more elegant manner. The method has been invaluable for pointing to the existence of a general factor, although without guaranteeing one.

Another measurement advance has been to specify theoretical constructs (ability dimensions) before conducting a factor analysis and then determine how well the hypothesized constructs reproduce the observed correlations among tests. This is the task of *confirmatory factor analysis*. It has become the method of choice for ascertaining which constructs a particular IQ test battery taps (Entry 11 in Figure 1.1), that is, its construct validity. Variations of the method provide a new, more exacting means of vetting tests for cultural bias (lack of construct invariance).

The following test-validation fallacy would seem to indicate, however, that nothing important has been learned about intelligence tests since Binet's time to sweep aside a century of construct validation. It ignores the discovery

of *g* and promotes outdated ideas to dispute the possibility that IQ tests could possibly measure such a general intelligence.

Test-Validation Fallacy: Contending Definitions Negate Evidence

This fallacy portrays lack of consensus in verbal definitions of intelligence as if this negated evidence for the construct validity of IQ tests. Critics of intelligence testing frequently suggest that IQ tests cannot be presumed to measure intelligence because scholars cannot agree on a verbal definition or description of it. By this reasoning, one could just as easily dispute that gravity, health, or stress can be measured. Scale construction always needs to be guided by some conception of what one intends to measure, but careful definition hardly guarantees that the scale will do so, as noted earlier. Likewise, competing verbal definitions do not negate either the existence of a suspected phenomenon or the possibility of measuring it. What matters most is not unanimity among proposed definitions or descriptions but construct validation or "dialogue with the data" (Bartholomew, 2004, p. 52).

Insisting on a consensus definition is an excuse to ignore what has already been learned, especially about *g*. To wit,

> Intelligence is an elusive concept. While each person has his or her own intuitive methods for gauging the intelligence of others, there is no a prior definition of intelligence that we can use to design a device to measure it. (Singham, 1995, p. 272; see also Appendix A, Example xvii)

Thus Singham (1995) suggested that everyone recognizes the phenomenon but that it will nonetheless defy measurement until we all agree on how to do so—which is never. Expanding on its critique of *The Bell Curve*, the editorial page of the *New York Times* ("The 'Bell Curve' agenda," 1994) stated, "Further, there is wide disagreement about what intelligence consists of and how—or even if—it can be measured in the abstract" (p. A16; see also Appendix A, Example xviii). The editorial had just remarked on the wide agreement among intelligence researchers that mental acuity—the supposedly unmeasurable—is influenced by both genes and environments.

Critics often appeal to the intelligence-is-marbles fallacy to propose new, "broadened conceptions" of intelligence, as if pointing to additional human competencies nullified the demonstrated construct validity of IQ tests for measuring a highly general mental ability, or *g*. Some such calls for expanding test batteries to capture more "aspects" or "components" of intelligence, more broadly defined, make their case by confusing the construct validity of a test (does it measure a general intelligence?) with its utility for predicting some social or educational outcome (how well does it predict job performance?). Much else besides *g* matters, of course, in predicting success in training, on the job, and any other life arena, which is why personnel selection professionals (e.g., Campbell & Knapp, 2001) routinely advise that selection batteries include a variety of cognitive and noncognitive measures

(e.g., conscientiousness). General intelligence is hardly the only useful human talent, nor need everything good be labeled intelligence to be taken seriously.

Yet critics implicitly insist that it be so when they cite the limited predictive validity of *g*-loaded tests to argue for "broadened conceptions of intelligence" before we can take tests seriously. One such critic, Rosenblum (1996), said this change would "enable us to assess previously unmeasured aspects of intelligence" (p. 622) as if, absent the relabeling, those other aspects of human competence are not or cannot be measured. He then chided researchers who, in "sharp contrast . . . stress validities of the traditional intelligence tests" (that is, stress what g-loaded tests do moderately well) and for "oppos[ing] public policies or laws" that would thwart their use in selection when tests do not provide, in essence, the be-all-and-end-all in prediction (see also the imperfect-prediction fallacy later in this chapter).

CAUSAL NETWORKS

Entries 1 through 7 in Figure 1.1 represent the core concepts required in any explanation of the causes of intelligence differences in a population (*vertical* processes, Entries 1–5 in Figure 1.1; Jensen, 1998) and the effects they produce on it collectively and its members individually (*horizontal* processes, Entries 5–7 in Figure 1.1). This schema is obviously a highly simplified rendition of the empirical literature (e.g., by omitting feedback processes and other personal traits that influence life outcomes), but its simplicity helps to illustrate how fundamental are the confusions perpetuated by the following three causal network fallacies.

Causal-Network Fallacies

Causal-Network Fallacy 1: Phenotype *Equals* Genotype

This fallacy portrays phenotypic differences in intelligence (Entry 5, Figure 1.1) as if they were necessarily genotypic (Entry 1, Figure 1.1). Intelligence tests measure only observed or *phenotypic* differences in intelligence. In this regard, IQ tests are like the pediatrician's scale for measuring height and weight (phenotypes). They allow physicians to chart a child's development, but such scales, by themselves, reveal nothing about why some children have grown larger than others. Differences in intelligence can likewise be real without necessarily being genetically caused, in whole or part. Only genetically informative research designs can trace the roles of nature and nurture in making some children larger or smarter than others. Such designs might include identical twins reared apart (same genes, different environments), adopted children reared together (different genes, same environment),

and other combinations of genetic and environmental similarity in order to determine whether similarity in outcomes within the pairs follows similarity of their genes more closely than it does their similarity in environments. Nonexperimental studies including only one child per family tell us nothing about the genetic or nongenetic roots of human variation.

The default assumption in all the social sciences, including intelligence testing research, is therefore that one is speaking only of phenotypes when describing developed differences among individuals and groups—unless one explicitly states otherwise. The phenotype–genotype distinction, which often goes without saying in scholarly circles, is not obvious to the public, however. Indeed, the average person may perceive the distinction as mere hairsplitting, because scientific research and common intuition both point to various human differences being heavily influenced by one's fate in the genetic lottery. In fact, it is now well established that individual differences in adult height and IQ—within the particular races, places, and eras studied so far—can be traced mostly to those individuals' differences in genetic inheritance (Plomin et al., 2001).

News of genetic causation of phenotypic variation in these peoples, places, and times primes the public to accept the fallacy that all reports of real differences are ipso facto claims for genetic ones. The *phenotype-equals-genotype* fallacy thus exposes scholars to false allegations that they are actually asserting genetic differences whenever they fail to repudiate them pointedly. For example, critics often insinuate that scientists who report racial gaps in measured intelligence (Entry 5 in Figure 1.1) are thereby asserting "innate" (genetic) differences (Entry 1 in Figure 1.1) between the races.

Duster (1995; see also Appendix A, Example xix) provided a fairly subtle example. In the context of discussing "those making the claims about the genetic component of an array of behavior and conditions (crime, mental illness, alcoholism, gender relations, intelligence)," he referred to "a sociologist, Robert Gordon (1987), who argues that race differences in delinquency are best explained by IQ differences between the races" (Duster, 1995, p. 1). Gordon's article, however, discussed only phenotypes, specifically, whether socioeconomic status or IQ is the better predictor of Black–White differences in crime and delinquency.

Some scholars have tried to preempt such false attributions by taking pains to point out that they are not claiming genetic causation for the phenotypic differences they observe, race related or not. Testing companies routinely evade the attribution by going further (Camara & Schmidt, 1999, p. 13). They align themselves with strictly nongenetic explanations by routinely blaming lower tested abilities and achievements on social disadvantages such as poverty and poor schooling, even when facts say otherwise for the population in question—for example, despite the evidence, for Whites in general, that shared family effects on the IQ and achievement of siblings mostly fade away by adolescence and that there are sizeable genetic correla-

tions among IQ, education, and social class in adulthood (Plomin & Petrill, 1997; Rowe, 1997; Rowe et al., 1998).

The *phenotype-*equals*-genotype* fallacy is reinforced by the confusion, noted earlier, between two empirical questions: (a) Do IQ differences represent real differences in ability, or, instead, do IQ tests mismeasure intelligence? For example, are they biased against certain races? (b) If the measured differences are real differences in intelligence, what causes them? For example, does poverty depress intelligence? The first question concerns a test's construct validity for measuring real differences; the second question concerns the role of nature and nurture in creating them. Even highly rigorous scholars can be read as confusing the two questions:

> A test is biased if it gives an advantage to one group rather than the other. In other words, we cannot be sure whether the score difference is due to ability to do the test or to environmental factors which affect the groups differently" (Bartholomew, 2004, pp. 122–123; see also Appendix A, Example xx)

This fallacy was also greatly reinforced by public commentary following publication of *The Bell Curve*. Although the book analyzed strictly phenotypic data, both its friends and detractors used its results on the relative predictive power of IQ versus social class to debate the relative contributions to social inequality of genes versus environments. They did this when they used IQ differences as a stand-in for genetic differences and social class as a stand-in for nongenetic influences. For example, one economist argued that Herrnstein and Murray "grossly underestimate the relative effect of environment versus intelligence in accounting for individual differences in various dimensions of achievement" (Loury, 1995, p. 19).

Causal-Network Fallacy 2: Biological *Equals* Genetic

This fallacy portrays biological differences (such as brain phenotypes, Entry 4 in Figure 1.1) as if they were necessarily genetic (Entry 1 in Figure 1.1). This is a corollary of the *phenotype-*equals*-genotype* fallacy, because an organism's observed form and physiology are part of its total phenotype. Like height and weight, many aspects of brain structure and physiology (Entry 4 in Figure 1.1) are under considerable genetic control (Entry 1 in Figure 1.1), but nongenetic differences, say, in nutrition or disease (Entry 2 in Figure 1.1) can also produce variation in these physical traits. When authors use the terms *biological* and *genetic* interchangeably (Bartholomew, 2004; see also Appendix A, Example xxi), they confuse phenotype with genotype.

Research in behavior genetics does, in fact, confirm a large genetic contribution to phenotypic differences in IQ, brain biology, and correlations between the two (Deary, 2000; Jensen, 1998). The genetic correlations between IQ and various brain attributes suggest potential mechanisms by which

genes could influence speed and accuracy of cognitive processing, yielding a higher intelligence. However, they do not rule out nongenetic effects. Instead, they tilt plausibility toward certain nongenetic mechanisms (micronutrients, etc.) and away from others (teacher expectations, etc.).

So far, however, this growing network of psychometric and biological evidence exists only for Whites. Extant evidence confirms mean racial differences in phenotypic intelligence and a few brain attributes, such as head size, but since the 1970s no scientific discipline has been willing to conduct genetic or brain research on non-White populations that could be tied to intelligence. The evidence for genetic influence on differences within the White population enhances the plausibility of a part-genetic-component rather than a no-genetic-component explanation for the average White–Black difference in phenotypic intelligence. Scholars legitimately differ in how skewed the evidence must be before they provisionally accept one hypothesis over another or declare a scientific contest settled. Nonetheless, until scientists are willing to conduct the requisite research, it remains fallacious to suggest that average racial differences in intelligence and brain physiology are necessarily owing to genetic differences between the races.

When scientists seem to overstate the evidence for a "controversial" conclusion, or are falsely depicted as doing so, their seeming overstatement is used to damage the credibility not only of that conclusion but also of all similar-sounding ones, no matter how well validated scientifically the latter may be and even when they have nothing to do with race—for example, the conclusion that IQ differences among Whites are substantially heritable.

The *biological*-equals-*genetic* corollary will become more common as knowledge of the physiological correlates of intelligence spreads. Protagonists in the nature–nurture debate have long conceptualized environmental influences as educational and cultural: Favorable social environments deliver more bits of skill and knowledge or they enhance the mind's learning and reasoning software. Judging by my own students' reactions, all mental behaviors that do not have any immediately obvious cultural origin (e.g., choice reaction time) tend to be perceived as necessarily genetic, as is everything physiological (e.g., brain metabolism). Treating the terms *biological* and *genetic* as synonyms reflects an implicit hypothesis, plausible but unproved. This implicit hypothesis may explain the strident efforts to deny any link between brain size and intelligence (e.g., Gould, 1981), as well as the just plain silly ones (Race and Intelligence, 2007; see also, Appendix A, Example xxii)—for example, that we should have "serious doubts" about such research because Albert Einstein "had a brain slightly below average for his size" (Race and Intelligence, 2007).

Causal-Network Fallacy 3: Environmental *Equals* Nongenetic

This fallacy portrays environments (Entry 3 in Figure 1.1) as if they were necessarily nongenetic (Entry 2 in Figure 1.1)—that is, unaffected by

and unrelated to the genotypes of individuals in them. It is the environmentalist counterpart to the hereditarian *biological*-equals-*genetic* fallacy. Environments are physically external to individuals, but, contrary to common belief, this does not make them independent of genes. Individuals differ widely in interests and abilities, partly for genetic reasons; individuals select, create, and reshape their personal environments according to their interests and abilities; therefore, as behavior genetic research has confirmed, differences in personal circumstances (e.g., degree of social support, income) are likewise somewhat genetically shaped (Entry 1 in Figure 1.1). Both childhood and adult environments (Entries 3 and 6 in Figure 1.1) are therefore influenced by the genetic proclivities of self and genetic kin. People's personal environments are their extended phenotypes (Dawkins, 1999).

Near-universal deference in the social sciences to the *environmental*-equals-*nongenetic* fallacy has fostered mostly causally uninterpretable research (on socialization theory, see Scarr, 1997; on family effects theory and passive learning theory, see Rowe, 1997). It has also freed testing critics to misrepresent the phenotypic correlations between social status and test performance as prima facie evidence that poorer environments, per se, cause lower intelligence. In falsely portraying external environments as strictly nongenetic, critics inappropriately commandeer all IQ–environment correlations as evidence for pervasive and powerful nongenetic causation.

Describing strictly phenotypic studies in this vein, *The Chronicle of Higher Education* (Monastersky, 2008) reported that "the new results from neuroscience indicate that experience, especially being raised in poverty, has a strong effect on the way the brain works" (¶ 1; see also Appendix A, Example xxiii). The article quoted one of the researchers as saying, "It's not a case of bad genes." It is likely, however, that the study participants who lived in better circumstances performed better because they had genetically brighter parents. Brighter parents tend to have better jobs and incomes and also to bequeath their offspring more favorable genes for intelligence. Parental genes can also enhance offspring performance more directly if they induce parents to create more cognitively effective child-rearing environments. In none of the studies had the investigators ruled out such genetic contributions to the child's rearing "environment." Adherence to the *environmental*-equals-*nongenetic* fallacy remains the rule, not the exception, in social science research.

Fischer et al. (1996) illustrated this fallacy when they argued that scores on the military's AFQT reflect differences not in intellectual ability but in the environments to which individuals have been exposed: "Another way to understand what we have shown is that test takers' AFQT scores are good summaries of a host of prior experiences (mostly instruction) that enable someone to do well in adult life" (Fischer et al., 1996, p. 68; see also Appendix A, Example xxiv).

Helms (2006) used the *environmental*-equals-*nongenetic* fallacy to argue a different point. Whereas Fischer et al. (1996) used it to claim that *g*-loaded

tests measure exposure to knowledge that we ought to impart equally to all, Helms used it to argue that the Black–White IQ gap reflects culturally caused differences in performance that have nothing to do with intellectual competence. In particular, racial differences in test scores must be presumed, "unless research proves otherwise, to represent construct irrelevant variance," that is, "systematic variance, attributable to the test taker's psychological characteristics, developed in response to socialization practices or environmental conditions" (p. 847). To make this claim, Helms must treat individuals as passive receivers of whatever influence happens by.

When combined, the three causal network fallacies can produce more convoluted ones. As noted earlier, protagonists in *The Bell Curve* debate often conjoined the phenotype-is-genotype fallacy with the *environmental-equals-nongenetic* fallacy when they used strictly phenotypic data to debate whether genes or environments create more social inequality.

POLITICS OF TEST USE

The previous sections on the measurement and correlates of cognitive ability have been directed to answering one question: What do intelligence tests measure? That is a scientific question with an empirical answer. However, the question of whether a cognitive test should be used to gather information for decision-making purposes is an administrative or political choice.

Standards-of-Evidence Fallacies

The decision to administer a test for operational purposes should rest on good science—principally, evidence that the test is valid for one's intended purpose. For example, does the proposed licensing exam accurately screen out practitioners who would endanger their clients, or would an IQ test battery help diagnose why failing students are failing? Validity is not sufficient reason for testing, however. The utility of tests in applied settings depends on practical considerations as well, including feasibility and cost of administration, difficulties in maintaining test security and operational validity, vulnerability to litigation or misuse, and acceptability to test takers (Murphy & Davidshofer, 2005). Valid tests may not be worth using if they add little to existing procedures, and they can be rendered unusable by high costs, chronic legal challenge, adverse publicity, and unintended consequences.

When used for operational purposes, testing is an intervention. Whether it be the aim of testing or just its consequence, test scores (Entry 9 in Figure 1.1) can influence the tested individuals' life chances (Entry 6 in Figure 1.1). This is why good practice dictates that test scores (or any other single indicator) be supplemented by other sorts of information when making decisions

about individuals, especially decisions that are irreversible and have serious consequences. Large-scale testing for organizational purposes can also have societal-level consequences (Entry 7 in Figure 1.1). For example, although personnel selection tests can improve workforce productivity, their use changes who has access to the best jobs.

Nor is the choice not to test a neutral act. If testing would provide additional valid, relevant, cost-effective information for the operational purpose at hand, then opting not to test constitutes a political decision to not consider certain sorts of information and the decisions they would encourage. Like other social practices, testing—or not testing—tends to serve some social interests and goals over others. This is why testing comes under legal and political scrutiny and why all sides seek to rally public opinion to their side to influence test use. Therefore, just as testing can produce a chain of social effects (Entry 9 → Entry 6 → Entry 7 in Figure 1.1), public reactions to those effects can feed back to influence how tests are structured and used, if at all (Entry 7 → Entry 8 → Entry 9 in Figure 1.1).

The measurement and causal-network fallacies described earlier are rhetorical devices that discourage test use by seeming to discredit scientifically the validity of intelligence tests. They fracture logic to make the true seem false and the false seem true in order to denigrate one or more of the three facts on which the democratic dilemma rests—the phenotypic reality, limited malleability, and practical importance of *g*. However, they otherwise observe the rules of science: Ideas must compete, and evidence matters.

The following standards-of-evidence fallacies violate these rules in the guise of honoring them. They accomplish this by invoking criteria for assessing the practical utility of tests as if they were criteria for assessing the scientific validity of the information they provide. This then allows critics to ignore the rule for adjudicating competing scientific claims—the preponderance of evidence, or which claim best accounts for the totality of relevant evidence to date. In this manner, critics can shelter favored ideas from open scientific contest while demanding that tests and test results meet impossibly rigorous scientific standards before their use can be condoned.

Scientific double standards are commonly triggered, for example, by insinuating that certain scientific conclusions pose special risks to the body politic. In other words, the standards-of-evidence fallacies invoke a criterion for test utility (alleged social risk) to justify their demand that particular tests or ideas be presumed scientifically inferior to—less valid than—all competitors until they meet insurmountable quality standards. Social risks must be weighed, of course, but for what they are—as elements in a political decision—and not as indicators of technical quality.

Standards-of-Evidence Fallacy 1: The Imperfect Measurement Pretext

This fallacy maintains that valid, unbiased intelligence tests should not be used for making decisions about individuals until the tests are made error

free. It labels highly *g*-loaded tests as flawed because they are not error free (reliability <1.0, or predictive validity <1.0). Nothing in human affairs is without error, of course, but the implication is that such tests allow socially unacceptable errors, even when they reduce error overall. The implied flaw is usually that tests rule out some candidates who would actually have performed well if hired or admitted (*false negatives*). Concern usually focuses on minority false negatives, in particular, even though valid tests tend to reduce such decision errors. The insinuation, however, is that valid, unbiased tests are biased, which allows opponents to call for suspending their use until they are cleansed of such "flaws."

FairTest (2007) argued just that: "[ACT] test scores should be optional in college admissions" because "ACT scores are imprecise" and the "ACT's flaws have serious consequences" (¶¶ 12, 14, 17; see also Appendix A, Example xxv). In like manner, an article in *The Chronicle of Higher Education* (Miller, 2001) reported that "educational researchers have begun describing testing's dark side": "Standardized tests, they say, are too limited, too imprecise, and too easily misunderstood to form the basis of crucial decisions about students. . . . [A] reliability of .9 ain't all it's cracked up to be" (p. A14; see also Appendix A, Example xxvi). Such critics express no concern over the precision or reliability of testing's alternatives, such as holistic admissions, which also have serious consequences but rest on subjective judgments framed as "individualized reviews" or assessments of the "whole person." As described by Jaschik (2007) in the journal *Inside Higher Ed*,

> In holistic admissions, colleges evaluating applicants replace grids of grades and test scores with more individualized reviews of would-be students. The practice is most commonly associated with liberal arts colleges or with public universities at which affirmative action has been banned. (¶ 1)

Although the imperfection pretext is most often used to justify eliminating valid tests, it is sometimes offered as a seemingly scientific rationale to increase a test's measurement error for the purpose of social leveling. For example, imperfect reliability of measurement is the rationale given for test score *banding*, which groups broad swaths of unequally qualified job applicants together as equally qualified (Cascio, Outtz, Zedeck, & Goldstein, 1991). Its purpose is to reduce disparate impact, and it does so by throwing away valid information, thereby reducing a test's reliability and validity. In like manner, the National Research Council of the National Academy of Sciences cited imperfect predictive validity to justify its 1989 recommendation that valid, unbiased employment tests be race normed (Hartigan & Wigdor, 1989; see also Appendix A, Example xxvii). Race norming reduces disparate impact by introducing systematic error designed to favor lower scoring races and disfavor higher scoring ones.

The demand for technical improvement is clearly pretextual. Testing is hardly the only useful source of information about students and employees,

but few are as reliable, construct valid, and predictive in education and employment settings as are *g*-loaded tests. Using a *g*-loaded test generally results in fewer false negatives (and fewer false positives) than not using one because the alternatives to testing tend to be less valid. Increasing a *g*-loaded test's validity would reduce the rate of false negatives (and false positives) in all groups, to be sure, but would thereby more accurately distinguish between less and more able individuals. As noted earlier, increasing the accuracy of a *g*-loaded test generally increases, not decreases, its disparate impact. This is why the fallacy demands perfect measurement—superior measurement is precisely what must be avoided.

This is not to say that all kinds of error are equal in affecting test utility, as illustrated by the trade-offs in medical diagnostics between *test specificity* (proportion of true negatives detected; e.g., true absence of HIV) and *test sensitivity* (proportion of true positives detected; e.g., actual presence of HIV). Balancing different kinds of error is a political, monetary, or ethical decision, however, not a technical one. The imperfection fallacy provides a pretext for imposing a political choice among social goods in the guise of insisting on greater scientific accuracy.

Standards-of-Evidence Fallacy 2: The Dangerous-Thoughts Trigger

This fallacy maintains that scientific conclusions purported to be divisive or dangerous should not be entertained until proved beyond all possible doubt. It sets as selective and insurmountable an evidentiary standard for unwelcome scientific conclusions as does the imperfection standard for *g*-loaded tests. Under this fallacy, opponents insinuate that an idea is fraught with danger to press their case for one-sided scientific rigor. The putative danger is rarely explained but connoted by allusions to physical harm (dangerous sports, risky human experimentation, genocide, etc.). Labeling a well-validated scientific conclusion dangerous allows any fear, any manufactured doubt, to trump the preponderance of the evidence for it, no matter how lopsided the evidence may be. The implicit premise seems to be that unsettling truths do no good and comforting lies no harm.

The dangerous-thoughts standard has appeared in various forms over the years. When rules governing research with human subjects were first formulated in the 1970s, there was an effort to bar research posing questions or suggesting answers that might offend minority groups. Many journal editors and manuscript reviewers act on the same impulse, and occasionally an editor will reject a submission explicitly on the grounds that "divisive" research should not be published unless it meets the most exacting technical standards. In the guise of heightened scientific rigor, the dangerous-thoughts fallacy shelters comforting ideas from competition. It is applied most aggressively today to stifle reportage and discussion of racial gaps on intelligence tests, especially their possible genetic component (Gottfredson, 2007).

In rejecting a manuscript testing the hypothesis that Black–White IQ differences represent differences in g, the editor of *American Psychologist* (C. Kiesler, 1980, personal communication to A. R. Jensen, January 17, 1980; see Appendix A, Example xxviii) explained to the author that because "this area is so controversial and important to our society, I should not accept any manuscript that is less than absolutely impeccable." Given the "hanging implication" of a genetic difference, one has to "assure one's self [*sic*] that other possibilities are not possible or plausible." More than 2 decades later, Hunt and Carlson (2007) used the same rationale for recommending special treatment of research on group differences: "We do not see any need for [Jensen's] potentially divisive 'default hypothesis' . . . in the absence of convincing evidence that rules out other hypotheses" (p. 210; see also Appendix A, Example xxix). In both cases, the author was expected to prove his hypothesis beyond all conceivable doubt before it would be allowed even to compete in the scientific arena. Simply insinuating harm is usually sufficient to trigger impossible scientific standards.

Authors are sometimes asked to pull the dangerous-thoughts trigger on themselves, ostensibly in the name of scientific balance. For example, when I recently submitted a manuscript analyzing the systematic public misrepresentation of current intelligence research, one reviewer asked that I also discuss the "sordid history of intelligence testing." Acceding to such requests does not enhance scientific balance but selectively burdens the research at hand by morally tainting it through guilt by association. The reviewer offered no examples, perhaps assuming them to be obvious.

Critics usually draw on two kinds of examples: either accusations that leading intelligence researchers have been scientifically dishonest or that their science has abetted mass oppression or murder. However, none of these lurid accusations have withstood scrutiny. In fact, my reading of the historical accounts (e.g., Anderson, 1997; Carroll, 1995; Lohman, 1997; Snyderman & Herrnstein, 1983; White, 2000; Wigdor & Green, 1991, chap. 1) and the archives of psychology (e.g., Terman, 1928, and other articles in the same volume) is that much of the field's supposedly sordid history has been, and continues to be, manufactured by the field's detractors. For instance, the claim that Cyril Burt committed scientific fraud now seems fraudulent itself (Fletcher, 1991; Joynson, 1989; Rushton, 2002; Samelson, 1992), and the only ideologically motivated mismeasurements of human skulls that Gould (1981, 1996) demonstrated were his own (Michael, 1988; Rushton, 1997; Samelson, 1982).

The reviewer mentioned earlier eventually suggested that I discuss eugenics as one example. Because eugenics has been associated in the public mind with genocide, I was, in essence, being asked to reinforce a falsehood that would morally taint my own scientific analysis. The falsehood is that claiming there is scientific evidence for genetic differences in valued human traits encourages mean-spiritedness and inhumane public policy, whereas environmentalism promotes justice and compassion. This notion has gained

false credibility (hence the request) partly because critics deploy historical examples selectively to taint as immoral the scientific evidence that they would override. For instance, they fail to acknowledge that environmentalist ideologies, not just hereditarian ones, have been used to justify genocide: Stalin's Communist Union of Soviet Socialist Republics and Pol Pot's Communist Cambodia, not just Hitler's Fascist Germany, perpetrated unspeakable atrocities to reshape their citizenries.

Standards-of-Evidence Fallacy 3: Happy-Thoughts Leniency

This fallacy maintains that mere theoretical possibility elevates the scientific credibility of a politically popular idea above that of an empirically plausible but unpopular conclusion. It is the obverse of the dangerous-thoughts fallacy: If some scientific conclusions must not be entertained until proved beyond all possible doubt, then their competitors may be accepted if they simply offer hypothetical situations that seem to contradict the evidence they would have us ignore. In the dangerous-thoughts examples, we saw arguments that genetic differences between races may not be entertained until conclusively proved, no matter the preponderance of evidence. Diamond (1999) illustrated happy-thoughts leniency by beginning with the "seemingly compelling" evidence that Australian Aborigines are observed to have lower intelligence than White immigrants to the continent because they have less favorable genes (p. 19; see also Appendix A, Example xxx). He then invoked the dangerous-thoughts fallacy to dismiss that evidence when he stated, "The objection to such racist explanations is not just that they are loathsome, but also that they are wrong" (p. 19). Finally, he introduced the happy alternative to which he invited us to ascribe greater credibility: "In fact . . . modern 'Stone Age' peoples are on the average probably more intelligent, not less intelligent, than industrialized peoples" (p. 19).

The most famous example of hypothesizing an implausible alternative reality is Lewontin's (1976) thought experiment about growing two handfuls of seed corn, one under excellent conditions and the other being deprived of essential nutrients. He used his thought experiment to argue that because it is possible experimentally to induce a 100% heritability for height differences within both groups but a 0% heritability for resulting height differences between them, we must dismiss the high within-race heritabilities of IQ as having no bearing on questions of racial differences or the malleability of intelligence. By Lewontin's reasoning, what is theoretically possible but empirically implausible for humans—namely, no genetic differences among races and no constraints on the malleability of intelligence—ought to be considered the most plausible scientific stance, until proved otherwise. In explaining why we can ignore the failures of compensatory education, Lewontin argued, "It is empirically wrong to argue that, if the richest environmental experience we can conceive does not raise IQ substantially, that we have exhausted the environmental possibilities " (p. 91).

Lewontin's (1976) happy possibility is commonly invoked for the same purposes. Consider an editorial in the *New York Times* ("The 'Bell Curve' Agenda," 1994) that sought to discredit *The Bell Curve*'s conclusion that intelligence has limited malleability:

> An example proves the point. Plants grown together under ideal conditions will achieve different heights based solely on individual genetic makeup. But lock half the plants in a dark closet and the difference in average height of the two groups will be due entirely to environment. (p. A16; see also Appendix A, Example xxxi)

Unlike Lewontin's hypothetical plants, humans are not randomly assigned to environments but to some extent select and shape their personal circumstances. His thought experiment thus works partly by inviting us to commit the *environmental*-equals-*nongenetic* fallacy.

The happy-thoughts fallacy also works by appealing to a false presumption rarely questioned—namely, that genetic influences limit human freedom and equality, whereas environmental influences do not. In my undergraduate courses on intelligence, I assign a paper in which I ask students to imagine changing any single fact about intelligence they wish. The charge is to describe what their new world will look like. Many choose to make intelligence differences entirely environmental. Yet instead of discovering their expected utopia, many find the opposite. In their new world, there is no genetically conditioned resilience to poor environments and no rising out of poverty, there is more assortative mating by wealth, and there are even laws holding parents accountable for children's "failure to learn." The students realize that genetic differences between parents and children guarantee some social mobility, but environmental causation does not. So just as supposedly dangerous thoughts about intelligence seem less fearsome when looked in the eye, so, too, do happy thoughts look less appealing.

Experts themselves sometimes seem to accept the three standards-of-evidence fallacies. They seem at once seduced by the appeal to scientific rigor and intimidated by the presumption that their ideas do social harm. Instead of rejecting the double standards, they seem defensive about not meeting them. In not questioning or probing the premise that democratic citizenries must be protected from certain ideas, they acquiesce to it.

CONCLUSION

All human groups exhibit large, enduring variations in intelligence that they must somehow accommodate for collective benefit. Mechanisms for accommodation evolve, as they must, when small populations grow and formerly distinct ones mix and jostle. The fallacies about intelligence testing all work to deny the need for accommodation by focusing hostility on the test-

ing enterprise, as if it were responsible for human inequality. This explains why its critics prefer to focus on intelligence testing's technical flaws, even though it has fewer than the alternatives they favor. This also explains why critics often respond to mounting scientific support for its construct validity, predictive value, and lack of bias with yet more strident critiques of the tests, test results, and persons giving them credence.

The 13 fallacies I have described seem to hold special power in the public media, academic journals, college textbooks, and the professions. I have also observed them frequently in conversations with journalists, college students, practitioners, and scholars in diverse fields. My aim in dissecting them has been to show how they work to persuade. Fallacies are tricks of illogic to protect the false from refutation. This is why they are more persuasive and more corrosive than outright falsehoods. Experts usually sense that fallacious arguments are specious but do not engage them for precisely that reason. Researchers would rather parse the evidence, not faulty reasoning about it. However, illogic does not yield to their showings of empirical evidence. Sophistry is best dealt with by recognizing it for what it is: arguments whose power to persuade resides in their logical flaws.

What can be done? First, fallacies must be anticipated. Not only is everyone susceptible to them, but antitesting fallacies are avidly pressed on the public. As teachers know, students do not come to academic subjects as blank slates but often with basic misconceptions that create barriers to learning unless the teacher takes them into account. When the topic is intelligence testing, we must assume that one or more of the foregoing fallacies will impede understanding unless neutralized.

Second, fallacies must be confronted to be neutralized. Their impact can be greatly reduced if everyone contributes to the effort. Small preventive acts by many people can add up to a big difference. Preventive actions include taking care not to repeat or acquiesce unthinkingly to fallacious claims, communicating in a manner that clarifies oft-conflated distinctions, openly questioning the false premises and illogic of common fallacies and objecting to their persistent use, and calling major perpetrators to account.

Antitesting fallacies are rhetorical gambits that serve political ends. They hobble good science, impede the proper use of tests, and distort understandings of human diversity. They probably also interfere with democratic peoples negotiating more constructive accommodations of their differences.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, L. W. (Ed.). (1997). Educational testing and assessment: Lessons from the past, directions for the future [Special issue]. *International Journal of Educational Research, 27,* 357–399.

Andrews, L. B., & Nelkin, D. (1996, January 5). *The Bell Curve:* A statement [Letter to the editor]. *Science, 271,* 13–14.

Bartholomew, D. J. (2004). *Measuring intelligence: Facts and fallacies.* Cambridge, England: Cambridge University Press.

The "Bell Curve" agenda. (1994, October 24). *New York Times,* p. A16.

Binet, A., & Simon, T. (1916). *The development of intelligence in children* (E. S. Kit, Trans.). Baltimore: Williams & Wilkins.

Blakemore, B. (Executive Producer). (1994, November 22). The American agenda [Television news broadcast]. On *World news tonight with Peter Jennings* (Transcript #4232). New York: ABC News.

Blits, J. H., & Gottfredson, L. S. (1990a, Winter). Employment testing and job performance. *The Public Interest, 98,* 18–25.

Blits, J. H., & Gottfredson, L. S. (1990b). Equality or lasting inequality? *Society, 27*(3), 4–11.

Brody, N. (1992). *Intelligence* (2nd ed.). San Diego, CA: Academic Press.

Brody, N. (1996). Intelligence and public policy. *Psychology, Public Policy, and Law, 2,* 473–485.

Brody, N. (2003). Construct validation of the Sternberg Triarchic Abilities Test (STAT): Comment and reanalysis. *Intelligence, 31,* 319–329.

Camara, W. J., & Schmidt, A. E. (1999). *Group differences in standardized testing and social stratification.* New York: College Examination Board.

Campbell, J. P., & Knapp, D. J. (Eds.). (2001). *Exploring the limits of personnel selection and classification.* Mahwah, NJ: Erlbaum.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Carroll, J. B. (1995). Reflections on Stephen Jay Gould's *The Mismeasure of Man* (1981): A retrospective review. *Intelligence, 21,* 121–134.

Carroll, J. B. (1997). Psychometrics, intelligence, and public perception. *Intelligence, 24,* 25–52.

Cascio, W. F., Outtz, J., Zedeck, S., & Goldstein, I. L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance, 4,* 233–264.

Ceci, S. J. (1996a). General intelligence and life success: An introduction to the special theme. *Psychology, Public Policy, and Law, 2,* 403–417.

Ceci, S. J. (Ed.). (1996b). Special theme: IQ in society. *Psychology, Public Policy, and Law, 2,* 403–645.

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, 149–160.

Colangelo, N., & Davis, G. A. (Eds.). (2003). *Handbook of gifted education* (3rd ed.). Boston: Allyn & Bacon.

Dawkins, R. (1999). *The extended phenotype: The long reach of the gene*. Oxford, England: Oxford University Press.

Deary, I. J. (2000). *Looking down on human intelligence: From psychometrics to the brain*. Oxford, England: Oxford University Press.

Deary, I. J., Whiteman, M. C., Starr, J. M., Whalley, L. J., & Fox, H. C. (2004). The impact of childhood intelligence on later life: Following up the Scottish Mental Surveys of 1932 and 1947. *Journal of Personality and Social Psychology, 86*, 130–147.

Detterman, D. K. (Ed.). (1994). *Current topics in human intelligence: Vol. 4. Theories of intelligence*. Norwood, NJ: Ablex.

Diamond, J. (1999). *Guns, germs, and steel: The fate of human societies*. New York: Norton.

Dionne, E. J., Jr. (1994, October 18). "Race and IQ: Stale notions." *Washington Post*, p. A7.

Duster, T. (1995). What's new in the IQ debate. *The Black Scholar, 25*(1), 25–31.

FairTest (2007, August 20). *The ACT: Biased, inaccurate, and misused* [FairTest University testing fact sheet]. Retrieved March 17, 2008, from http://www.fairtest.org/act-biased-inaccurate-and-misused

Fischer, C. S., Hout, M., Jankowski, M. S., Lucas, S. R., Swidler, A., & Voss, K. (1996). *Inequality by design: Cracking the bell curve myth*. Princeton, NJ: Princeton University Press.

Flanagan, D. P., Genshaft, J. L., & Harrison, P. L. (Eds.). (1997). *Contemporary intellectual assessment: Theories, tests, and issues*. New York: Guilford Press.

Fletcher, R. (1991). *Science, ideology and the media: The Cyril Burt scandal*. New Brunswick, NJ: Transaction Press.

Flynn, J. R. (2007). *What is intelligence?: Beyond the Flynn effect*. New York: Cambridge University Press.

Frisby, C. L. (Ed.). (1999). Straight talk about cognitive assessment and diversity [Special issue]. *School Psychology Quarterly, 14*(3).

Frisby, C. L., & Reynolds, C. R. (Eds.). (2005). *Comprehensive handbook of multicultural school psychology*. New York: Wiley.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic Books.

Gardner, J. W. (1984). *Excellence: Can we be equal and excellent too?* (rev. ed.). New York: Norton.

Gordon, R. A. (1987). SES vs. IQ in the Race-IQ Delinquency Model. *International Journal of Sociology and Social Policy, 7*, 30–96.

Gottfredson, L. S. (Ed.). (1986). The *g* factor in employment [Special issue]. *Journal of Vocational Behavior, 29*(3).

Gottfredson, L. S. (1994). The science and politics of race-norming. *American Psychologist*, 49, 955–963

Gottfredson, L. S. (1996). Racially gerrymandering the content of police tests to satisfy the U.S. Justice Department: A case study. *Psychology, Public Policy, and Law, 2*, 418–446.

Gottfredson, L. S. (Ed.). (1997a). Intelligence and social policy [Special issue]. *Intelligence, 24*(1).

Gottfredson, L. S. (1997b). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*, 13–23.

Gottfredson, L. S. (2004). Intelligence: Is it the epidemiologists' elusive "fundamental cause" of social class inequalities in health? *Journal of Personality and Social Psychology, 86*, 174–199.

Gottfredson, L. S. (2007). Applying double standards to "divisive" ideas: Commentary on Hunt and Carlson (2007). *Perspectives on Psychological Science, 2*, 216–220.

Gould, S. J. (1981). *The mismeasure of man*. New York: Norton.

Gould, S. J. (1996). *The mismeasure of man* (Rev. ed.) New York: Norton.

Hart, M. H. (2007). *Understanding human history: An analysis including the effects of geography and differential evolution*. Augusta, GA: Washington Summit Publishers.

Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. Washington, DC: National Academy Press.

Hayden, E. C. (2007, October 17). So similar, yet so different. *Nature*, 762–763.

Helms, J. E. (2006). Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective. *American Psychologist, 61*, 845–859.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.

Holt, J. (1994, October 19). Anti-social science? *New York Times*, p. A23.

Horn, J. L. (1988). Thinking about human abilities. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 645–685). New York: Plenum Press.

Howe, M. J. A. (1997). *IQ in question: The truth about intelligence*. London: Sage.

Hunt, E. (1996). When should we shoot the messenger? Issues involving cognitive testing, public policy, and the law. *Psychology, Public Policy, and Law, 2*, 486–505.

Hunt, E., & Carlson, J. (2007). Considerations relating to the study of group differences in intelligence. *Perspectives on Psychological Science, 2*, 194–213.

Jaschik, S. (2007, March 2). Making holistic admissions work. *Inside Higher Ed*. Retrieved March 2, 2008, from http://www.insidehighered.com/news/2007/03/02/holistic

Jencks, C., & Phillips, M. (Eds.). (1998). *The Black–White test score gap*. Washington, DC: Brookings Institution Press.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Jensen, A. R. (1981). *Straight talk about mental testing*. New York: Free Press.

Jensen, A. R. (1985). The nature of the Black–White difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences*, 8(2), 193–263.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences*. New York: Elsevier.

Jensen, A. R., & Sinha, S. N. (1993). Physical correlates of human intelligence. In P. A. Vernon (Ed.), *Biological approaches to the study of human intelligence* (pp. 139–242). Norwood, NJ: Ablex.

Joynson, R. B. (1989). *The Burt affair*. London: Routledge.

Jung, R. E., & Haier R. J. (2007). The parieto-frontal integration theory (P-FIT) of intelligence: Converging neuroimaging evidence. *Behavior and Brain Sciences, 30*, 135–187.

Kirsch, I. S., & Mosenthal, P. B. (1990). Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly, 25*, 5–30.

Lewontin, R. C. (1976). Race and intelligence. In N. J. Block & G. Dworkin (Eds.), *The IQ controversy* (pp. 78–92). New York: Pantheon Books.

Lohman, D. F. (1997). Lessons from the history of intelligence testing. *International Journal of Educational Research, 27*, 359–377.

Loury, L. D. (1995, August). An exchange. IQ, race, and heredity: Charles Murray and his critics. *Commentary*, 18–19.

Lubinski, D. (Ed.). (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) "General Intelligence, Objectively Determined and Measured." *Journal of Personality and Social Psychology, 86*, 96–111.

Mainstream science on intelligence. (1994, December 13). *Wall Street Journal*, p. A18.

Marks, J. (1995). *Human biodiversity: Genes, race, and history*. New York: Aldine de Gruyter.

Michael, J. S. (1988). A new look at Morton's craniological research. *Current Anthropology, 29*, 349–354.

Miller, D. W. (2001, March 2). Scholars say high-stakes tests deserve a failing grade. Studies suggest students an educator are judged by faulty yardsticks. *The Chronicle of Higher Education*, p. A14.

Monastersky, R. (2008, February 18). Researchers gain understanding of how poverty alters the brain. *The Chronicle of Higher Education*. Available at http://chronicle.com/daily/2008/02/1705n.htm

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Neisser, U. (Ed.). (1998). *The rising curve: Long-term gains in IQ and related measures*. Washington, DC: American Psychological Association.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77–101.

Park, M. A. (2002). *Biological anthropology* (3rd ed.). Boston: McGraw-Hill.

Plomin, R., DeFries, J. C., McClearn, G. E., & McGuffin, P. (2001). *Behavioral genetics* (4th ed.). New York: Worth.

Plomin, R., & McClearn, G. E. (Eds.). (1993). *Nature, nurture, and psychology*. Washington, DC: American Psychological Association.

Plomin, R., & Petrill, S. A. (1997). Genetics and intelligence: What's new? *Intelligence, 24*, 53–77.

Race and intelligence: Politicizing the findings [Interview by F. Chideya with Bill Tucker, author of *The intelligence controversy*]. (2007, October 23). *News & Notes* [Radio program]. Audio retrieved June 27, 2008, from http://www.npr.org/templates/story/story.php?storyId=15560405

Rosenblum, V. G. (1996). On law's responsiveness to social scientists' findings: An intelligible nexus? *Psychology, Public Policy, and Law, 2*, 620–634.

Rowe, D. C. (1997). A place at the policy table? Behavior genetics and estimates of family environmental effects on IQ. *Intelligence, 24*, 53–77.

Rowe, D. C., Vesterdal, W. J., & Rodgers, J. L. (1998). Herrnstein's syllogism: Genetic and shared environmental influences on IQ, education, and income. *Intelligence, 26*, 405–423.

Rushton, J. P. (1997). Race, intelligence, and the brain: The errors and omissions of the "revised" edition of S. J. Gould's *The Mismeasure of Man* (1996) [Book review]. *Personality and Individual Differences, 23*, 169–180.

Rushton, J. P. (2002). New evidence on Sir Cyril Burt: His 1964 speech to the Association of Educational Psychologists. *Intelligence, 30*, 555–567.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56*, 302–318.

Samelson, F. (1982, February 5). Intelligence and some of its testers [Review of the book *The mismeasure of man*]. *Science, 215*, 656–657.

Samelson, F. (1992). Rescuing the reputation of Sir Cyril [Burt] [Review of the books *The Burt affair* and *Science, ideology, and the media: The Cyril Burt scandal*]. *Journal of the History of the Behavioral Sciences, 28*, 221–233.

Sarich, V., & Miele, F. (2004). *Race: The reality of human differences*. Boulder, CO: Westview.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed.) San Diego, CA: Sattler.

Scarr, S. (1997). Behavior-genetic and socialization theories of intelligence: Truce and reconciliation. In R. J. Sternberg & E. L. Grigorenko (Eds.), *Intelligence, heredity, and environment* (pp. 3–41). Cambridge, England: Cambridge University Press.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implication of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82,* 719–730.

Singham, M. (1995). Race and intelligence: What are the issues? *Phi Delta Kappan,* December, 271–278.

Smith, W. J., & Lusthaus, C. (1995). The nexus of equality and quality in education: A framework for debate. *Canadian Journal of Education, 20,* 378–391.

Snyderman, M., & Herrnstein, R. J. (1983). Intelligence tests and the immigration act of 1924. *American Psychologist, 38,* 986–995.

Snyderman, M., & Rothman, S. (1987). Survey of expert opinion on intelligence and aptitude testing. *American Psychologist, 42,* 137–144.

Snyderman, M., & Rothman, S. (1988). *The IQ controversy: The media and public policy.* New Brunswick, NJ: Transaction.

Society of Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* New York: McMillan.

Sternberg, R. J. (1997). *Successful intelligence: How practical and creative intelligence determine success in life.* New York: Plume.

Sternberg, R. J., & Grigorenko, E. L. (Eds.). (2001). *Environmental effects on cognitive abilities.* Mahwah, NJ: Erlbaum.

Sternberg, R. L., & Grigorenko, E. L. (Eds.). (2002). *The general intelligence factor: How general is it?* Mahwah, NJ: Erlbaum.

Sternberg, R. J., Wagner, R. K., Williams, W. M., & Horvath, J. A. (1995). Testing common sense. *American Psychologist, 50,* 912–927.

Terman, L. M. (1928). The influence of nature and nurture upon intelligence scores: An evaluation of the evidence in Part 1 of the 1928 Yearbook of the National Society for the Study of Education. *Journal of Educational Psychology, 19,* 362–373.

Vernon, P. A. (Ed.). (1993). *Biological approaches to the study of human intelligence.* Norwood, NJ: Ablex.

Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49,* 436–458.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale—Third edition (WAIS–III) and Wechsler Memory Scale—Third edition (WMS–III) technical manual.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (2002). *Wechsler Preschool and Primary Scale of Intelligence—Third edition (WPPSI–III) technical and interpretive manual.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children—Fourth edition (WISC–IV) technical and interpretive manual.* San Antonio, TX: Psychological Corporation.

White, S. H. (2000). Conceptual foundations of IQ testing. *Psychology, Public Policy, and Law, 6,* 33–43.

The White House. (2001). *Foreword by President George W. Bush* [No Child Left Behind Act of 2001]. Retrieved June 27, 2008, from http://www.whitehouse.gov/news/reports/no-child-left-behind.html

Wigdor, A. K., & Garner, W. R. (Eds.). (1982). *Ability testing: Uses, consequences, and controversies: Part I. Report of the committee. Part II. Documentation section.* Washington, DC: National Academy Press.

Wigdor, A. K., & Green, B. F. (Eds.). (1991). *Performance assessment for the workplace* (Vol. 1). Washington, DC: National Academy Press.

Williams, W. M. (1996). Consequences of how we define and assess intelligence. *Psychology, Public Policy, and Law, 2,* 506–535.

Williams, W. M. (Ed.). (2000). Special themes: Ranking ourselves: Perspectives on intelligence testing, affirmative action, and educational policy. *Psychology, Public Policy, and Law, 6,* 1–252.

Yerkes, R. M. (Ed.). (1921). *Psychological examining in the United States Army.* Washington, DC: National Academy of Sciences.