MANKIND QUARTERLY Mankind Quarterly, (), 1-3 Edition
ISSN:

Editorial
It's All about Prediction

Gerhard Meisenberg

The ability to predict real-world outcomes is a defining characteristic of science. A scientific hypothesis not only explains observed phenomena parsimoniously. It must also be able to predict the probability of future events above chance, and the results of experiments designed to test the hypothesis. A hypothesis that cannot predict future outcomes and therefore cannot be tested by observation or experiment does not qualify as science. It is guesswork, belief, opinion, prejudice or superstition but not science.

Predicting the social and behavioural life outcomes of individuals is a central task of the social and behavioural sciences. To this end, psychologists have designed predictive tests whose purpose is not only the *description* of people's well-being, character or abilities, but also the *prediction* of their future. These tests are a kind of crystal ball in which the scientifically trained soothsayer can discover a person's secrets and predict the course of his future life — above chance, but of course not in absolute terms.

Of all the tests that psychologists have concocted, those intended to assess intelligence are more predictive and more useful than those assessing other psychological traits, in large part because on intelligence tests, unlike personality tests, people cannot "fake good". They can only fake bad, but are rarely motivated to do so. The scores on intelligence tests do indeed predict success in life well above chance. They can predict a child's learning in school, and to a limited but still important extent, intelligence measured early in life predicts adult earnings, creative achievement, probability of staying out of prison, health, longevity, and much more.

But what is it that these tests measure? Some psychometricians insist on the reality and importance of "general intelligence", also called the *g* factor (e.g., Jensen, 1998). They sometimes treat *g* as a real cognitive ability, the function of a neural circuitry somewhere in the convolutions of the brain that is required for all cognitive tasks. However, such a view is neurologically improbable. For all we know, the human brain has circuitries that are dedicated to specific functions such as categorization, analogical reasoning, language learning, the encoding, storage and retrieval of long-term memories, manipulation of the content of visual and verbal working memory, social skills, good judgement, and much more.

At its core, the *g* factor is merely a description of the fact that at the population level, there are positive correlations among the scores of all the subtests of which intelligence tests are composed: tests of categorisation, inductive and analogical reasoning, word knowledge, jigsaw puzzle ability, mental rotation, mechanical comprehension, and the like. Technically, the *g* factor is extracted from these scores by factor analysis. This method produces a kind of weighted average where the subtest scores that have the highest correlations with the other subtest scores — presumed to be the "purest" or most accurate measures of intelligence — are given the greatest weight.

It is therefore likely that *g* is not a general ability, but a statistical phenomenon caused by the fact that many and perhaps most of the conditions that boost one specialized ability also boost other specialized abilities to various extents (Conway & Kovacs, 2015). Going to school, living in an intellectually enriched environment, absence of nutritional deficiencies in infancy and childhood, and having a low load of genetic mutations that impair brain development and function, all these are non-specific influences that can cause positive correlations among test scores because they all enhance multiple cognitive brain functions. Although specialized abilities contribute to real-world outcomes beyond *g* (Coyle, 2018), the *g* factor is nevertheless the most predictive metric of IQ tests because intelligent responses to most of the challenges and opportunities that people encounter in life require the coordinated functioning of many specialized abilities.

This is the background on which we can try to explain group differences on cognitive tests. Test score differences between groups, for example between males and females, schoolchildren and pensioners, ethnic and racial groups and the populations of different countries, are universal but can occur at three levels. The first possibility is a difference in scores on the *g*-factor, meaning that one group scores higher than the other on most or all subtests. It indicates that one or more of those factors that have non-specific effects on brain function, such as genes, nutrition and schooling, are different between the groups. The second possibility is that differences on subtest scores are larger than those on *g*. This indicates that the groups are not much different on broad-band thinking and learning ability, but that the differences are in specialized abilities. Neither of these two kinds of group difference implies that the test is "biased". Group differences are expected, and the ultimate criterion of test bias is predictive power. As long as a learning ability test predicts learning in school and a mathematics ability test predicts success in a mathematics study program equally for everyone, independent of group membership, the test is not biased no matter how big the group differences are.

A third possibility is that group differences are specific to individual test items, technically known as differential item functioning (DIF). Does this mean that those items showing the group differences are "biased", that it would be unfair to use them on a test? Again, the ultimate criterion is, in theory, the correlation of item response with the real-world outcomes that the test is intended to predict. Because this would be impractical, psychometricians use a short-cut: They determine whether the probability of answering an item correctly is predicted by the score on the (sub)test alone — assumed to be an approximation of the "latent ability" construct the test is intended to measure —, or also by group membership independent of this hypothetical latent ability. If group membership has an additional effect, the item measures not only the latent ability construct, but also something extraneous, something else that is different between the different groups.

In this issue of *Mankind Quarterly*, Ismael Albursan and his colleagues in Saudi Arabia and Sudan are looking at possible sources of differences between boys and girls in the items of Raven's Coloured Progressive Matrices (CPM), an intelligence test for children. Like the adult versions of the Raven test, the CPM is intended to measure core aspects of "general" intelligence such as pattern recognition, inductive reasoning, and abstraction. It is not intended to measure specialized abilities. Unlike the Wechsler tests for example, it is not composed of multiple subtests, each measuring a different cognitive ability or skill. Studies so far have shown that sex differences on the total score are minimal, interpreted as minimal sex differences in "general intelligence".

The Albursan et al. study is unique in many respects. It used an impressive sample size of 2624 children aged 4 to 12 years, more than in comparable studies elsewhere. The sample is from preschools and schools in the Khartoum area of Sudan, a place where studies of this kind have never been done before. Finally, they describe sex-related differential item functioning not only for the entire group, but also relate such item specificity to the child's age. Indeed, their main result is that the number of DIF items, small in younger children, increases substantially in the oldest age group.

The authors refrain from speculation about the possible causes of such item-specificity. Familiarity with the test items is an improbable explanation. However, we know from earlier studies that although sex differences on across-the-board intelligence tests are very small, there are consistent differences in specialized abilities such as processing speed (favouring females) and spatial relations (favouring males). Perhaps the most parsimonious explanation for differential item functioning is that the Raven test measures not a single general ability, but two or more specialized cognitive skills, at least one on which males do better and another one on which females do better. Some items tap into an ability on which females are better, and others tap into an ability on which males are better.

And why should this differentiation by sex be greater in children aged 10 to 12 than in those aged 4 to 9? Hormones are an easy guess. The result suggests that testosterone, estradiol and other sex steroids promote not only sexual but also cognitive maturation at puberty in a sexually dimorphic way, thereby enhancing cognitive sex differences that are still minimal in younger children. Or is it a cultural effect whereby the specific environment to which children are exposed in Sudan stimulates some cognitive skills more strongly in boys and others more strongly in girls?

What needs to be done now is to assess the cognitive processes that are required to answer the items on the Raven tests, abilities such as holistic perception of visual patterns, analytical ability to abstract distinctive properties from a visual image, or verbal strategies. We also need cross-cultural comparisons to see which ones of the sex differences are universal and therefore in some way part of "human nature", and which ones are culture-bound. As far as test development is concerned, should all items with sex-related differential item functioning be removed? Probably not, if they predict the outcomes that the test is intended to predict and if male-favouring and female-favouring items are about equally matched on the test. What ultimately counts is the test's, not the item's, ability to predict real-world outcomes such as school grades independent of sex. It's all about prediction.

References

Conway, A. R., & Kovacs, K. (2015). New and emerging models of human intelligence. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(5), 419-426.

Coyle, T. R. (2018). Non-g factors predict educational and occupational criteria: More than *g*. *Journal of Intelligence*, *6*(3), 43.

Jensen, A. (1998). *The g Factor: The Science of Mental Ability*. Westport CT: Praeger.