Logical Fallacies Used to "Discredit" Intelligence Testing

Linda S. Gottfredson

University of Delaware

> *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 also its most publicly reviled. Measurement technology is far more sophisticated than in decades past, but anti-testing 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 wrong-headed and its products misused, but that does not distinguish mental measurement from any other expert endeavor. The problem with intelligence testing is instead, one suspects, 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], and 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 of screening, selecting, and placing 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.

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 for 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* upon blacks and Hispanics. That is, they yield lower average scores for them than for Asians and whites. In employment settings, different 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 chapter 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 will be described). Other socially important sociopsychological measures are likewise *norm-referenced*, not *criterion-referenced*. 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 different members of a group fall above or below the group's average on some social indicator (academic achievement, health) or hierarchy (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 applicants to college 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 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. But when individuals differ substantially in merit, however it is defined, societies

cannot simultaneously and fully satisfy both 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. Gardner, 1984). This is the *democratic dilemma,* which is created by differences in human talent. In many applications, the democratic dilemma's chief source today is the wide dispersion in human intelligence, because higher intelligence is well documented as providing individuals with more practical advantages in modern life than any other single indicator, including social class background.

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

One effect of research with intelligence tests has been to make such choices and their societal consequences clearer and more public. There now exists a sizeable literature in personnel selection psychology, for example, that estimates the costs and benefits of sacrificing different levels of test validity to improve racial balance by different degrees when selecting workers for different kinds of jobs (e.g., Schmitt, Rogers, Chan, Sheppard, & Jennings, 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 practice (e.g., require racial preferences? or 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 non-use all provide decision-makers tools for tilting tradeoffs 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 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 mollify 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 the evidence for the validity of the tests and their results is ambiguous, unsettled, shoddy, or dishonest. Although mistaken, that 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.

Exhibit 1 and Table 1 summarize key aspects of the literature. Exhibit 1 reprints a statement by 52 experts that summarizes 25 of the most elementary and firmly-established conclusions about intelligence and intelligence testing. Received wisdom outside the field is often quite the opposite, in large part because of the fallacies I will describe. Table 1 illustrates how the scientific debates involving intelligence testing have advanced during the last half century. The list is hardly exhaustive and no doubt reflects the particular issues I have followed in my career, but it likewise 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 Exhibit 1 and Bartholomew, 2004; Brody, 1992; Carroll, 1993; Deary, 2000; Deary et al., 2004; Gottfredson, 1997, 2004; Jensen, 1980, 1998; Murphy & Davidshofer, 2005; Neisser et al., 1996; Plomin, DeFries, McClearn, & McGuffin, 2001; Schmidt & Hunter, 1998; Wigdor & Garner, 1982).

Exhibit 1 and Table 1 go about here

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? As this question illustrates, 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 remain open or had been answered more to the critics' liking. The clearest recent example is the 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[1] as long-discredited pseudoscience when, in fact, they represent some of the most elemental scientific conclusions about tests and intelligence (Carroll, 1997). Statements from the American Psychological Association (Neisser et al., 1994) and the previously mentioned group of experts (see Exhibit 1) who attempted to set the scientific record straight in both public ("Mainstream Science on Intelligence," *Wall Street Journal*, December 13, 1994, p. A18) and scientific venues (Gottfredson, 1997) 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, 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 do not. As noted above, 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 therefore include all highly *g*-loaded tests below when I refer to tests of intelligence.

Public opinion is always riddled with error, of course, no matter what the issue. But fallacies are not simply mistaken claims or intentional lies, which could be effectively answered with facts disproving to them. Instead, they tend to systematically corrupt public understanding. They not only present falsehoods as truths, but reason falsely about the facts, thus making those they persuade largely insensible to correction. 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 anti-testing 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 helps to distinguish the usual targets by grouping them into three arenas of research and debate: Can intelligence be measured, and how? What are the causes and consequences of human variation in intelligence? And, what are the social aims and impact of using intelligence tests—or *not* using them—as tools in making decisions about individuals and organizations? These are labeled in Figure 1, respectively, as the measurement model, the causal network, and the politics of test use. Key phenomena (really, fields of inquiry) within each arena are distinguished by numbered entries to more easily illustrate which fact or field each fallacy aims to discredit. The arrows (→) represent the relations among the phenomena at issue, such as the impact of genetic differences on brain structure (Entry 1 → Entry 4), or the temporal ordering of advances in measurement (Entries 8 → 9 → 10 → 11).

Figure 1 goes about here

I. Measurement

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. A 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 were following different internal compasses as they engaged the world around them. Because they are inferred, traits are called *theoretical constructs*. They therefore represent causal hypotheses about why 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 must be detected via their effects on something that is observable. Intelligence tests consist of a set of tasks that reliably instigates performances requiring intelligence 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, interative process of evaluating and revising tests.

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, the so too have fallacies about it multiplied and mutated. Figure 1 delineates the broad stages (Entries 8-11) in this coevolution of intelligence measurement and the fallacies about 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.

A. Test design.

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, 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) to reveal and quantify differences among school children in that hypothetical trait (Entry 5), as Binet's observations had led him to conceive it. The French Ministry of Education had asked Binet to develop an objective way of identifying 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, pp. 42-43), 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.

This conception provided a good starting point for designing tasks that might effectively activate intelligence and cause it to leave its fingerprints in observable behavior. Binet and Simon's 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 "points to nose, eyes, and mouth" (age 3), "counts thirteen pennies" (age 6), "notes omissions from pictures of familiar objects" (age 8), "arranges five blocks in order of weight" (age 10), and "discovers 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 to which children are routinely exposed outside of schools and 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 is 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., WISC 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): 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 .8 among themselves (which is not far below the maximum possible in view of their reliabilities of over .9), so they are capturing the same phenomenon. Mere similarity of results among IQ tests is necessary but not sufficient to confirm that they tests measure the intended construct.

Today, item content, test format, and administration procedure (Entry 8) 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 they all 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 (WPPSI), school-age children (WISC), and adults (WAIS), are administered orally to test takers one-on-one, item by item for an hour or more, by highly trained professionals who follow written scripts governing what they must and must not say to the individual in order 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, that is, one which consistently measures the intended construct and nothing else. Such strategies cannot guarantee this happy result, of course. That 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 whose superficial appearances make them highly vulnerable to fallacious reasoning of the following sorts.

Test-design fallacy # 1: Yardstick is construct. Portraying the superficial appearance of a test (Entry 8) as if it mimicked the inner essence of the phenomenon it measures (Entry 5).

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. And yet, some critiques of intelligence testing rest precisely on such reasoning. For example, one often hears that intelligence tests cannot possibly measure a general intelligence because they are composed of items with narrow, esoteric, or academic content of little practical value, their tasks are more structured than real-life problems, they are unrealistic because they always have right and wrong answers, and they do not give credit for experience.

Instead, the argument continues, IQ tests measure only what their superficialities suggest, for example, just an aptness with paper-and-pencil tasks, a narrow academic ability, familiarity with the tester's culture, facility only with well-defined tasks with unambiguous answers, and the like. Not only are these inferences unwarranted, but their premises are often wrong. In actuality, most items on individually-administered batteries require neither paper nor pencil, 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.). The superficial mechanics of IQ testing tell us nothing about the essence of the construct that IQ tests capture. Nor does the manifest content of their items. [2]

Figuring out what construct(s) a particular test actually measures requires extensive *validation* research, which involves collecting and analyzing test results in many different circumstances and populations. As described later, such research shows that ostensibly different tests can be used to measure the same latent ability. The Yardstick-Is-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. Portraying intelligence as if it were just an aggregation of separate specific abilities, not a phenomenon in itself (Entry 5), because IQ tests calculate IQs by adding up scores on different subtests (Entry 9).

The overall IQ is typically calculated by, in essence, adding up a person's scores on the various subtests in a battery. This manner of calculating scores from IQ tests (the *measure*) is often mistaken as mirroring how general intelligence itself (the hypothetical entity or *construct*) is constituted. Namely, 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 a conglomeration or aggregate of many separate elements like marbles in a bag. While this may have been a viable hypothesis in Binet's time, the conglomeration view has now been decisively disproved. Appealing to the mechanics by which a scale succeeds in quantifying something—heat, speed, height—says nothing about the integrity or organizing principles of the phenomenon being measured. As discussed further below, *g* (Entry 10) is not the sum of separate cognitive abilities, but is the common core of them all. In this sense, general intelligence is *psychometrically unitary.*[3]

B. Test scoring.

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 correctly, or *raw* score, has no intrinsic meaning. Nor does percentage correct, because the denominator (total number of test items) has no substantive meaning either. 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 to 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 attempted to provide interpretable intelligence test results by assigning a *mental age* (MA) to each item on their test (the age at which the average child answers it correctly). Because mental capacity increases over childhood, a higher score can be interpreted as a sign of more advanced cognitive development. To illustrate, if 8-year-olds answer 20 items correctly, on the average, then a raw score of 20 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.[4] Thus, if John scores at the average for children aged 10 years, 6 months, he has a mental age of 10.5. How we interpret his mental age depends, of course, on how old he 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).

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 (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 became 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 the 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 zero 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 will score within two standard deviations of the mean, that is, between IQs 68-132 (when SD is set to 16) or between IQs 70-130 (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 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, first note that temperature is criterion referenced. 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 outcome within an age cohort, 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 an SAT-Verbal score of 600 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?[5]

Now, IQ tests are not intended to isolate and measure highly specific skills and knowledge. That 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 leaves IQ tests as black boxes into which people can project all sorts of unwarranted hopes and fears. Psychometricians speaking in statistical tongues may be perceived as psycho-magicians practicing dark arts.

But 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 chunks, 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 4 years ago, she grew taller at a rate of 1 inch per year, or she runs 1 mile/hour faster than her sister. And we can 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 ft for all ages and variability in height to be the same for four-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. But better this "deviation height" scoring than assigning ages to height scores and dividing that "height age" by chronological age to get an HQ (HA/CA), which 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 we can learn from them. It also nourishes certain fallacies about intelligence testing because, without the absolute results to contradict them, critics can falsely represent IQ scores (relative standing in ability) as if they gauged absolute levels of ability in order to ridicule and discredit them. The following 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 tradeoffs that democratic societies would rather not ponder. All three operate by confusing different sorts of variation: absolute growth vs. growth relative to age peers, the required components for development (hence, no variation at all) vs. individual differences in development, and differences within a species vs. differences between species.

Test-scoring fallacy #1: Growth proves malleability. Using evidence of developmental change as if it were proof that intelligence level is changeable (malleable).

This common fallacy creates the illusion of malleability despite evidence to the contrary. Figure 2 distinguishes the two phenomena that the fallacy confuses: absolute growth vs. relative growth. The three curves represent the typical course of cognitive development for individuals at three levels of relative ability: IQs 70, 100, and 130. All three sets of individuals develop along similar trajectories, their mental capabilities rising in childhood, leveling off in adulthood, and then falling somewhat in old age. The trajectories for brighter individuals are steeper and therefore level off at a higher point. This typical pattern has been ascertained from various specialized tests whose results were not age-normed. As noted earlier, no current tests can gauge absolute level of intelligence ("raw mental power" in Figure 2), so we cannot be sure what the shape of the curves is, but available data and common observation leave no doubt that they tend to rise steeply in childhood and fall late in life.

Figure 2 goes about here

IQ tests cannot chronicle amount of growth and decline over a lifetime, because they are not ratio measures. They compare individuals only to 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.

IQ level is not made malleable by any means yet devised, but many a critic has sought to nullify this fact by pointing to the obvious but irrelevant fact that individuals grow and learn. The Growth-Proves-Malleability Fallacy succeeds by using the word "change" for two entirely different phenomena as if they were one and the same. It points to developmental "change" within individuals to suggest, wrongly, that the differences between individuals may be readily "changed." Asserting that IQ is stable (unchanging) despite this obvious growth (change) therefore makes one appear foolish or doggedly ideological. Ratio measurement would make the fallacy as transparent for intelligence as it would be for height: children grow, so their heights can all be made the same. The fallacy's political implication is that the democratic dilemma need not exist because cognitive inequality need not exist.

Test-scoring fallacy #2: Interactionism disproves heritability. Portraying the gene-environment partnership in creating a phenotype as if conjoint action within the

individual precluded teasing apart the roots of phenotypic differences between individuals.

While the Growth Fallacy seems to discredit a phenotypic finding (stability of observed IQ), the Interactionism Fallacy targets a genetic finding (substantial heritability of IQ differences). It states an irrelevant truth to reach an irrelevant conclusion in order to peremptorily dismiss all estimates of heritability. The irrelevant truth: An organism's development requires genes and environments to act in concert. The two forces are inextricable, mutually dependent, constantly interacting. Development is their mutual product, like the dance of two partners. The irrelevant conclusion: It is therefore impossible to apportion credit for the product to each partner separately, say, 40% of the dance to the man and 60% to the woman. The inappropriate generalization: Behavior geneticists cannot possibly do what they claim to, namely, decompose phenotypic variation within a particular population into its genetic and nongenetic components. The fallacy creates this 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).

The field of behavior genetics seeks to explain, not the human theme, but variations on it. It does so by measuring phenotypes for pairs of individuals that differ systematically in genetic and environmental relatedness. Such data allow decomposition of phenotypic variation in behavior into its environmental (Entry 2 in Figure 1) and genetic (Entry 1) sources. The field has actually gone far beyond estimating the heritabilities of traits. For instance, it can determine to what extent the *co-variation* between two outcomes (e.g., a correlation between intelligence and occupational level) represents a genetic correlation between them (e.g., intelligence and occupation share some genetic roots).

Critics reinforce the Interactionism Fallacy by caricaturing the unwanted conclusions about heritability. Researchers refer to percentages of phenotypic variation in a population that can be traced to genotypic variation in the population. But critics transmogrify this into the absurd 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 cellular conditions,[6] 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 human genes (*alleles*) cause individuals to respond differently in the *same* circumstances.

Thus the Interactionism Fallacy's one-two punch: Caricature the unwelcome scientific conclusion, X, into absurdity, and then say something incontestable about something else, Y, in words that suggest X, as if the irrelevant truth about Y scientifically disproved X. It provides a scientific-sounding excuse to denigrate all evidence for a genetic influence (Entry 1 in Figure 1) on intelligence (Entry 5) as patently absurd.

> *Test-scoring fallacy #3: 99.9% similarity trumps differences. Portraying the study of human genetic variation as irrelevant or wrong-headed because humans are 99.9% alike genetically, on average.*

Of recent vintage, the 99.9% Fallacy impugns even investigating human genetic variation by implying, falsely, that a 0.1% difference in genetic profiles (3 million base pairs) is trivial. (For comparison, the human and chimpanzee genomes differ by about 1.3%.) The fallacy works by looking at human variation as if from a great distance, against the full length of evolutionary time. 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 99.9% genetic similarity makes us human, *Homo sapiens sapiens,* but the other 0.1% makes us individuals.

The identical parts of the genome are called the *non-segregating* genes, which are said to be 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 which 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 segregating genes that scientists search for the genetic roots of those phenotypic differences. The 99.9% Fallacy would put this search off-limits.

C. Test-validation.

Validating a test refers to determining which sorts of inferences may properly be drawn from the test's scores, most commonly whether it measures 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. And 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). The tests also tend to predict these outcomes better than does any other single predictor, including family background. This evidence makes it plausible that IQ tests measure differences in a general intelligence, but it is not sufficient to prove 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: Is there one intelligence, or many? Do IQ tests measure the same intelligence or different intelligences in different sex, race, ethnic, and age groups? Do they measure intelligence at all, or just familiarity with the culture or social privilege? Advances in measurement have provided new ways to adjudicate such claims. Entries 10 and 11 in Figure 1 represent two advances in identifying, isolating, and contrasting the constructs that cognitive tests may be measuring: factor analysis and latent trait modeling. Both provide tools for scrutinizing tests and test items in action (Entry 9) 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 re-examined. Successive rounds

of such psychometric scrutiny reveal a lot not only about tests but also about the phenomenon they poke and prod into expressing itself.

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 abilit(ies). Intelligence researchers developed the method of factor analysis to extract those common factors from any large, diverse set of mental tests administered to representative samples of individuals (Entry 10). With this tool, the researchers can ask: 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 only a single general ability rather than many independent co-equal ones, 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, called *g,* consistently emerges at the most general level (Carroll's Stratum III),[7] and (c) the group factors in Stratum II, such as verbal or spatial ability, correlate moderately highly

with each other because all reflect mostly *g*—explaining why Carroll refers to them as different "flavors" of the same *g*.[8] There are many different cognitive abilities but all turn out to be suffused with *g*. The most important distinction among them, overall, is how broadly applicable they are, from all-purpose (*g*) to narrow and specific (e.g., associative memory, reading decoding, pitch discrimination).

Although the *g* factor is highly correlated with the IQ (usually .8 or more), the distinction between *g* (Entry 10) and IQ (Entry 9) cannot be overstated. The IQ is nothing but a test score, albeit one with social portent and, for some purposes, considerable practical value. *g,* however, 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), through the brain (Entry 4), and into the currents of social life (Entries 6 and 7). 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, so it is not a statistical chimera. Its nature is not constructed or corralled by how we choose 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, and problem solving—the construct—that researchers and lay persons alike usually associate with the term intelligence. 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 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 phenomenon being measured, *g*, from the devices used to measure it. Its discovery shows that the underlying phenomenon that IQ tests measure (Entry 10) has nothing to do with the manifest content or format of the test (Entry 8): it is not restricted to paper-and-pencil tests, to timed tests, ones with numbers or words, academic content, or whatever. The active ingredient in intelligence tests is something deeper and less obvious—namely, the cognitive complexity of the various tasks to be performed. 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.

This separation between measure and phenomenon also affords the possibility of examining how well different tests and tasks measure *g* or, stated another way, how heavily each draws upon or taxes *g* (how *g loaded* each is) and thereby advantages individuals of higher *g*. Just as we can characterize individuals by *g* level, we can now characterize tests and tasks by their *g* loading and learn 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. Perhaps 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 *g* and something in addition to *g.* Using hierarchical factor analysis, psychometricians can strip the

lower-order factors and tests of their *g* components in order to reveal what each measures uniquely and independently of all others. 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 for all the other factors combined. Hierarchical factor analysis also reveals 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 non-cognitive abilities such as pitch discrimination) actually measure mostly *g*, not the specialized ability that their names suggest.

All the factor analyses mentioned so far employed *exploratory* factor analysis (EFA), which extracts a parsimonious set of factors to explain the commonalities running through tests and causing them to intercorrelate. It posits no constructs but waits to see which dimensions emerge from the process (Entry 10). It is a data reduction technique, which means that it provides fewer factors than tests in order 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, though 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* (CFA). Exploratory factor analysis generates hypotheses about latent traits, but CFA tests them (Entry 11). It has become the method of choice for ascertaining which constructs a particular IQ test battery taps, that is, its construct validity. Multi-group confirmatory factor analysis (MGCFA) has become the method of choice for ascertaining whether a test measures exactly the same constructs in different ages, races, and sexes. It thus constitutes another, more

exacting means of vetting tests for cultural bias.[9] The finding of *construct invariance* (lack of bias) has been the rule in studies so far, which also strengthens the case for a culture-independent intelligence continuum. CFA and MGCFA belong to a larger family of methods for modeling latent traits, often called *structural equation modeling* (SEM). SEM is used to test causal models, for instance, of *g*'s origins and effects.

The following two fallacies would have us believe, however, that nothing important has been learned about intelligence tests since Binet's time in order to sweep aside a century of construct validation. Both ignore the discovery of *g* and promote outdated musings in order to dispute the possibility that IQ tests could possibly measure such a general intelligence.

Test-validation fallacy #1: Contending definitions trump evidence. Portraying lack of consensus in verbal definitions of intelligence as if that 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 definition of it. By this reasoning, one could just as easily dispute that gravity or health 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 the suspected phenomenon or the possibility of measuring it. What matters most is not the diversity of proposed definitions, but construct validation or "dialogue with the data" (Bartholomew, 2004, p. 52). Insisting on a consensus definition is an excuse to ignore what has been learned already, especially about *g*, and leaves the field at the mercy of the most capricious and undisciplined definition mongers

Test-validation fallacy #2: Non-identical results proves tests invalid. Portraying differences in results from two intelligence tests as evidence against their both measuring intelligence.

This fallacy rests on the false premise that cognitive tests either measure intelligence or they don't. It thereby encourages the mistaken belief that different tests should produce identical results if they are all good measures of intelligence. When a person's results differ across tests, as they usually do to some extent, test critics invite us to doubt the validity of all the tests. Or, when racial gaps differ in size across tests, critics invite us to infer pervasive test bias (the larger gaps supposedly signaling more bias). These hypotheses may have been plausible decades back, but no longer. The differences in results to which critics point not only fail to discredit the construct validity of cognitive tests, but they confirm it. Why? Because the differences conform to theoretical expectation.

The hierarchical structure of mental abilities discovered via factor analysis (e.g., 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 and the size of group differences that will be found in new samples. First, tests will intercorrelate more highly when they are more *g* loaded; these will also be the more general tests (higher in Carroll's hierarchy; see Entry 10 in Figure 1). Second, controlling for *g* loading, tests will correlate more highly when their non-*g* components are more similar, which is most likely for broad tests at the Stratum II level. Third, groups of test takers who differ in average *g* will have proportionately larger test score gaps on progressively more *g*-loaded tests, and, fourth, groups who differ in their profile of Stratum II abilities (controlling for *g*) will outscore other groups (of equal average *g*) on tests in their strong suits.

To illustrate, the WAIS Vocabulary and Block Design subtests are far more *g* loaded that the Digit Span subtest. The former two thus correlate more highly with each other than with Digit Span. Blacks and whites differ in average *g*, so black-white differences are larger on the former two than on the third. Regarding profile differences, equally *g*-loaded tests often differ in their *specificities*, or non-*g* components of variance (see Entry 10 in Figure 1). Tests are more highly correlated when they tap the same broad non-*g* components (controlling for *g* loading). For example, Stratum II tests with a verbal component correlate more highly with each other than with tests having a strong spatial or memory component. Although men and women score about the same in *g* level, on average, they have different profiles of abilities at the Stratum II level. Females score somewhat higher on tests with a strong verbal component and males on tests with a strong spatial component. Today, we must question the validity of tests whose results do not differ in expected ways.

II. Causal Network Fallacies

Entries 1-7 in Figure 1 represent the core concepts required in any explanation of the causes of intelligence differences in a population (*vertical* processes, Entries 1-5; Jensen, 1998) and the effects they produce on it collectively and its members individually (*horizontal* processes, Entries 5-7). This schema is obviously a highly simplified rendition of the empirical literature (for example, 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 four causal-network fallacies.

Causal fallacy #1: Phenotype is genotype fallacy. Portraying phenotypic differences in intelligence (Entry 5) as if they were necessarily genotypic (Entry 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 similarity of their environments. Non-experimental 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 hair-splitting, 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.

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-Is-Genotype Fallacy thus exposes scholars to false allegations that they are actually asserting genetic differences whenever they fail to pointedly repudiate them. For example, critics often insinuate that scientists who report racial gaps in measured intelligence (Entry 5) are thereby asserting "innate" (genetic) differences (Entry 1) between the races. Some scholars try to preempt such attributions by taking pains to point out 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. They align themselves with *nongenetic* explanations by routinely blaming lower tested abilities and achievements on social disadvantages such as poverty and poor schooling, even when facts say otherwise (e.g., shared family effects on IQ and achievement [among whites] mostly fade away by adolescence, and there are sizeable *genetic* correlations among IQ, education, and social class in adulthood; Rowe, Vesterdal, & Rodgers, 1998).

Causal fallacy #2: Biological is genetic fallacy. Portraying biological differences (brain phenotypes, Entry 4) as if they were necessarily genetic (Entry 1).

This is a corollary of the Phenotype-Is-Genotype Fallacy, because an organism's form and physiology are part of its total phenotype. Like height and weight, many aspects of brain physiology (Entry 4) are under considerable genetic control (Entry 1), but nongenetic differences, say, in nutrition or disease (Entry 2) can also produce variation in physical traits. The finding that IQ scores correlate with various aspects of brain physiology—including total brain volume, volume of frontal lobes, neural conduction speed, complexity of brain waves, and

rate of glucose metabolism (negatively)—increases the plausibility that individual differences in general intelligence are substantially genetic in origin. Research in behavior genetics does, in fact, confirm a large genetic contribution to IQ differences, brain biology, and correlations between the two. The genetic correlations between IQ and brain suggest potential mechanisms by which genes could influence speed and accuracy of cognitive processing, yielding a higher intelligence. But 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 evidence exists only for whites. Extant evidence confirms mean racial differences in phenotypic intelligence and a few brain attributes, such as head size, but no scientific discipline has been willing since the 1970s to conduct genetic or brain research on nonwhites that could be tied to intelligence. The evidence for genetic influence on differences within the white population enhances the plausibility of a part-genetic 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 all similar-sounding ones, no matter how well validated scientifically the latter may be and even when they have nothing to do with race (e.g., the conclusion that IQ differences among whites are substantially heritable).

Causal fallacy #3: Social effect proves social cause. Portraying socioeconomic consequences of intelligence differences (Entry 6) as if they were evidence for socioeconomic causes of intelligence differences (Entry 3).

When distinct phenomena are called by the same name, they naturally risk being confused. Such is the case for *socioeconomic status (SES*), which figures prominently in debates over intelligence. Scholars often use the term to refer to both the possible causes and the possible consequences of IQ variation: respectively, (a) individuals' rearing circumstances (e.g., parents' income), which might influence their cognitive development in childhood, and (b) individuals' own achievements and circumstances in adulthood (e.g., occupational status), which greater cognitive aptness might have facilitated.

This distinction between possible social origins (Entry 3) and social consequences (Entry 6) of intelligence differences is obvious once pointed out, but it becomes lost in common discourse when socioeconomic variables are not explicitly qualified as *childhood* or *adult*. Conflating the two allows critics to use the robust correlations between intelligence level and socioeconomic outcomes (Entry 5 → Entry 6) as if they constituted evidence for a socioeconomic cause of intelligence differences (Entry 3 → Entry 5). This fallacy is sometimes used to press the false claim that intelligence level could be changed if only social conditions were changed. It is also used by some to press a very different false claim, namely, that IQ tests do not measure intellectual competence at all, but only social class privilege masquerading as merit in order to legitimize social class advantages (see Bowles and Gintis, 1972/1973, for the iconic statement). The first false claim evades the democratic dilemma by postulating high malleability of intelligence, and the other by denying that the trait truly exists. The first attributes

social inequality to the advantages of social privilege, and the second to oppression by the privileged.

Causal fallacy #4: Environment is nongenetic. Portraying environments (Entry 3) as if they were necessarily nongenetic (Entry 2).

This fallacy is the environmentalist counterpart to the hereditarian Biological-Is-Genetic Fallacy. It helps to bolster the Social-Cause Fallacy by excluding the possibility that an environmental correlation can reflect a genetic influence. Environments are physically external to individuals but, contrary to common belief, that does not make them nongenetic. 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). Both childhood and adult environments (Entries 3 and 6) are therefore influenced by the genetic proclivities of self and genetic kin. People's personal environments are their *extended phenotypes.*

Near-universal deference in the social sciences to the Environment-Is-Nongenetic Fallacy has fostered mostly causally uninterpretable research (see Scarr, 1997, on Socialization Theory, and Rowe, 1997, on Family Effects Theory and Passive Learning Theory). 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—or, alternatively, that tests must be biased when certain social groups score lower than others, on average. In falsely portraying external environments as strictly nongenetic, critics

commandeer all IQ-environment correlations as evidence for pervasive and powerful nongenetic causation or test bias.

When combined, the foregoing causal-network fallacies can produce more convoluted ones. To take one example, protagonists in *The Bell Curve* debate often conjoined the Phenotype-Is-Genotype Fallacy with the Environment-Is-Nongenetic Fallacy. All sides used IQ as a stand-in for genetic influence and social class as a stand-in for nongenetic influence in order to debate whether genes or environments create more social inequality.

III. Politics of Test Use

The previous sections on the measurement and correlates of cognitive ability were 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.

Test utility.

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? But validity is not sufficient reason for testing. The utility of tests in applied settings depends on practical considerations, too, including feasibility and cost of administration, difficulties in maintaining test security and operational validity, vulnerability to litigation, 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 in practice 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) can influence the tested individuals' life chances (Entry 6). This is why good practice dictates that test scores (or any other 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). For example, personnel selection tests can improve workforce productivity and change who has access to the best jobs. Like other social practices, testing (or not testing) tends to serve some social interests and goals over others. That 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), 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).

The measurement and causal-network fallacies 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*. But they otherwise observe the rules of science: ideas must compete, and evidence matters.

The following test-utility fallacies violate these rules in the guise of honoring them. They begin by ignoring the rule for adjudicating competing scientific claims: the *preponderance of*

evidence. Which claim best accounts for the totality of relevant evidence to date? They then demand that certain tests and test results meet especially rigorous scientific standards before their use can be condoned. The double standards for competing ideas are triggered by insinuating that one competitor poses special risks to the body politic. In other words, the test-utility fallacies invoke a political criterion for test utility (alleged social risk) to justify demanding that particular tests or ideas be presumed scientifically inferior to all competitors until they meet insurmountable quality standards.

Test-utility fallacy #1: Imperfection Standard. Maintaining that g*-loaded tests should not be used for making "high stakes" decisions until they are error-free.*

The Imperfection fallacy labels highly *g*-loaded tests as "flawed" because they are not error-free (reliability <1.0, or validity <1.0). Nothing is 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*). The concern is usually with minority false negatives, in particular. The insinuation is that valid, unbiased tests are biased, which allows opponents to call for suspending their use until they are cleansed of such "flaws."

The demand for technical improvement is clearly pretextual. 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. That is

precisely why the fallacy must use perfection to overrule the good, especially when the good gets better. This is not to say that all kinds of error are equal, as illustrated by the tradeoffs 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). But balancing different kinds of error is a political, monetary, or ethical decision, not a technical one. The Imperfection Standard provides a pretext for imposing a political choice of social goods in the guise of advocating technical improvement.

Test critics sometimes use the Imperfection Standard to justify *increasing* 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 reducing a test's reliability and validity. In like manner, the National Research Council (NRC) of the National Academy of Sciences (NAS) cited imperfect predictive validity to justify its 1989 recommendation that valid, unbiased employment tests be race-normed (Hartigan & Wigdor, 1989). Race-norming reduces disparate impact by introducing systematic error designed to favor some races and disfavor others.

Test-utility fallacy #2: Dangerous ideas standard. Maintaining that "destructive" or "divisive" scientific conclusions should not be entertained until proved beyond all conceivable doubt.

This fallacy sets a similarly selective and equally insurmountable evidentiary standard as does the Imperfection Standard, but for unwelcome scientific conclusions. Opponents insinuate

that an idea is fraught with danger in order to press their case for one-sided scientific rigor. The putative danger is never explained, but just connoted by references to physical harm (dangerous sports, risky human experimentation, genocide, etc.). Labeling a scientific conclusion dangerous allows any fear, any manufactured doubt, to trump the preponderance of the evidence for it, no matter how lop-sided the evidence may be. The premise seems to be that disturbing truths do no good and comforting lies do no harm.

The Dangerous Ideas Standard has appeared in different guises over the years. When rules governing research with human subjects were first formulated in the 1970s, there was an effort to bar research whose questions or answers 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 Danger Standard 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).

Experts themselves sometimes seem to accept these test-utility fallacies. They seem at once seduced by the appeal to scientific rigor and intimidated by the premise that their ideas might do social harm. So, 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, which 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 testing 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 respond to mounting scientific support for its construct validity, predictive value, and lack of bias with yet more strident critiques of the tests, tests results, and persons giving them credence.

The fourteen 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 make the true seem false and to protect the false from refutation. That 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. But 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 anti-testing fallacies are avidly pressed upon the populace. 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 we intervene.

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 make a big difference. Preventive actions include taking care not to unthinkingly repeat or acquiesce to fallacious claims, communicating in a manner that clarifies oft-conflated distinctions, openly questioning the false premises and illogic of common fallacies, objecting to their persistent use, and calling major perpetrators to account.

Anti-testing 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.

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

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

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.

Bowles, S. & Gintis, H. (1972/1973). IQ in the U.S. class structure. *Social Policy, 3(4 & 5)*, 65–96.

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

Carroll, J. B. (1993). *Human cognitive abilities*. Cambridge University Press.

Carroll, J. B. (1997). Psychometrics, intelligence, and public perception. *Intelligence, 24(1)*, 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.

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

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: W. W. Norton.

Gottfredson, L. S. (1994). The science and politics of race-norming. *American Psychologist, 49*(11), 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*(3/4), 418-446.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24(1),* 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. *Perspectives on Psychological Science, 2*(2).

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

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.

Herrnstein, R. J. & Murray, C. (1994). *The bell curve: Intelligence and class structure in*

American life. New York: Free Press.

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

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.

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

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. E., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996, February). Intelligence: Knowns and unknowns. *American Psychologist, 51(2)*, 77–101.

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

Rowe, D. C. (1997). A place at the policy table? Behavior genetics and estimates of family environmental effects on IQ. *Intelligence, 24(1)*, 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.

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

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: 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.

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

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

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.

Exhibit 1: Mainstream Science on Intelligence (Reprinted with permission from the *Wall Street Journal*, December 13, 1996, p. A18)

Since the publication of "The Bell Curve," many commentators have offered opinions about human intelligence that misstate current scientific evidence. Some conclusions dismissed in the media as discredited are actually firmly supported.

This statement outlines conclusions regarded as mainstream among researchers on intelligence, in particular, on the nature, origins, and practical consequences of individual and group differences in intelligence. Its aim is to promote more reasoned discussion of the vexing phenomenon that the research has revealed in recent decades. The following conclusions are fully described in the major textbooks, professional journals and encyclopedias in intelligence.

The Meaning and Measurement of Intelligence

1. Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings—"catching on," "making sense" of things, or "figuring out" what to do.
2. Intelligence, so defined, can be measured, and intelligence tests measure it well. They are among the most accurate (in technical terms, reliable and valid) of all psychological tests and assessments. They do not measure creativity, character, personality, or other important differences among individuals, nor are they intended to.
3. While there are different types of intelligence tests, they all measure the same intelligence. Some use words or numbers and require specific cultural knowledge (like vocabulary). Other do not, and instead use shapes or designs and require knowledge of only simple, universal concepts (many/few, open/closed, up/down).
4. The spread of people along the IQ continuum, from low to high, can be represented well by the bell curve (in statistical jargon, the "normal curve"). Most people cluster around the average (IQ 100). Few are either very bright or very dull: About 3% of Americans score above IQ 130 (often considered the threshold for "giftedness"), with about the same percentage below IQ 70 (IQ 70-75 often being considered the threshold for mental retardation).
5. Intelligence tests are not culturally biased against American blacks or other native-born, English-speaking peoples in the U.S. Rather, IQ scores predict equally accurately for all such Americans, regardless of race and social class. Individuals who do not understand English well can be given either a nonverbal test or one in their native language.
6. The brain processes underlying intelligence are still little understood. Current research looks, for example, at speed of neural transmission, glucose (energy) uptake, and electrical activity of the brain.

Group Differences

7. Members of all racial-ethnic groups can be found at every IQ level. The bell curves of different groups overlap considerably, but groups often differ in where their members tend to cluster along the IQ line. The bell curves for some groups (Jews and East Asians) are centered somewhat higher than for whites in general. Other groups (blacks and Hispanics) are centered somewhat lower than non-Hispanic whites.

8. The bell curve for whites is centered roughly around IQ 100; the bell curve for American blacks roughly around 85; and those for different subgroups of Hispanics roughly midway between those for whites and blacks. The evidence is less definitive for exactly where above IQ 100 the bell curves for Jews and Asians are centered.

Practical Importance

9. IQ is strongly related, probably more so than any other single measurable human trait, to many important educational, occupational, economic, and social outcomes. Its relation to the welfare and performance of individuals is very strong in some arenas in life (education, military training), moderate but robust in others (social competence), and modest but consistent in others (law-abidingness). Whatever IQ tests measure, it is of great practical and social importance.

10. A high IQ is an advantage in life because virtually all activities require some reasoning and decision-making. Conversely, a low IQ is often a disadvantage, especially in disorganized environments. Of course, a high IQ no more guarantees success than a low IQ guarantees failure in life. There are many exceptions, but the odds for success in our society greatly favor individuals with higher IQs.

11. The practical advantages of having a higher IQ increase as life settings become more complex (novel, ambiguous, changing, unpredictable, or multifaceted). For example, a high IQ is generally necessary to perform well in highly complex or fluid jobs (the professions, management); it is a considerable advantage in moderately complex jobs (crafts, clerical and police work); but it provides less advantage in settings that require only routine decision making or simple problem solving (unskilled work).

12. Differences in intelligence certainly are not the only factor affecting performance in education, training, and highly complex jobs (no one claims they are), but intelligence is often the most important. When individuals have already been selected for high (or low) intelligence and so do not differ as much in IQ, as in graduate school (or special education), other influences on performance loom larger in comparison.

13. Certain personality traits, special talents, aptitudes, physical capabilities, experience, and the like are important (sometimes essential) for successful performance in many jobs, but they have narrower (or unknown) applicability or "transferability" across tasks and settings compared with general intelligence. Some scholars choose to refer to these other human traits as other "intelligences."

Source and Stability of Within-Group Differences

14. Individuals differ in intelligence due to differences in both their environments and genetic heritage. Heritability estimates range from 0.4 to 0.8 (on a scale from 0 to 1), most thereby indicating that genetics plays a bigger role than does environment in creating IQ differences

among individuals. (Heritability is the squared correlation of phenotype with genotype.) If all environments were to become equal for everyone, heritability would rise to 100% because all remaining differences in IQ would necessarily be genetic in origin.

15. Members of the same family also tend to differ substantially in intelligence (by an average of about 12 IQ points) for both genetic and environmental reasons. They differ genetically because biological brothers and sisters share exactly half their genes with each parent and, on the average, only half with each other. They also differ in IQ because they experience different environments within the same family.

16. That IQ may be highly heritable does not mean that it is not affected by the environment. Individuals are not born with fixed, unchangeable levels of intelligence (no one claims they are). IQs do gradually stabilize during childhood, however, and generally change little thereafter.

17. Although the environment is important in creating IQ differences, we do not know yet how to manipulate it to raise low IQs permanently. Whether recent attempts show promise is still a matter of considerable scientific debate.

18. Genetically caused differences are not necessarily irremediable (consider diabetes, poor vision, and phenylketonuria), nor are environmentally caused ones necessarily remediable (consider injuries, poisons, severe neglect, and some diseases). Both may be preventable to some extent.

Source and Stability of Between-Group Differences

19. There is no persuasive evidence that the IQ bell curves for different racial-ethnic groups are converging. Surveys in some years show that gaps in academic achievement have narrowed a bit for some races, ages, school subjects and skill levels, but this picture seems too mixed to reflect a general shift in IQ levels themselves.

20. Racial-ethnic differences in IQ bell curves are essentially the same when youngsters leave high school as when they enter first grade. However, because bright youngsters learn faster than slow learners, these same IQ differences lead to growing disparities in amount learned as youngsters progress from grades one to 12. As large national surveys continue to show, black 17-year-olds perform, on the average, more like white 13-year-olds in reading, math, and science, with Hispanics in between.

21. The reasons that blacks differ among themselves in intelligence appear to be basically the same as those for why whites (or Asians or Hispanics) differ among themselves. Both environment and genetic heredity are involved.

22. There is no definitive answer to why IQ bell curves differ across racial-ethnic groups. The reasons for these IQ differences between groups may be markedly different from the reasons for why individuals differ among themselves within any particular group (whites or blacks or Asians). In fact, it is wrong to assume, as many do, that the reasons some individuals in a population have high IQs but others have low IQs must be the same reason why some populations contain more such high (or low) IQ individuals than others. Most experts believe that environment is important in pushing the bell curves apart, but that genetics could be involved too.

23. Racial-ethnic differences are somewhat smaller but still substantial for individuals from the same socioeconomic backgrounds. To illustrate, black students from prosperous families tend to score higher in IQ than blacks from poor families, but they score no higher, on average, than whites from poor families.

24. Almost all Americans who identify themselves as black have white ancestors—the white admixture is about 20%, on average—and many self-designated whites, Hispanics, and others likewise have mixed ancestry. Because research on intelligence relies on self-classification into distinct racial categories, as does most other social-science research, its findings likewise relate to some unclear mixture of social and biological distinctions among groups (no one claims otherwise).

Implications for Social Policy

25. The research findings neither dictate nor preclude any particular social policy, because they can never determine our goals. They can, however, help us estimate the likely success and side-effects of pursuing those goals via different means.

Table 1

Examples Illustrating How Scientific Debate on Intelligence and IQ Tests Has Advanced Over the Last 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 constitute or contribute to overall intelligence? (Intelligence is conceptualized as a complex conglomeration of specific abilities.)	To what extent does *g* constitute the core of different specific abilities? (*g* is conceptualized as psychometrically [not biologically] unitary, and specific abilities as complex conglomerations of higher-order abilities, with *g* being their major component.)
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	Raw scores on IQ tests have risen over time (the *Flynn Effect*), so do IQ tests measure different things in different epochs, or has

disparate impact against blacks and Hispanics?	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? (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?

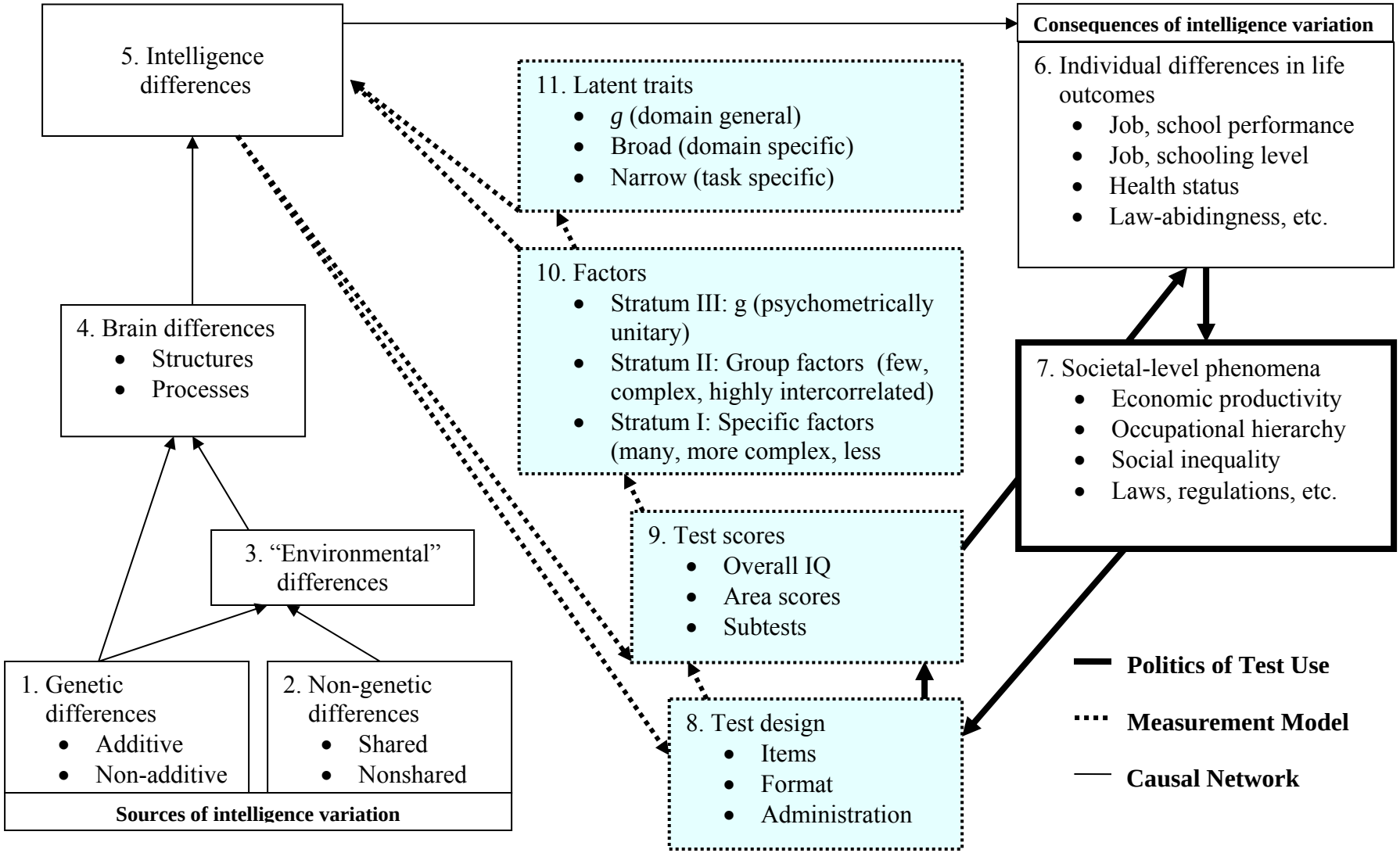
Does a whole battery of different ability tests (verbal, spatial, etc.) predict outcomes (e.g., educational or occupational) substantially better than just an overall IQ score?	Which classes of cognitive and non-cognitive tests provide incremental validity, when used with *g*, in predicting performance on different classes of tasks (instrumental, socioemotional)?
Do IQ tests predict performance of non-academic 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 (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/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 clinician's 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 non-cognitive talents, and selection into gifted programs rely on teacher, parent, and self ratings?

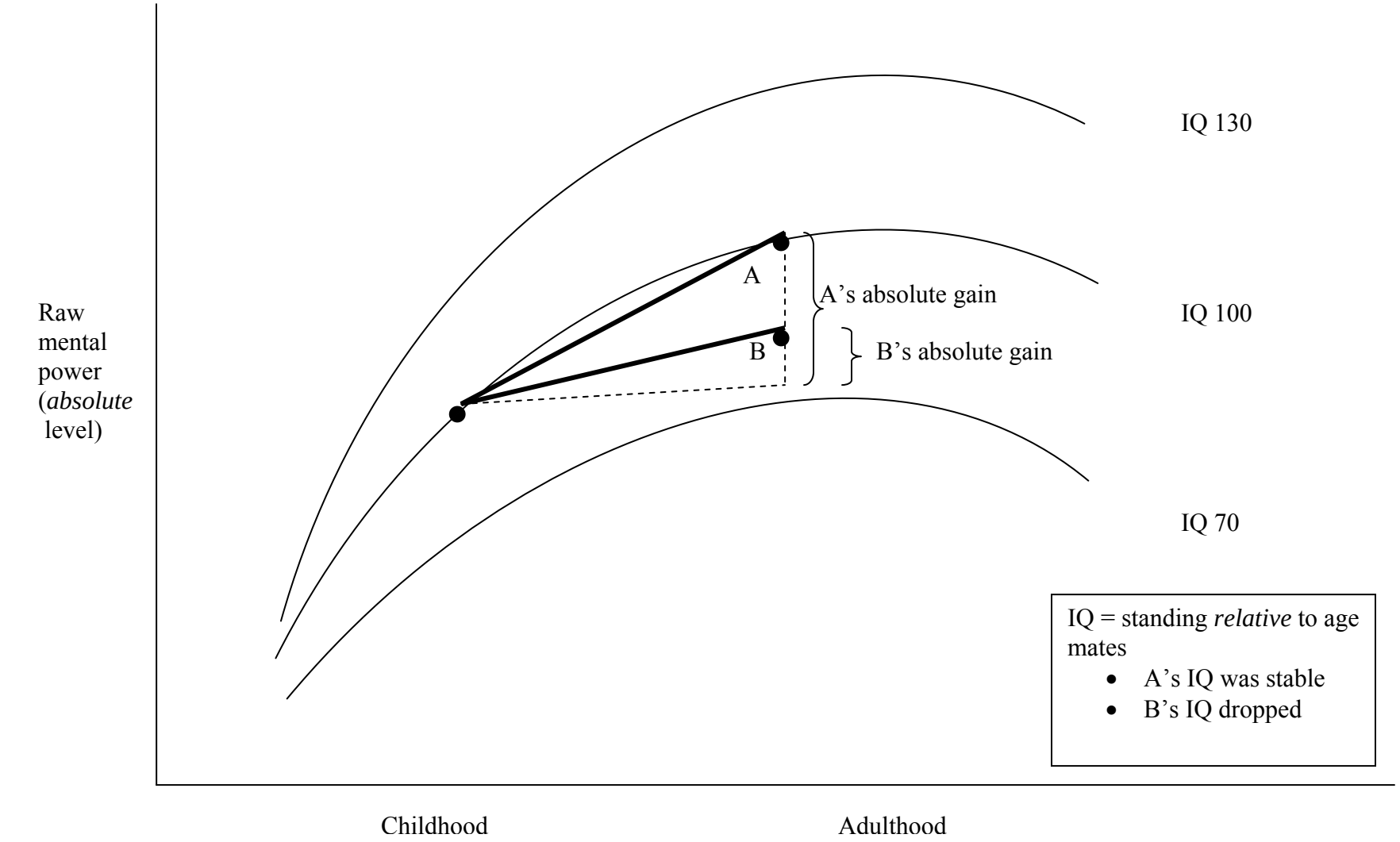
Should employers give less weight to technical expertise and more to organizational citizenship when hiring employees in order to improve racial balance?	Should colleges give less weight to cognitive abilities and more to non-cognitive strengths when admitting students in order to improve racial balance?
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 non-cognitive tests should employers use *instead of* cognitive tests when selecting employees?	Which non-cognitive tests should employers use *in addition to* cognitive tests when selecting employees?
Should courts prohibit schools from administering IQ tests to black students?	Which IQ tests and test norms should courts use in determining an individual's eligibility for the death penalty (i.e., whether they score above IQ 70).

Figure Captions

Figure 1. Three foci of fallacious reasoning: Measurement of intelligence, causes and consequences of intelligence differences, and the politics of test use.

Figure 2. An oft-muddled distinction: Changes in relative vs. absolute levels of intelligence.


5. Intelligence differences
Consequences of intelligence variation
6. Individual differences in life outcomes
• Job, school performance
• Job, schooling level
• Health status
• Law-abidingness, etc.
11. Latent traits
• g (domain general)
• Broad (domain specific)
• Narrow (task specific)
10. Factors
• Stratum III: g (psychometrically unitary)
• Stratum II: Group factors (few, complex, highly intercorrelated)
• Stratum I: Specific factors (many, more complex, less
4. Brain differences
• Structures
• Processes
7. Societal-level phenomena
• Economic productivity
• Occupational hierarchy
• Social inequality
• Laws, regulations, etc.
3. "Environmental" differences
9. Test scores
• Overall IQ
• Area scores
• Subtests
1. Genetic differences
• Additive
• Non-additive
2. Non-genetic differences
• Shared
• Nonshared
8. Test design
• Items
• Format
• Administration
Sources of intelligence variation
Politics of Test Use
Measurement Model
Causal Network


Raw mental power (*absolute* level)
IQ 130
IQ 100
IQ 70
A
B
A's absolute gain
B's absolute gain
IQ = standing *relative* to age mates
• A's IQ was stable
• B's IQ dropped
Childhood
Adulthood

Endnotes

[1] Briefly, Herrnstein and Murray (1994) state 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 being more intelligent or smarter than others, individuals' IQ scores are relatively stable throughout their lives, properly administered IQ tests are not demonstrably culturally biased, and individual differences in intelligence are substantially heritable.

[2] This is often referred to as "face validity," "content validity," or "fidelity." All are important considerations for tests of achievement or for tests meant to predict performance in a particular content domain. All refer to test content looking like the knowledge or skill to be measured, for example, repairing a jet engine, typing a business letter, or knowing algebra. All tend to increase public acceptance of a test, but neither their presence nor absence tells us anything about the validity of a test for measuring an unobservable construct.

[3] There is much debate, however, about whether *g* is unitary at the *physiological* level (Entry 4). This is a very different matter (Jensen, 1998, 2006).

[4] In more general terms, an *age equivalent* is derived This is analogous to the *grade equivalent,* which is frequently used today in reporting academic achievement in elementary school: "Susie's grade equivalent (GE) score on the district math test is 4.3, that is, she scored at the average for children in the third month of Grade 4."

[5] Thermometers illustrate another limitation of IQ tests. We cannot be sure that IQ tests provide interval-level measurement rather than just ordinal-level measurement (i.e., rank order). 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 like an interval-level scale has been a reasonable and workable assumption for many purposes, but we really do not know if a 10-point difference measures off the same intellectual difference at all ranges of IQ.

[6] The technical term *gene-environment interaction* usually refers to a particular kind of non-additive genetic effect, namely, where environmental (nongenetic) effects are conditional on genotype (e.g., an allele may make one more susceptible to adverse environments).

[7] The Horn-Cattell model claims there are two *g*s, fluid and crystallized, but can do so only by stopping the factoring process just below this most general level.

[8] These Stratum II abilities appear to incorporate four of 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.

[9] Item response theory (IRT) is used to examine the properties of individual items, not the construct validity of an entire test.