Error in the measurement of low IQ: implications for research, clinical practice and diagnosis

By

Simon Whitaker

Word count: 2737

Summary

The paper considers the effect of error in the measurement of low IQ for research, clinical practice and diagnosis. Test error has most impact on diagnosis of intellectual disability (ID) in an individual.

Introduction

Recent concerns have been expressed in the literature about the accuracy of IQ tests when used in the low range (Flynn 2007, 2012, Gordon et al 2010, Whitaker 2008, 2010, 2013). This paper expands on this work in considering the differential impact of these errors for research, clinical practice and diagnosis.

Chance and systematic errors

It is usual to divide the factors that result in error in IQ tests into chance and systematic. Chance errors are due mainly to minor variables acting randomly that cause IQ scores to be either higher or lower by a small amount. In the main they result from a lack of internal consistency and lack of stability of the tests. The degree of internal consistency is calculated by split-half or coefficient alpha reliabilities, and the stability by the test re-test reliability. How much these errors affect scores is traditionally represented by the 95% confidence interval, which is the range of scores around the measured IQ, where the true IQ is thought to have a 95% chance of lying. Although the 95% confidence interval for the WISC-IV and WAIS-IV is usually cited as being about eight to ten points, from the high score to the low score (Wechsler 2004, Wechsler et al 2008), it has been pointed out (Whitaker 2008, 2010, 2013) that this estimate largely ignores the lack of stability of the test and is based on data from individuals with average intellectual abilities. He suggests that a better estimate of the 95% confidence interval in the low range should use data taken from studies using individuals with low intellectual ability and include errors due to both a lack of internal consistency and a lack of stability. When this is done it gives a 95% confidence interval of about 26 points.

Systematic errors are thought to be due to a smaller number of non-random factors that affect specific tests and result in one test scoring, on average, higher or lower than other tests. We know about some of these factors, such as the Flynn effect (Flynn 2007) and the floor effect (Whitaker and Gordon 2012), which could possibly could be corrected for, but there are others that are not understood. Although in the average IQ range the differences between tests are only about two or three points, in the low range the differences are much more significant. For example, Gordon et al (2010) found the WISC-IV to measure 12 points lower than the WAIS-III with 16-year-olds, Silverman et al (2010) found Stanford Binet tests to measure 17 points lower than the WAIS tests with adults and Grondhuis and Mulick (2013) found the Stanford Binet Five measured 22 points lower than the Leiter-R with autistic children.

Measured IQ vs. True Intellectual Ability

Because of these errors a distinction will be made here between measured IQ (the score that would be obtained on current IQ tests in the particular circumstances in which it was used) and true intellectual ability (the score that would be obtained by a perfectly standardized IQ test, measuring to an accuracy of one point, given under ideal conditions).

The use of tests with groups vs. with individuals

In looking at the differential effect of these errors it is important to understand that their effects are different when applied to groups as opposed to individuals.

--------------- Put Table 1 about here ---------------

Groups

With a group, the chance errors will tend to cancel each other out so that the mean score is only affected by a relatively small amount. Table 1 illustrates this. It uses dummy data for 10 randomly chosen measured IQs between 45 and 70 and a pattern of corresponding true intellectual abilities that could occur with a 95% confidence interval of 26 points, however, assumes there is no systematic error. Although the average difference between the scores is 4.6, the difference between the means was only 0.2 of a point. So averaging measured IQs reduces chance error and the larger the group on which this average is based, the smaller the effect of chance error. But Table 1 assumes no systematic error, which would not actually occur and would not be reduced by averaging measured IQ scores.

Individuals

It can be seen from Table 1 that, although the difference between the means is trivial (0.2) the disparity between the score for some individuals is not. The mean difference is 4.6 points and two subjects differed by 9 points or more. So only taking into account chance error, measured IQ is not a good estimate of true intellectual ability. However, there is also systematic error, which also reduces the accuracy of a test.

Whitaker (2010, 2013) had calculated the effective 95% confidence intervals if both chance and systematic error are taken into account, which are slightly different for both the WISC-IV and the WAIS-III: For the WISC-IV the effective 95% confidence extends from 14 points below the measured IQ to 23 points above it and for the WAIS-III it was 16 points above measured IQ to 26 points below it. Although the WAIS-IV has not been examined to the same extent as the WAIS-III with regard to how it compares with the WISC-IV at low IQ levels, Whitaker (2012) has suggests that there are likely to be the same inaccuracies with the WAIS-IV as there are with the WAIS-III.

Even though some correction can be made to scores to compensate for these errors we don't know how accurate these corrections are and a lot of error cannot be corrected for (Whitaker 2013). So a measured IQ could vary from true intellectual ability by the order of 20 points.

Test specific IQ

A distinction is made above between true intellectual ability, which is currently not measurable, and measured IQ, which is measurable but is subject to chance and systematic error. A further distinction that can be made is that between measured IQ, which implies any test, and measured IQ on a specified test. Specifying a test and then referencing all scores against what is known about the psychometrics of that test would eliminate the effect of systematic error. If it is known how much other tests differ from the specified test then their scores could be adjusted so that they are equivalent to the specified test. There would seem to be circumstances where it is reasonable to do this, for example if a psychologist is operating in a service mainly for adults, it may be reasonable to mainly use the WAIS-IV and reference all scores against WAIS-IV scores. If there are historical WISC scores then these could be adjusted by adding the appropriate number of points, approximately 10, to make them equivalent to WAIS-IV scores.

Impact of chance and systematic errors

Research

In research IQ is used as both a descriptor of individuals and as a dependent and at times as an independent variable (Laird and Whitaker 2011).

Intelligences as a descriptor

IQ or some other measure of intellectual ability, such as mental age, is commonly used as a descriptor of subjects with ID in the research literature (Laird and Whitaker 2011). The accuracy of this description will depend on whether the test used is specified, whether there is a single test or more than one used and whether a group or an individual is being described. The greatest effect will be on an individual case study where the measured IQ will be subject to most errors. If there is a group of subjects and different IQ tests are used the chance errors will be reduced, but systematic errors will remain and one will not be able to talk about a test specific IQ. The most accurate description would be for a group of individuals who have been assessed on the same IQ test and where a mean IQ is reported, which would be an accurate test specific mean IQ.

Intelligence as an experimental variable

If the same test is used with a group it would be reasonable to use IQ as dependent variable. The chance error would have little effect, systemataic error would have no effect on mean test specific IQ scores, and statistical tests would be able to show whether the difference in scores between two or more groups was statistically significant. However, with individuals it would be much more difficult to show significant change even when the same test is used. A significant result would be one that was greater than that which could be reasonably expected to occur by chance. Whitaker (2008) found that the 95% confidence interval for test re-test reliability in the low range was 12.5 points, which, as IQ is measured to a whole point, would require that there would need to be a 13 point change in IQ score for it to be significant at the 5% level (two tailed) or 11 points (one tailed).

The obvious consequence of these errors for research is that intellectual abilities of subjects are wrongly assessed leading to the wrong conclusion being drawn. An example of a study where a failure to appreciate that different IQ tests do not agree with each other in the low range is cited by Laird and Whitaker (2011). Russell et al (1997) investigated whether schizophrenia reduced IQ. They compared the IQs of adults who had developed schizophrenia with their IQs as children before developing schizophrenia. However, as children they were mainly assessed on the WISC-R and as adults on the WAIS-R. They reported the mean WISC-R IQ to be 84.1 and the mean WAIS-R IQ to be 82.2 and concluded that schizophrenia did not result in a significant reduction in IQ. However, they failed to consider that WISC-R might systematically measure lower than the WAIS-R by about 10 points at these IQ, which it is likely to do. Spitz (1989) found that for IQs in the 60s (on the WAIS-R) the WAIS-R measured 15 points lower than the WISC-R, though the effect was less at higher IQs, therefore a difference between the two tests of 10 points at these higher IQ levels seems a reasonable estimate. So if these 10 points are added to the WISC-R score to make it equivalent to a WAIS-R score, the WAIS-R equivalent IQ as children would be 94.1, just short of 12 points greater than the adult WAIS-R measured IQ of 82.2 and good evidence that schizophrenia does reduce intellectual ability.

Clinical use

The clinical use of IQ tests is predominantly with individuals and usually with a specified IQ test. So a psychologist working primarily with children with ID may mainly use the WISC-IV and would be able to compare individual WISC-IV scores against research evidence about the psychometrics of the WISC-IV in the low range. However, even when using a single specified test one should be very careful about how much weight one gives to IQ scores in making clinical decisions. For example, one clinical use of an IQ assessment is to find an individual's cognitive strengths and weaknesses in terms of differences between index scores and between subtest scores. However, using the data from the comparison between the WISC-IV and WAIS-III on 16-year olds in special education, Whitaker and Gordon (2009) calculated the strengths and deficits profile for each individual on both tests and found very little agreement between the profiles on the different tests. Whether this lack of agreement was due to a difference between the WISC-IV and WAIS-III, a lack of stability of the subtest score or index scores, or something else, is not yet clear, however, the result must cast doubt on how valid it is to use such analysis in the low range. Therefore a clear possible consequence of the clinical use of IQ tests is that the wrong inference could be drawn about the capabilities of an individual from a measured IQ score.

Diagnosis

In effect both ICD-10 and DSM-5 specify an IQ cut-off point of 70 or 75[1] as a necessary criterion for a diagnosis without specifying a test that should be used. This clearly implies that they are referring to true intellectual ability rather than a test specific IQ. Therefore an assessment done in order to see if an individual's IQ is above or below a cut-off point will be subject to both chance and systematic error, only some of which could be corrected for (Whitaker 2013). A large proportion of diagnoses are therefore likely to be wrong (Whitaker in press). The consequences of making a wrong diagnosis can also be much greater than the consequences of test error in research or the clinical use of IQ assessments. If an individual has a measured IQ above 70 yet has a true intellectual ability below 70 then there is a danger they will not be given a diagnosis of ID. This could result in them not getting the service they need to be able to cope or even not been spared the death penalty if they had been convicted of a capital crime in the USA (Whitaker 2013). If, on the other hand, they have a true intellectual ability above 70 but a measured IQ below there is a danger that they are given a diagnosis, which they may well find stigmatizing and wish to avoid.

[1] Both state IQ 70. DMS-V specifies a margin of error of 5 points and ICD-10 says it's an approximate IQ but convention would suggest that they also imply a 5 point margin of error. This gives an effective IQ cut-off point of 75.

References

Flynn, J.R. (2007). What is Intelligence: Beyond the Flynn Effect. Cambridge: Cambridge University Press.

Flynn, J.R. (2012). Are We Getting Smarter? Rising IQ in the Twenty-First Century. Cambridge: Cambridge University Press.

Gordon, S., Duff, S. Davison, T and Whitaker, S. (2010). Comparison of the WAIS-III and WISC-IV in 16 year old special education students. Journal of Applied Research in Intellectual Disability, 23, 197-200.

Grondhuis, S.N. & Mulick, J.A. (2013). Comparison of the Leiter International Performance Scale—Revised and the Stanford-Binet Intelligence Scales, 5th Edition, in Children with Autism Spectrum Disorders. American Journal on Intellectual and Developmental Disabilities: January 2013, Vol. 118, No. 1, pp. 44-54.

Laird, C. and Whitaker, S. (2011). The use of IQ and descriptions of people with intellectual disabilities in the scientific literature. British journal of developmental disabilities, 57, 175-183.

Russell, A.J., Munro, J.C., Jones, P.B., Hemsley, D.R., and Murray, R.M. (1997). Schizophrenia and the myth of intellectual decline. American Journal of Psychiatry, 154, 635-639.

Silverman, W., Miezejeski, C., Ryan, R., Zigman, W., Krinsky-McHale, S & Urv, T. (2010). Standford-Binet and WAIS IQ differences and their implications for adults with intellectual disability (aka mental retardation). Intelligence, 38, 242-248

Spitz, H.H. (1989). Variations in the Wechsler interscale IQ disparities at different levels of IQ. Intelligence, 13, 157-167

Wechsler, D. (2004). Wechsler Intelligence Scale for Children – Fourth UK Edition: Administrative and Scoring Manual. London: The Psychological Corporation.

Wechsler, D. Coalson, D. L. & Raiford, S. E. (2008). WAIS-IV Technical and Interpretive Manual San Antonio, Texas: Pearson.

Whitaker, S. (2008). The stability of IQ in people with low intellectual ability: An analysis of the literature. Intellectual and Developmental Disabilities, 46, 120-128.

Whitaker, S. (2010). Error in the estimation of intellectual ability in the low range using the WISC-IV and WAIS-III. Personality and Individual Differences, 48, 517–521.

Whitaker, S. (2010). The measurement of low IQ with the WAIS-IV: a critical review. Clinical Psychology Forum, 231, 45-48.

Whitaker, S. (2013). Intellectual Disability: An Inability to Cope with an Intellectually Demanding World. London: Palgrave McMillan.

Whitaker, S. (in press). How accurate are modern IQ tests at categorising people as ID or non ID? Clinical Psychology Forum.

Whitaker, S. and Gordon, S. (2009). Profile Analysis on the WISC-IV and WAIS-III in the low intellectual range: Is it valid and reliable? Clinical Psychology and People with Learning Disabilities, 7, 35-38.

Table 1

Subject	Measured IQ	True Intellectual Ability	Difference Between Scores
1	60	70	10
2	55	61	6
3	63	67	4
4	58	61	3
5	69	70	1
6	54	53	1
7	50	48	2
8	62	59	3
9	67	60	7
10	56	47	9
Mean	**59.4**	**59.6**	**4.6**