ASSESS AND ASSIST INDIVIDUALS, NOT SEXES

LINDA S. GOTTFREDSON
UNIVERSITY OF DELAWARE

Halpern highlights the bane of all mental testers—tests tell truths that many people prefer not to hear. In this case, the unwelcome truth is that males and females tend to differ in their profiles of skills, abilities, and knowledge. Halpern suggests that the Education Act of 2001 will thrust this unwanted news before more people, because it requires more widespread and systematic assessment of students' academic proficiencies. The risk is that a legislative effort meant to *improve* all students' learning could, by aggravating the "politically explosive" issue of group differences in ability, be side-tracked into contentious debates over *equalizing* test scores.

THE FACTS AT ISSUE

I appreciate Halpern's providing a comprehensive, matter-of-fact overview of sex differences, and one, moreover, that does not seek political cover by immediately discounting the differences as socially constructed. Her account is comprehensive in a number of ways. Halpern describes the wide array of competencies on which the sexes have been shown to differ, sometimes hardly at all but occasionally a lot. She describes the developmental course of these competencies, to the extent that we know it. And she takes care to describe the disabilities, not just the strengths, that characterize each sex more than the other. In short, she creates a detailed portrait of the empirical reality of intellectual differences between American boys and girls at different ages.

Not only does Halpern describe the pattern of sex differences well, but she also leaves no doubt that they are rooted to some extent in biology. She does so by describing their developmental course partly in biological terms, and by introducing a psychobiosocial model of development that stresses the ceaseless interaction between nature and nurture. Her acceptance of a biological-genetic component to human abilities reflects an obeisance to scientific evidence that is lacking in most public reports of sex differences (including reports from national testing organizations), which generally attempt to defuse potential criticism by quickly asserting that the group differences they report result entirely from discriminatory social practices.

Direct all correspondence to: Linda S. Gottfredson, School of Education, University of Delaware, e-mail: gottfred@udel.edu

Issues in Education, Volume 8, Number 1, 2002, pages 39–47

ISSN: 1080-9724

Halpern also clarifies some basic confusions about tests that often derail productive discussion of test score differences. One of the most mischievous misconceptions is that average group differences in test scores signal that a test is *necessarily* biased against the lower-scoring group. Jensen (1980) has dubbed this the "egalitarian fallacy," because it merely assumes a preferred answer to an empirical question, in this case, whether the sexes possess equal average levels of the skills being assessed. Test *bias*, as Halpern points out, is a systematic mismeasurement of abilities that results in underestimating the true abilities of members of some groups relative to others. An example of the egalitarian fallacy would be to conclude that rulers must be biased against women (give them artificially low scores) because they tell us that men tend to be taller than women. Some differences, however, are real. Whether we think this situation *fair* or not is an entirely different matter. As Halpern suggests, diatribes against rulers would be no way to achieve equality in height.

Next, Halpern reminds readers what good testing practice requires most of all, namely, that any inferences drawn from test scores be warranted. That is the basic issue of *validity*, as described in any good textbook on the subject. The issue is this. We might have in mind many possible uses of a test, but the test scores will likely support only some of those uses, that is, be valid for some purposes but not others. So, for instance, if the girls in Mr. Jones' English class score better than the boys on a professionally developed test of writing, we might properly infer that the girls in the class actually do write better than the boys, on the average. However, we could not infer, on the basis of the test scores alone, that Mr. Jones has taught the girls more effectively than the boys, that girls received more encouragement from their parents to write well, or that boys are genetically less apt. The latter three inferences cannot be justified on the basis of test scores alone, and thus would reflect *invalid uses of the test scores*.

Finally, Halpern points to an underappreciated phenomenon with big sociopolitical consequences: any average difference between two score distributions (two bell curves) are magnified at the tails of those distributions. More specifically, if one demographic group has a higher average score or is more variable than another, then the proportions of the two groups that are above (or below) some particular score become increasingly lopsided the more extreme the score being considered. For example, as she notes, the ratio of males to females scoring above certain levels in SAT-Math rises sharply with SAT level: 2:1 above 500, 5:1 above 600, and 17:1 above 700.

THE POLITICS OF THE FACTS

In short, I applaud Halpern for describing some of the basic facts about which knowledgeable experts tend to agree but which may discomfit segments of the public: most generally, that there are average sex differences in many skills, abilities, and knowledges, and that they have a partly biological basis. Her laudable aim is that this information not be misconstrued and that tests providing it not be misused.

Such truth-telling can be a bruising public service to render, however, precisely because it shines light on facts that some would angrily deny or dispute, whatever the evidence proves. It therefore is not surprising that researchers differ in how they couch unwelcome

news. Sometimes, in attempting to make it more palatable to potential critics, they inadvertently send mixed messages that allow their audience to minimize or dismiss the very information they wish to convey. I would therefore like to clarify what might seem to be such messages in Halpern's presentation. They pertain to the interpretation of sex differences as well as to how we should respond to them. My concern is that mixed messages retard rather than speed public understanding and constructive action on a vexed issue.

Can We Do We Have Fair, Meaningful, and Valid Tests for Boys and Girls?

Halpern asks whether we can design "assessments that are fair, meaningful, and valid for girls and boys," but she provides no explicit answer to her question. She explains the distinctions between test bias, validity, and fairness and the importance of each, but reports no professional conclusions about how standardized tests fare against these criteria. This silence, together with her reference to a cacophony of litigants wanting "fairer" achievement tests, might leave the mistaken impression that the possibility of designing valid, fair, and meaningful tests is still an open question scientifically. It is not. Her own implicit acceptance of the reality of sex differences suggests that she herself considers that question settled in the affirmative. Indeed, there already exist many valid, meaningful tests, including the ones from which she draws her results.

As Halpern suggests, there are special techniques for determining whether a particular test is culturally biased. What she does not say, however, is that professionally developed tests *do* pass the technical procedures designed to reveal cultural bias. This concern was settled to most researchers' satisfaction two decades ago for the major mental tests when administered to native-born English-speaking Americans, including blacks and women (Jensen, 1980; Wigdor & Garner, 1982). If such procedures were to reveal a particular test to be biased against some particular group, a psychologist who used it with affected individuals would be deemed professionally irresponsible. Institutions that used biased tests to select workers or students could be sued for illegal discrimination. Professional test standards also require that publishers provide evidence that their tests actually measure or predict what they claim (i.e., that the tests have *construct or predictive validity*). The market is awash with technically excellent tests of all sorts.

Test *fairness* will always be a contentious issue, however, because it is not a technical question. Rather, it is a political matter because it involves balancing the conflicting social, educational, economic, and other aims that different people have for test use or non-use. Some people, for instance, insist that technically excellent (valid and unbiased) tests are unfair when members of some races fail them in disproportionate numbers. In such cases, they argue, it is fairer (to groups) to use tests that are both less valid and biased in favor of the lower-scoring groups, but which yield more equal results across groups (e.g., race-normed tests; Gottfredson, 1994). Other people vehemently disagree, insisting that fairness (to individuals) requires that tests be not only unbiased but also as predictively valid as possible.

Test developers and publishers are required by professional test standards to encourage proper use of their tests, but they obviously cannot control the uses and misuses to which schools, employers, and others actually put their tests or the often errant conclusions that

politicians, social activists, and others draw from them. Their lack of direct control is the reason why experts should speak out, as Halpern has, to foster appropriate interpretation and use of test scores. Testing is ripe for misuse when political ends are at stake, as they always will be when valid tests reveal group disparities in developed competence in important arenas of life. An ambiguous message about whether tests can be made valid, meaningful, and fair for groups invites efforts to corrupt tests in the name of improving their fairness and technical adequacy (see Gottfredson, 1996, for an egregious example). It invites the game-playing that Halpern rightly deplores.

Is it Useful to Examine Genetic vs. Environmental Causes of Ability Differences?

The second mixed message concerns the biology of sex differences. Halpern suggests that ability differences have genetic as well as environmental origins, which is now so well-established a fact that it is no longer an interesting question among persons acquainted with the evidence (Plomin, DeFries, McClearn, & McGuffin, 2001). They have moved on to multivariate genetics (to what extent do different traits share the same genetic roots?), developmental genetics (to what extent are stability and change in cognitive development due to genetic vs. environmental forces?), and the genetics of environments (to what extent do our genetic proclivities shape our choice and construction of personal environments?).

Although Halpern describes how mental development has a biological basis, she nonetheless suggests that it is impossible to determine to what extent variability in that development is genetic rather than non-genetic in origin (cf. her fuller discussion in Halpern, 1997). Implying that behavior genetic information is irrelevant diverts attention from a hot-button issue (are group differences genetic?), but behavior genetics has, in fact, developed powerful analytic tools for decomposing the variance among individuals (and between groups) into its various genetic and non-genetic components. Behavior genetic methods are far from useless in understanding cognitive development—quite the opposite.

Halpern is correct that genes and environments perform an intricate and extended dance together in forming the human brain and body, a dance that begins at conception and ceases only with death. But that does not mean that one or the other, nature or nurture, might not be leading the dance, making the most of the partner that chance has assigned. The question behind the nature-nurture debate is not "how does the *typical* dance develop over a lifetime?" (in which each partner responds to the other in a "reciprocally continuous feedback loop"), but "why do some dances develop so differently than others?" That is, after all, the burning question raised by group differences in ability. *Why* do girls and boys, women and men, differ so much in their pattern of interests, abilities, and achievements? Is the cause sex discrimination, differential socialization, genetic differences, or some combination of them? What steps could parents, schools, or others take to narrow the gaps, should they wish to? Turning readers' attention to *normative* processes that affect everyone's development will not still their concern over students' *differences* in development. Only behavior genetic research can unravel the much-tangled sources of those differences.

The best-known distinction that behavior genetic research draws is between the genetic and environmental (non-genetic) sources of ability differences. Such research can estimate the overall *heritability* (or *environmentality*) of any measurable personal attribute, such as intelligence, vocational interests, personality, attitudes, life events, and rearing "environments." A crude but useful rule of thumb is that all enduring personal traits are about 50% heritable in typical Western populations. This means that equalizing our environments (opportunities, instruction, and so on) would not equalize our expressed abilities and personalities, only narrow the differences somewhat.

Genetic variability is important in educational contexts because it makes students differentially responsive to instruction. That variability probably also produces an imbalance in the *proportion* of boys relative to girls who are especially sensitive or insensitive to instruction in specific kinds of skills. If so, equalizing instruction may do little to eradicate sex differences. The large female superiority in writing skill exists despite much instruction in writing, so there is little reason to expect that introducing instruction in spatial skills would do much to reduce the large male superiority in spatial transformation. In fact, instruction can widen gaps in complex skills among differentially apt students.

By the same token, 50% heritability means that there is also considerable *non*-genetic influence on personal traits. Behavior genetics draws another distinction that subdivides this environmental component. The distinction is less obvious but often more important than the one between nature and nurture. It is that some non-genetic effects are *shared* whereas others are *non-shared*. Shared influences are ones that affect all children in a family and therefore make siblings more alike, whereas non-shared influences affect individuals uniquely, one at a time, and therefore make siblings less alike. Most developmental theories in the social sciences, such as those emphasizing social class or parental child-rearing practices, assume that environmental influences on mental abilities are shared.

Behavior genetics has found the opposite: non-shared effects are the rule and shared effects the exception. Shared effects are sometimes sizeable in childhood (e.g., for intelligence), but their influence soon fades. This is illustrated by adoptive siblings raised in the same home, because their early correlation in IQ dissipates by adolescence, at which time they are no more alike in IQ (or personality) than random strangers. In contrast, identical twins reared *apart* (MZAs) are much more similar than fraternal twins reared together (DZTs), and they are almost as similar as identical twins reared together (MZTs).

It is important to point out that the story is a bit different for general vs. specific abilities (i.e., general intelligence vs. verbal or spatial ability) and for mental abilities of either kind vs. academic achievements (e.g., mathematics). First, the heritability of the three categories differs, with general intelligence being highly heritable, specific abilities somewhat less so, and particular academic achievements being only moderately heritable. Second, there are no lasting shared effects on intelligence and only "negligible" ones on specific abilities, but there are some shared effects, albeit small, on achievement. In all three categories, most environmental effects are non-shared. It is therefore not surprising that the moderately high correlations among intelligence and the specific abilities are almost entirely genetically mediated, as are the more modest correlations of both with academic achievement (e.g., Plomin et al., 2001, ch 10; Wadsworth, De Fries, Fulker, & Plomin, 1995). The sex differences that Halpern discusses seem to range broadly over the three

classes of competency, so it is likely that the skills in question are differentially susceptible to environmental effects, especially of the shared sort.

Twins are generally cited to illustrate the importance of genetic factors, as I did above, but identical twins reared *together* (MZTs) are useful for illustrating the importance of non-shared *environmental* factors (Pike, Reiss, Hetherington, & Plomin, 1996): these twins differ somewhat in basic traits all their lives despite sharing identical genes and the same rearing conditions. Other family studies confirm the enduring importance of non-shared influences, which sometimes rival the importance of genetic effects. What these non-shared influences consist of is still a puzzle: what varies within families, but not between them on the average? Behavior genetics is well suited to tackling this unexpected puzzle, and it has started to do in earnest, for example, in the Nonshared Environment and Adolescent Development (NEAD) project (Reiss, Neidhiser, Hetherington, & Plomin, 2000).

It is hard to overstate the importance of the discovery that shared environmental influences pale in importance relative to non-shared ones in explaining divergence in development. It means that social scientists have been looking in all the wrong places for the most important environmental effects on enduring personal traits, such as personality and intelligence. It turns out, perhaps counterintuitively, that both genes and environments operate to make individuals unique, with the moderate family resemblance in our most basic psychological traits being almost entirely genetic by adulthood.

In short, behavior genetic research should be welcomed into, not banished from, the collective effort to understand individual and group differences in development. Failing to appreciate the great genetic diversity among students leads to false expectations that we can equalize the skills of individuals and groups by equalizing their learning opportunities. It leads us to ignore how students' genetic propensities often lead them to seek out, reject, construct, or remake their own formative experiences, including how they use or misuse their time in the classroom (Scarr, 1996). It is literally impossible to provide equal environments to beings who exploit them differently. As for the myth of powerful shared environments, it produces false expectations that a society can reconfigure its members' abilities in specified ways, if only it "had the will" to do so. While behavior genetics has thus revealed some of the constraints we face in reshaping each other, its discovery of important non-shared effects has also opened hitherto unsuspected vistas for spotting specific environmental influences on development.

While eschewing discussion of genetic causes of group differences, Halpern turns our attention to "an exciting area of recent research" that proposes a non-genetic, discriminatory part-cause for such differences: stereotype threat. It is the basis for her recommendation that schools provide anti-prejudice instruction, presumably to remedy male as well as female deficits in skill. As Halpern herself notes, however, the hypothesized effect has not been found for either blacks or females in real-life testing situations. Moreover, stereotype threat researchers have yet to show that this unconscious phenomenon—to the extent it actually occurs—represents more than the transient induction of anxiety, too much of which can temporarily interfere with the *effective use* of one's skills.

Many social scientists nonetheless clasp this small and inconsistent body of research on stereotype threat like a life-preserver salvaged from the sinking ship of shared family effects theory, which has failed to explain more than a small portion of either individual or

group differences in mental abilities. Enthusiastic appeals to stereotype threat encourage us to believe something implausible, namely, that vague psychological threats might create the big group differences in cognitive ability (and the small ones, too, presumably) that intense direct instruction in preschool interventions and K-12 schooling do virtually nothing to remediate. Those appeals encourage social blame that may be unwarranted and interventions that may be useless.

In any case, if stereotype threat truly were a widespread and important phenomenon, it would show up as *test bias*, that is, as causing an underestimation of what girls and blacks (and boys) *actually* accomplish in the classroom and elsewhere. I have seen no such evidence. As already discussed, decades of research on test bias have yielded no evidence of bias in the major tests of mental ability and academic achievement. (The frequent underprediction of females' college grades by the SAT seems due mostly to females' entering easier majors and being more conscientious in their studies [Young, 2001].)

RESPONDING TO THE FACTS

The challenge with tests is not whether fair and meaningful ones can be created (they have been), but whether they will be *used* in a fair, valid, and meaningful manner. That is why Halpern's recommendations actually focus on how we should *interpret* test scores, especially average differences by sex. Because group differences in developed abilities are such a sensitive political issue, urging proper interpretation is a special challenge. The reporter of facts must simultaneously encourage people to take them seriously (the differences are real, sometimes large, and often have practical consequences) while simultaneously discouraging destructive overreaction, including angry denial, sad resignation, and smug acceptance. It is hard to find the right balance, and scholar-reporters have different instincts about where it lies. Halpern has found a good balance. I will review her recommendations, sometimes refining or reconfiguring them slightly to reflect my own observations of how the facts are often misinterpreted.

Provide the Broad Context for Interpreting Differences

Halpern provides essential context in five ways. Each works to prevent people from jumping to false conclusions, whether by attributing too much importance to ability differences or too little. Halpern has focused primarily on the former problem, so I have appended cautions meant to forestall the latter as well.

- There are wider mental ability differences within groups than between them, but, as Halpern notes, small average differences can produce striking disparities at the upper and lower ends of the distribution.
- Many different skills are useful in today's economy, but some are more useful or highly recompensed than others (because they are more general or scarcer).
- Virtually all students can improve their skills, but we cannot expect students of different ability levels to profit equally from the same instruction or practice.
- Differences in ability are not "fixed" or "immutable," but some (the more general ones) are more resistant to change than others.

- Differences are not necessarily deficiencies, but they sometimes are (e.g., dyslexia and low intelligence do, in fact, create big functional disadvantages).

Follow Standard Professional Practice When Acting Upon Students' Test Scores

Many of Halpern's suggestions for using test results appropriately are, in essence, simply a call for following established professional principles. None is specific to sex differences—which is precisely the point. This guidance requires attending to each student's uniqueness in order to optimize every student's growth. It is meant to forestall treating individuals as clones of their groups, which they never are.

- Do not judge individuals by group averages, because there is enormous variability within all groups.
- Be alert for change after assigning students to treatments based on test scores, because individuals (and the two sexes) mature at different rates.
- Provide all students an opportunity to develop to their fullest. I would add: realize, however, that differences in ability create different needs for development (e.g., complexity of instruction, amount of assistance).
- Strive for the best achievement from *all* students. Realize, however, that differences in ability forecast different levels of best achievement, all else equal.

Don't Let the Tail of Group Differences Wag the Dog of Test Development and Use

"What should schools do about sex disparities, if anything?" is a political question about which reasonable people will disagree. It often comes camouflaged as a scientific matter, however, when technical pretexts are provided for corrupting valid tests for ideological ends. For instance, all too many employers and educational institutions have started picking and choosing among tests and assessments (or the items on them) to create assessment batteries yielding more "politically correct" results, even when that requires *introducing* bias into the tests and *reducing* their validity, all the while trumpeting technical superiority for their new procedures. The last two decades have produced a sorry legacy of "scientifically-justified" attempts to reconfigure tests (racially gerrymander their content) or to rescore them by giving bonus points for belonging to a lower scoring group (race-norming). When exposed as having no real scientific justification, the ideologically-driven degradation of tests has provoked not only a Congressional ban on race-norming employment tests but also outrage and high-profile embarrassment in some professional circles (Gottfredson, 1994, 1996).

The public schools should avoid any temptation to go down that road. Their charge is to serve individuals, not sexes or races. Halpern's recommendation to "use a variety of measures to capture the strengths of both sexes" might be misinterpreted as an invitation to compile sex-balanced test batteries in order to get more politically palatable outcomes, regardless of their educational relevance. I would therefore restate her recommendation to emphasize that educational relevance must always be the guiding principle in test selection and use: any test battery should assess the *important* skills that schools *can and should* teach, some of which *may* have been overlooked before.

CONCLUSION

People do not differ in cognitive ability simply because society treats them differently, and treating them the same will not level their differences in ability. Neither, even, will treating them in a compensatory manner, in which the least able are given the most resources and the most able the fewest. Tests expose the basis for this democratic dilemma: people differ in consequential cognitive abilities. Schools do not create this situation, but they are often held accountable for it nonetheless. In self-defense, schools may be tempted to hold the "ruler" or "thermometer" accountable.

Schools would do better to follow Halpern's urging to focus on students as individuals. Providing equal opportunity to diverse students does not mean providing *identical* opportunity for all. Rather, it means providing a diverse menu of opportunities, from which students and their mentors can select the most fitting for the child's particular needs and interests. Such action will never produce the anthill of identically-skilled beings that some seem to desire, but it will promote the individual growth—the "best achievements"—that dedicated educators seek from their students.

REFERENCES

Gottfredson, L. S. (1994). 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 U.S. Justice Department: A case study. *Psychology, Public Policy, and Law, 2*(3/4), 418-446.

Halpern, D. F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist, 52*, 1091-1102.

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

Pike, A., Reiss, D., Hetherington, E. M., & Plomin, R. (1996). Using MZ differences in the search for nonshared environmental effects. *Journal of Child Psychology and Psychiatry, 37*(6), 695-704.

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

Reiss, D., Neiderhiser, J. M., Hetherington, E. M., & Plomin, R. (2000). *The relationship code: Deciphering genetic and social patterns in adolescent development*. Cambridge, MA: Harvard University Press.

Scarr, S. (1996). How people make their own environments: Implications for parents and policy makers. *Psychology, Public Policy, and Law, 2*(2), 204-228.

Wadsworth, S. J., DeFries, J. C., Fulker, D. W., & Plomin, R. (1995). Covariation among measures of cognitive ability and academic achievement in the Colorado Adoption Project: Sibling analysis. *Personality and Individual Differences, 18*(1), 63-73.

Wigdor, A. K., & Garner, W. R. (Eds.) (1982). *Ability testing: Uses, consequences, and controversies*. Washington, DC: National Academy Press.

Young, J. W. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis*. (College Board Research Report No. 2001-6) New York: College Entrance Examination Board.