Rising Scores on Intelligence Tests

Test scores are certainly going up all over the world, but whether intelligence itself has risen remains controversial

Ulric Neisser

This article originally appeared in the September-October 1997 issue of American Scientist.

Average scores on intelligence tests are rising substantially and consistently, all over the world. These gains have been going on for the better part of a century-essentially ever since tests were invented. The rate of gain on standard broad-spectrum IQ tests amounts to three IQ points per decade, and it is even higher on certain specialized measures. In the Netherlands, for example, all male 18-year-olds take a test of abstract-reasoning ability as part of a military-induction requirement. Because the same test is used every year, it is easy to see the mean score rising, in this case, at about seven points per decade.

The cause of these enormous gains remains unknown. At this point, no one even knows whether they reflect genuine increases in intelligence or just the gradual spread of some specialized knack for taking tests. Greater sophistication about tests surely plays some role in the rise, but there are other possible contributing factors: better nutrition, more schooling, altered child-rearing practices and the technology-driven changes of culture itself. Right now, none of these factors can be ruled out; all of them may be playing some part in the increasing scores. Whatever the causes may be, the sheer size of the gains forces us to reconsider many long-held assumptions about intelligence tests and what they measure.

To focus on standardized tests-as this article does-is not to suggest that they measure every form of intelligence. Indeed, they surely do not. Supporters and opponents of testing are often at odds, but no serious scholar claims either that IQ tests measure nothing important or that they measure everything important. At the very least, they do tap certain abilities that are relevant to success in school and do so with remarkable consistency. On the other hand, many significant cognitive traits-creativity, wisdom, practical sense, social sensitivity-are obviously beyond their reach. Because there are no established measures of these subtler traits, no one knows if they are changing too. Standardized-test scores are all that we have, and they are certainly going up. This article will explore the paradoxes created by that rise and the factors that may be responsible for it.

The ABCs of Intelligence

Because there are many different forms of mental ability, there are also many different kinds of tests and test items. Some are verbal and others are visual in format. Some tests consist only of abstract-reasoning problems, and others focus on such special competencies as arithmetic, spatial imagery, reading, vocabulary, memory or general

knowledge. The broad-spectrum tests, which establish actual IQ scores, typically include a wide variety of items. Before considering these general instruments, however, we must take a brief look at the relations among different specialized tests and how those relations are traditionally interpreted.

The degree to which any two tests measure something in common can be indexed by their correlation r, which in principle ranges from -1 to +1. A positive r means that individuals who score high on one test also tend to score high on the other; a negative r, which rarely occurs in this context, means that high scores on one test go with low scores on the other. When the same group of individuals takes a number of different tests, one can compute an r for each pair of tests considered separately, and the result is a correlation matrix. For intelligence tests, the correlation matrix tends to consist of r's that are all positive, but well below 1.00.

Early in this century, the British psychologist Charles Spearman made the first formal factor analyses of such correlation matrices. He concluded that a single common factor accounted for the positive correlations among tests-a notion still accepted in principle by many psychometricians. Spearman christened it g for "general factor." In any test battery, the test that best measures g is-by definition-the one that has the highest correlations with all the others. The fact that most of these g-loaded tests typically involve some form of abstract reasoning led Spearman and his successors to regard g as the real and perhaps genetically determined essence of intelligence.

Although that view remains widely held, it is not a necessary conclusion. Other factor analyses of such data are possible and have been proposed. Today, some psychometricians regard g as little more than a statistical artifact, whereas others seem even more convinced than Spearman himself that it reflects a basic property of the brain. Whatever g may be, at least we know how to measure it. The accepted best measure is a (usually untimed) test of visual reasoning called Raven's Progressive Matrices, which was first published in 1938 by Spearman's student John C. Raven and is now available in several different levels of difficulty. As we shall see, Raven's test plays a central role in recent analyses of the worldwide rise in test scores.

In contrast to specialized instruments like the Raven, the tests most widely used in America include a wide variety of different items and subtests. The best known of these "IQ tests" are the Stanford-Binet and the various Wechsler scales. The Wechsler Intelligence Scale for Children (WISC), for example, has five "verbal" subtests (information, comprehension, arithmetic, vocabulary and explaining similarities) and five "performance" subtests in which a child must copy designs using patterned blocks, put several related pictures in their proper order and so on. A child's scores on these subtests are added up, and the tester converts the total to an IQ by noting where it falls in the established distribution of WISC scores for the appropriate age.

That distribution itself-the crucial reference for assigning IQ scores-is simply the empirical result that was obtained when the test was initially standardized. By convention, the mean of each age group in the standardization sample defines an IQ score of 100; by further convention, the standard deviation of the sample defines 15 IQ points. Given appropriate sampling and a normal distribution, this implies that about two-thirds of the population in any given age group will have IQs between 85 and 115.

IQ defined in this way reflects relative standing in an age group, not absolute achievement. The mean-scoring eight-year-old attains a higher raw score on the WISC than the mean-scoring seven-year-old, but both have IQs of 100 because they are at the middle of their distributions. So in one sense (as measured by raw scores), a normal child becomes systematically more intelligent with age; in another sense, his or her intelligence remains relatively stable. Although raw scores rise systematically throughout the school years, IQs themselves rarely change much after age 5 or 6.

The Flynn Effect

IQ tests do not remain fixed forever. Most of the major ones have been updated from time to time. The 1949 WISC, for example, was superseded by the WISC-R in 1974 and by the WISC-III in 1991. The revised versions are standardized on new samples and scored with respect to those samples alone, so the only way to compare the difficulty of two versions of a test is to conduct a separate study in which the same subjects take both versions. Many such studies have been carried out, and James Flynn, a political scientist at the University of Otago in New Zealand, summarized their results in 1984. In virtually every instance, the subjects achieved higher scores on the older version of a test. For example, in David Wechsler's own study of his revised adult test-the Wechsler Adult Intelligence Scale-Revised (WAIS-R)-a group of subjects who averaged 103.8 on the new WAIS-R had a mean of 111.3 on the older WAIS. This implies that the actual IQ-test performance of adults rose by 7.5 points between 1953 (when the old WAIS was standardized) and 1978 (when the WAIS-R was standardized), which is a rate of about 0.3 IQ points per year.

These gains are not limited to the WAIS, to adults or to the United States. In an influential series of papers, Flynn showed that the increasing raw scores appear on every major test, in every age range and in every modern industrialized country. (The rise itself is now often called "the Flynn effect.") The increase has been continuous and roughly linear from the earliest days of testing to the present. On broad-spectrum tests such as the WISC and the WAIS, Americans have gained about 3 IQ points per decade, or 15 points over a 50-year period. It is interesting to compare this total with the much-discussed gap between the mean test scores of Caucasian and African Americans, which is also about 15 points (one standard deviation of the IQ distribution). Given that the IQ of the population as a whole has increased by a similar amount just since the 1940s, that gap does not seem so large.

The pattern of score increases for different types of tests is somewhat surprising. Because children attend school longer now and have become much more familiar with the testing of school-related material, one might expect the greatest gains to occur on such content-related tests as vocabulary, arithmetic or general information. Just the opposite is the case: "Crystallized" abilities such as these have experienced relatively small gains and even occasional declines over the years. The largest Flynn effects appear instead on highly g-loaded tests such as Raven's Progressive Matrices. This test is very popular in Europe; the Dutch data mentioned earlier came from a 40-item version of Raven's test. Using the 1952 mean to define a base of 100, Flynn has calculated average Dutch Raven IQs for subsequent years. The mean in 1982 was 121.10-a gain of 21 points in only 30 years, or about seven points per decade. Data from a dozen other countries show similar trends, which seem to be continuing into the 1990s. Whatever g may be, scores on tests that measure it best are going up at twice the rate of broad-spectrum tests like the WISC and WAIS, while the tests most closely linked to school content show the smallest gains of all.

Increase or Artifact?

These gains are far too rapid to result from genetic changes. There evidently are substantial environmental influences on g, even if we do not clearly understand them at the present time. Moreover, the sheer size of the gains undermines the very concept of "general intelligence." To see why, consider that individuals with IQ scores over 130 are typically regarded as "very superior" and those with scores under 70 as "intellectually deficient." In any normal distribution of scores with a mean of 100 and a standard deviation of 15, about 2.25 percent of the population can be expected to fall into each of those categories. This was indeed roughly the case for the first Stanford-Binet standardization sample in 1932. An ongoing rise of 0.3 IQ points per year means, however, that if a representative sample of the American children of 1997 were to take that 1932 test, their average IQ would come out to be around 120. This would put about one-quarter of the 1997 population above the 130 cutoff for "very superior"-10 times as many as in 1932. Does that seem plausible?

If we go by more recent norms instead, we arrive at an equally bizarre conclusion. Judging the American children of 1932 by today's standards-considering the scores they would have obtained if they had somehow taken a test normalized this year-we find that their average IQ would have been only about 80! Hardly any of them would have scored "very superior," but nearly one-quarter would have appeared to be "deficient." Taking the tests at face value, we face one of two possible conclusions: Either America is now a nation of shining intellects, or it was then a nation of dolts.

If we focus instead on the most g-loaded tests, the gains seem even more preposterous. The mean Raven IQ in the Netherlands rose by 21 points between 1952 and 1982; extrapolating backward, there has probably been something like a 35-point increase since

the 1930s. Dutch 19-year-olds today are getting scores that would have been more than two standard deviations above the mean in their grandfathers' time. The size of these gains boggles the mind. If they do not reflect some trivial artifact, we (and especially the Dutch!) must be living in a truly remarkable age of genius. As Flynn puts it, the data imply that dozens of nations should now be in the midst of "a cultural renaissance too great to be overlooked." Because that does not seem to be happening, Flynn concludes that the tests do not measure intelligence but only a minor sort of "abstract problem-solving ability" with little practical significance.

Flynn's extreme position can help us organize the range of hypotheses that have been proposed to explain the rise in scores. Some hypotheses-including increases in test-taking sophistication and in the motivation to score well-are consistent with the view that there has been no real rise in intelligence at all. Other possibilities-including the impact of worldwide improvements in health and nutrition-conflict with Flynn's conclusion, suggesting that intelligence has really gone up. Several more subtle hypotheses-that the gains have been produced by changes in schooling, in child-rearing practices or by more general aspects of culture-lie in between these two extremes. Let us consider these possibilities one at a time.

Test-Taking Sophistication

The first large-scale application of IQ testing occurred during World War I, when psychologists employed by the U.S. Army tested more than one million raw recruits. A generation later, World War II psychologists found that the draftees of 1941-45 were achieving substantially higher scores than those of 1917-18. This rise--now considered as an early manifestation of the Flynn effect--seemed easy to explain at first: The draftees of World War II were better educated and far more familiar with mental tests, which by then had become an accepted part of American culture. This explanation seemed plausible enough at the time, but that was half a century and 15 IQ points ago. For how many generations can we continue to appeal to increasing sophistication about tests?

It is true that teaching today has become increasingly geared to certain kinds of achievement tests, and many students have learned test-taking strategies that were less widely known in the 1940s. But this hypothesis, like most others that appeal to the effects of schooling, predicts that the largest gains should appear in subjects most closely related to school content. Why then do the greatest increases appear on abstract-reasoning tests such as the Raven? Moreover, even children who take the very same test a second time usually gain only 5 or 6 points by doing so, which seems to set an upper limit on the effects of test sophistication. In short, increased familiarity with tests in general cannot fully explain the Flynn effect.

Nutrition

Today's average adult from an industrialized nation stands much taller than the comparable adult of a century ago. That increase in stature-almost certainly the result of

general improvements in nutrition and health-has come at a rate of more than a centimeter per decade. Available data suggest that these gains have been accompanied by analogous increases in head size, and presumably by an increase in the average size of the brain. Richard Lynn of the University of Northern Ireland argues that this is the only significant cause of the Flynn effect: Larger brains produce higher levels of intelligence. According to his interpretation, the rise in test scores is no artifact; it indexes genuine gains in cognitive ability.

The most obvious objection to Lynn's proposal is that it explains too much. To treat the gains as genuine and essentially biological is to concede that we are indeed vastly more intelligent than our grandparents-a conclusion that seems intuitively unpalatable. There is also an empirical problem: It is difficult to demonstrate the direct connection between diet and intelligence that the hypothesis requires. Severe malnutrition in childhood almost certainly produces negative cognitive effects, but the fact that it usually occurs together with other forms of deprivation makes those effects difficult to analyze. Experimental studies involving dietary supplements have rarely included adequate control groups, and the few studies with what seem to be positive results have proved difficult to replicate. Although it is sometimes suggested that malnutrition produces a greater effect on visual-spatial than on verbal skills-which would be compatible with the pattern of observed long-term gains in test scores--this too has not been firmly established. Taken together, the evidence linking nutritional levels to intelligence is shaky at this point.

Schooling

In the countries where IQ scores are rising, people are also staying in school much longer than their parents and grandparents did. Could sheer duration of schooling be responsible for the gains? This hypothesis is plausible because the general effect of schooling on IQ is very well established. Many studies indicate that children who do not attend school for one reason or another score lower on the tests than their regularly attending peers. One especially unfortunate example of that principle appeared in the 1960s, when some Virginia counties closed their public schools to avoid racial integration. Compensatory private schooling was available only for white children. On average, the African-American children who received no formal education during that period fell back at a rate of about six IQ points per year.

Fortunately, such episodes are rare. School attendance through the elementary grades is virtually universal in all modern industrialized countries. In the United States, for example, more than three-quarters of the population goes on to finish high school. This universality makes it difficult to separate the contributions of age and schooling to mental development, because the "average eight-year-old" is an eight-year-old who has been in school for two or three years. Some separation is possible, however, because admission to first grade in most school systems is governed by an arbitrary age cutoff, such as six years old by September 1 of a given year. This practice ensures that the children in any

given grade will vary in age by up to a year and that there will be children in different grades who are very nearly the same age.

In 1987, Sorel Cahan and Nora Cohen of the Hebrew University in Jerusalem took advantage of these birth-date distributions in an ingeniously designed study. They administered 12 different brief tests-presented together in a single session-to some 10,000 fourth-, fifth- and sixth-grade children in the Jerusalem schools. Using a complex statistical analysis based on birth dates and school admission, Cahan and Cohen compared the effects of a year of school (controlling for age) with those of a year of age (controlling for school) on each test separately. As one might expect, schooling mattered more than age for every test of verbal or numerical skills. More surprising, perhaps, is that schooling also made a contribution-albeit smaller-to performance on several nonverbal tests of abstract and visual reasoning. One of those nonverbal tests consisted entirely of items from the Raven.

Despite those data, schooling is not an altogether satisfactory explanation of the secular rise in test scores. For one thing, elementary school children tested with the WISC show gains comparable to those of adults who take the WAIS. Since elementary education was already universal in the 1930s, the WISC gains cannot be attributed to increased years of schooling. Another argument is based on further analysis of the Dutch Raven data. Flynn reports that grouping the subjects by educational level makes very little difference: The gains appear almost undiminished in each such group considered individually. Finally, the effects of schooling do not fit the overall pattern of test-score gains; schooling affects tests of content more than tests of reasoning, and the rise in test scores shows exactly the opposite pattern.

Child-Rearing Practices

When societies modernize, child-rearing practices change along with everything else. Parents everywhere are now interested in their children's intellectual development and are probably doing more to encourage it than they did in the past. We have no systematic data on that point, but we do know that modern volumes of child-care advice are available almost everywhere. For instance, the works of Dr. Spock have been translated into dozens of languages, and millions of children spend hours every day watching Sesame Street and other educational programs. A principal purpose of all that early stimulation is to raise children's overall intelligence. Is there reason to believe that it has done so?

The effect of early childhood experience on IQ scores has been a vexed question for many years. We now know that preschool (age 3-4) intervention programs like "Head Start" do not produce lasting changes in IQ, although they do confer other important benefits. It is possible, however, that more intensive interventions earlier in childhood would produce more substantial effects. For example, in the North Carolina "Abecedarian Project"-an all-day program that provided various forms of environmental enrichment to

57 children from infancy onward (mean starting age 4.4 months) and compared their test performance to a matched control group-differences between groups became apparent before the end of the first year. The difference did not diminish over time; the IQ difference between the groups was still present at age 12. Nevertheless, even this very intensive intervention only resulted in a gain of five IQ points, and not all such projects have been successful. Thus it seems unlikely that changes in the intellectual character of early childhood experience can explain a major share of the Flynn gains.

The Visual and Technical Environment

Child-rearing practices and modes of schooling are just instances of culture, and culture in a general sense has undergone enormous changes in all the "modern" countries where test scores have risen. Everything is changing: what people aspire to, whom they respect, how they live, how much they know about the world, what they do with their time, what skills they have acquired and how they treat other people and expect to be treated themselves. Given the scope of these changes, it is often suggested that modernization results in fundamentally different modes of thought. What might those new modes be? Although this question may be too vague to allow a general or comprehensive answer, one possibility leaps immediately to mind.

Perhaps the most striking 20th-century change in the human intellectual environment has come from the increase in exposure to many types of visual media. From pictures on the wall to movies to television to video games to computers, each successive generation has been exposed to far richer optical displays than the one before. People once regarded pictures as museum pieces or as occasional decorations for the homes of the rich; now they are everywhere, and everybody takes their own photographs. Schoolchildren of all ages devote far more time to visual "projects" today than they did a generation ago. (They devote correspondingly less time to the old "three Rs" of reading, writing and arithmetic, with the predictable consequence that skills in those domains have diminished.)

Beyond merely looking at pictures, we analyze them. Picture puzzles, mazes, exploded views and complex montages appear everywhere-on cereal boxes, on McDonald's wrappers, in the instructions for assembling toys and in books intended to help children pass the time. Even the answer sheets for standardized tests-often on pages separate from the questions-assume that the test-takers can locate the right places to record their responses. And static displays such as pictures and diagrams are only the beginning. We have had movies since the 1920s, television since the 1950s and video games since the 1970s. Patricia Greenfield of the University of California at Los Angeles argues that children exposed to these media develop specific skills of visual analysis, skills in which they routinely excel their elders. The assumption that children can program a VCR more effectively than their parents has become a cliche of American society, one that recognizes an important generational shift.

It is possible, then, that exposure to complex visual media has produced genuine increases in a significant form of intelligence. This hypothetical form of intelligence might be called "visual analysis." Tests such as Raven's may show the largest Flynn gains because they measure visual analysis rather directly; tests of learned content may show the smallest gains because they do not measure visual analysis at all.

Although little direct evidence exists for the visual-analysis hypothesis, it does offer the advantage of focusing our attention on the diversity of mental abilities. Flynn's argument that real intelligence cannot have gone up as much as scores on the Raven assumes that there is a "real intelligence"-some unitary quality of mind not unlike Spearman's g. Abandoning that assumption, we may think instead that different forms of intelligence are developed by different kinds of experience. The paradox then disappears: We are indeed very much smarter than our grandparents where visual analysis is concerned, but not with respect to other aspects of intelligence. This is hardly a final answer, but it may be a useful way of thinking about the worldwide rise in test scores.

Bibliography

· Brody, N. 1992. Intelligence. Second Edition. New York: Academic Press.
· Cahan, S., and N. Cohen. 1989. Age versus schooling effects on intelligence development. Child Development 60:1239-1249.
· Campbell, F. A., and C. T. Ramey. 1994. Effects of early intervention on intellectual and academic achievement: a follow-up study of children from low-income families. Child Development 65:684-698.
· Carpenter, P. A., M. A. Just and P. Shell. 1990. What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test. Psychological Review 97:404-431.
· Ceci, S. J. 1996. On Intelligence: A Bioecological Treatise on Intellectual Development. Expanded edition. Camidge, Mass: Harvard University Press.
· Cronbach, L. J. 1970. Essentials of Psychological Testing. New York: Harper and Row.
· Flynn, J. R. 1984. The mean IQ of Americans: massive gains. Psychological Bulletin 95:29-51.
· Flynn, J. R. 1987. Massive IQ gains in 14 nations: what IQ tests really measure. Psychological Bulletin 101:171-191.
· Flynn, J. R. 1994. IQ gains over time. In R. J. Sternberg (ed.). Encyclopedia of Human Intelligence. New York: MacMillan, pp. 617-623.
· Lynn, R. 1990. The role of nutrition in secular increases in intelligence. Personality and Individual Differences 11:273-285.
· Neisser, U., G. Boodoo, T. J. Bouchard, A. W. Boykin, N. ody, S. J. Ceci, D. F. Halpern, J. C. Loehlin, R. Perloff, R. J. Sternberg and S. Urbina. 1996. Intelligence: knowns and unknowns. American Psychologist 51:77-101.
· Schmidt, I. M., M. H. Jorgensen and K. F. Michaelsen. 1995. Height of conscripts in Europe: Is postneonatal mortality a predictor? Annals of Human Biology 22:57-67.

· Tuddenham, R. 1948. Soldier intelligence in World Wars I and II. American Psychologist 3:54-56.