Ability tests, the internet, and practice tests: A recipe for invalidity

David Lohman
Professor Emeritus, The University of Iowa

September 2013

Ability test scores provide essential information for inferences about giftedness. They are most important for those who, through lack of experience or opportunity, have not yet developed unusually high levels of competence in a socially valued domain. Therefore, ability tests are most useful for young children and for students of any age who have not developed strong academic skills (or other valued forms of expertise). Indeed, poor and minority students typically perform better on ability tests than on well-constructed achievement tests. On the other hand, ability test scores of such students still trail the scores of their more advantaged classmates, especially on ability tests that emphasize the kinds of verbal reasoning skills that are required for and developed through formal schooling. Also, since ability tests are usually best understood as measuring the predictor, not the criterion, children who evidence high levels of accomplishment should not be denied advanced or more challenging instruction simply because their scores on the predictor (i.e. ability test) are strong, but lower than desired.

Because ability tests measure traits that, although not fixed, are relatively stable over short periods of time, many erroneously assume that the tests themselves are impervious to external influences. This is not the case. Ability tests are sophisticated, but relatively fragile instruments. Nonverbal tests that attempt to reduce the overt impact of language and education are particularly sensitive to practice and coaching. Even without feedback, re-test gains of five to ten IQ points are commonly observed on nonverbal tests. Practice with feedback and deliberate coaching can produce even larger gains, especially for more able students.

For a long time, test users have ignored the effects of practice and minimized the larger effects of coaching by keeping secret the contents of the tests.[1] Aside from rare instances of cheating, children approached the test with no special preparation. This is no longer the case. The internet has lifted the veil of secrecy that once shrouded ability tests. The recent proliferation of practice materials sold over the internet and of coaching schools that operate in many urban areas has seriously undermined the fairness of both group and individually administered ability tests when test scores are used as the primary criterion for high-stakes admissions decisions. For a price, savvy parents with resources can virtually assure their child a high score and thus of placement in the gifted program.

[1] See E.L. Thorndike (1919). "Tests of intelligence" (Teachers College Newsletter). Also Kulik & Kulik (1984). Effects of practice on aptitude and achievement tests. AERJ, 21, 435-447.

Test developers, test users, and educational policy-makers can make changes to current policies and practices that address this problem. Test developers can provide free practice activities for all students, thereby partially re-leveling the playing field; they can develop multiple forms of their tests, thereby reducing the advantage gained by access to the items on any one form; and on some tests they can report indices that caution users when analyses of item performance suggest coaching or cheating. Although helpful, none of these changes will completely solve the problem. The most important change can only be made by those who set the policies that specify how scores on ability tests are used in the talent identification process.

An IQ score of 130 (or national percentile rank of 97) on an ability test has long been required for admission to G&T programs in many states. Although experts have cautioned against this sort of high-stakes use of ability test scores, the convenience of the policy has outweighed its negative consequences. In some states, one of the more desired outcomes has been an increase in funding: every child who scores exceed the standard brings additional resources to the program. A less desired outcome is that students with artificially high scores are often less prepared for the demands of the G&T program than some of their classmates who were not admitted. Even more problematic is the decrease in the diversity of the student population that is served by the program. The parents of poor and under-served minority students typically are not the purchasers of either practice tests or admission to test-preparation classes for their children. Unequal practice thus not only invalidates scores for those who receive it, but effectively unravels the often considerable efforts that programs have made to diversify the population of students that they serve.

Reducing the stakes attached to scores on the ability test is the key to fairer and more defensible policies. If the stakes are reduced, then the ability test score can become one of the more lenient criteria in the selection procedure rather than the most restrictive criterion. A good example is Renzulli's recommendation to use ability tests to help identify a large and diverse talent pool. Children whose needs can be addressed by the talent development program are assigned to different types and levels of intervention using evidence on interests, accomplishments, and scholastic achievement. The scores on a multi-factor ability test are set as a lower bound, typically using local stanines or local percentile ranks that make eligible for consideration the top 15 - 25 percent of the student population. This seemingly liberal standard actually comports well with what we know about the plurality of abilities and their imperfect relationships achievements in different domains. Using local norms allows all schools to identify those students whose levels of cognitive or academic development are significantly in advance of their classmates. However, local norms are difficult to implement when the goal of the program is to identify gifted students, since giftedness is then easily conflated with IQ's or national (rather than local) percentile ranks.

Setting a more generous cut score on the ability test simultaneously reduces the stakes placed on the ability test and increases the importance of evidence of accomplishment and creative production. Reducing the stakes on the ability test also reduces the need for and value of external practice on the test, thereby preserving the validity of the scores students obtain on it. Emphasizing accomplishments in science, mathematics, and the arts encourages substantive preparation in these domains rather than unproductive cramming for an ability test.

Therefore, paradoxically, the most effective way to preserve the integrity of ability tests and of the important information on giftedness that they provide is to rely less on the attainment of a particular IQ or national percentile rank score and instead to rely more heavily on other sources of information. Although such policies conflict with established practices in programs that have a single pull out class for gifted children, they mesh well with programs that offer different kinds and levels of enrichment and acceleration that serve a broader and more diverse student population.

The current policies on test use for gifted identification usually follow the recommendations of state and national organizations that serve gifted children. These policies reflected the best thinking of experts trained in the pre-internet era. These experts could not have foreseen how the internet would change the landscape. Therefore, you have a critical role in advocating for policies on test use that better comports with the realities of test use (and misuse) that we now see.

In the mean time, here are some things that you can do:

1) Avoid (or change) policies that encourage retesting. These tend to advantage the advantaged.
2) Reduce practice effects by using the free Practice Activities for developing thinking skills and test-wiseness.
3) On CogAT, confirm the integrity of the scores
    a) No Warnings – especially:
        i) "Inconsistent Response Pattern" (one subtest out of line?)
        ii) "Many Items Omitted" (slow and accurate response style – levels 9+)
    b) Confidence intervals for any battery unusual?
    c) Extreme or large weakness in one area when using composite or "and" rule
    d) Avoid retesting by ignoring score on battery that is questionable; use remaining two batteries and/or their partial composite (see Norms Manual)

4) If you must retest:
    a) Know that scores are likely to increase - especially on picture-based and unfamiliar quant tests
    b) Use stanines or broad-bands of PR scores rather than a single, fixed SAS score
    c) If you must retest with the form of the same test, wait at least 3 months
    d) Use an alternative form of the same test
    e) On CogAT – go up one level (50% new items)
        i) Cautions & advantages for going from Level 8 to Level 9
    f) Use a different test that is not announced in advance