Error in the estimation of intellectual ability in the low range using the WISC-IV and WAIS-III

By

Simon Whitaker

In press Personality and Individual Differences

Abstract

The error, both chance and systematic, in the measure of true intellectual ability in the low IQ range is quantified and combined to find an overall confidence interval. The chance error was due to: lack of stability, scorer error and lack of internal consistency. The systematic error was due to: the Flynn effect, a floor effect and that error apparent from the lack of agreement between the WISC-IV and WAIS-III. For low Full Scale IQs the WAIS-III can only be considered accurate to within 18 points above the measured IQ and 28 points below, and the WISC-IV to 16 points below the measured IQ and 25 points above it. The implications for the diagnosis of intellectual disability are considered.

Key word: WISC-IV, WAIS-III, Low IQ, Test error.

Introduction

A diagnosis of intellectual disability (ID) or as it was previously known, mental retardation (MR), can have a major effect on people's lives. On the positive side it can provide services, finance, help in schools, and even prevent the recipient from being executed (Flynn 2006; Flynn 2007; Schalock et al 2007). On the negative side it may be regarded as a stigmatizing label that an individual may seek to avoid (Baroff 1999).

Currently, a necessary though not sufficient part of the diagnosis of ID is having an IQ below a specified figure, usually 70, or two standard deviations (SDs) below the norm (American Association on Mental Retardation 2002; American Psychiatric Association 2000; Department of Health 2001; British Psychological Society 2001). This specification of a specific IQ figure implies that an individual has a "true intellectual ability" that can be measured and quantified in terms of an IQ score. True intellectual ability can be defined as the IQ score that an individual would obtain if he/she was assessed using a perfectly standardized IQ test with no measurement error.

It is acknowledged that current tests do not measure IQ to a level of accuracy of one point: there is a margin of error, usually considered to be about five points either side of the obtained IQ, which should be taken into account when making a diagnosis of ID (The American Association on Mental Retardation 2002). However, Whitaker (2003, 2008a) has suggested that the margin of error in the low IQ range is much greater than five points and is indeed so large that it is unreasonable to have a specific IQ figure as part of the diagnostic criteria of ID.

The purpose of this paper is to quantify the various sources of error in the measurement of low intellectual ability in order to derive a margin of error of the assessment of true intellectual ability using current IQ tests. Most of the examples given will relate specifically to the measurement of Full Scale IQ (FS IQ) on the Wechsler assessments, the Wechsler Adult Intelligence Scale third edition (WAIS-III Wechsler 1997) and the Wechsler Intelligence Scale for Children fourth edition (WISC-IV Wechsler 2003). This is because the Wechsler assessments are the most widely used assessments of low intellectual ability and the bulk of research has been done using these assessments. However, the same arguments apply equally well to other assessments.

Errors in the measurement of IQ

Error in the measurement of IQ is due to non-intellectual variables that affect an IQ score and can be considered to be of two broad types: chance errors and systematic errors.

Chance Error in the Measurement of IQ

Chance errors are usually due to a large number of relatively small factors that may occur during an assessment. According to Anastasi and Urbina (1997) three broad types can be detected and quantified.

Firstly, there is error derived from a lack of internal consistency. This is the degree to

which items on a subtest are measuring the same psychological factor. The degree of error due to a lack of internal consistency is derived by subtracting the split-half reliability coefficient from one.

A second type of chance error is temporal error due to variation in the conditions under which assessments are administered. For example, factors such as the level of distraction, how the client felt on the day and the way the assessment was administered will all vary to some extent between assessments and will all affect the score. An estimate of temporal error is given by a test re-test reliability coefficient, or stability coefficient, obtained by correlating the scores of the same individuals when given the same assessment on two occasions. The degree of error is then obtained by subtracting this test re-test reliability coefficient from one.

Thirdly, there is scorer error, which is due to inconsistency in scoring the assessment. An estimate of this is obtained by correlating the results of two independent scorers scoring the same assessments and subtracting this correlation coefficient from one.

According to Anastasi and Urbina (1997), as well as other theorists in psychometrics such as Crombach, Gleser, Nanda and Rajaratnam (1972) and Shavelson and Webb (1991), errors due to a lack of internal consistency, temporal changes and scorer error are mutually exclusive. Therefore, in order to get an estimate of the total error in the assessment of true intellectual ability all three errors should be summed.

Standard Error of Measurement and 95% confidence interval

An estimate of error in an IQ assessment is usually given by the standard error of measurement (SEM), and the 95% confidence interval. SEM is the theoretical SD of test scores that would be expected to occur if the tests were repeatedly given to the same client or if different combinations of possible test items were used. Anastasi and Urbina (1997) give the following formula to calculate SEM:

$$SEM = SD \sqrt{(1-r)}$$

where SD is the standard deviation of the test, which for the Wechsler assessments is 15 and r is a reliability coefficient. The 95% confidence interval is the interval around the measured IQ in which there is a 95% probability that the true IQ falls. It is calculated by multiplying SEM by 1.96 and then adding and subtracting the resulting figure from the obtained IQ score to get the upper and lower limits of the interval.

Values for both SEM and the 95% confidence interval given in the manuals may well be low estimates of the true confidence intervals in the assessment when it is used on a clinical population of people with low intellectual ability. There are a number of reasons for supposing this.

First, the figures in manuals were obtained using the standardization sample in which

subjects were mainly in the normal range of intellectual ability and so may not necessarily apply to individuals with low intellectual ability (c.f. Anastasi and Urbina 1997). Secondly, the assessments of the standardized sample would have been done under near optimal conditions. The subjects would have been well motivated and in good health, distraction would have been at a minimum and assessors would have been well trained. These conditions would not always occur when the assessments are used in clinical practice and so would be subject to more error than was apparent when the test was standardized. Thirdly, the SEM and 95% confidence interval are usually calculated using a single reliability figure and so do not take into account all sources of chance error.

Chance error on the Wechsler assessments

The reliability used to calculate the SEM and 95% confidence intervals for individual subtests in the WAIS-III and the WISC-IV, with the exception of Symbol Search, Coding (Digit Symbol) and Cancellation, is the split-half reliability. As Symbol Search, Coding and Cancellation are speed tests it is not possible to get a split-half reliability and test re-test reliability is used instead. The reliability value used to calculate the SEM and 95% confidence interval for IQs and Index scores is an aggregation of the reliability figures for subtests used to measure the IQ or index score. Therefore the SEM and 95% confidence intervals for IQ and Index scores are mainly based on the split-half reliability of subtests. This is the case for nine out of the 10 subtests used to measure FS IQ on the WISC-IV, and 10 of the 11 subtests used to derive FS IQ on the WAIS-III. However, split-half reliability only accounts for error due to a lack of internal consistency. It does not give any indication of error due to temporal changes or scorer error. Therefore, if measured IQ is affected significantly by changes in the conditions under which the assessment is given or by inconsistencies in scoring, the confidence interval given in the manual will be a significant underestimate.

The degree to which measured IQ varies over time is indicated by the stability coefficient which is the test re-test reliability score. A recent meta analysis (Whitaker 2008b) of the stability of low IQ (< 80) found a weighted mean stability coefficient of .82 for FS IQ for an average re-test interval of 2.8 years. It is possible that some of this variation in scores over the time was due to genuine changes in intellectual ability and not due to measurement error, however, as there was no statistically significant relationship between inter-test intervals and stability coefficients of the individual studies in the meta analysis, it is likely that little of the variance was due to change in actual intellectual ability. However, it is also possible that being a mean figure it may well not be representative of all groups.

If SEM is calculated on the basis of a stability figure it can be used to give an indication of proportion of the IQs that would change by particular amounts on the second testing as the SEM is the SD of variation between the two assessments. For example, one would expect that 32% would vary by more than 1 SEM. The SEM based on the above mean stability figure is 6.4, which corresponds to a 95% confidence interval 13 points either side of the measured FS IQ. Whitaker (2008b) notes that this figure would seem to be reasonably accurate as it predicts the proportion of FS IQs that change by specific

amounts quite well. It would be expected that for a SEM of 6.4, sixty one percent of FS IQs would change by less than six points and 13% would change by 10 points or more. It was found that in those studies that reported on the proportion of IQs that changed by specific amounts, 57% changed by less than six points and 14% changed by 10 points or more.

The above confidence interval, based on error due to a lack of stability, may well include scorer error if the assessments were scored by different people on the two occasions they were given, but will not take into account the error due to a lack of internal consistency. Therefore an estimate of the full chance error will be given by adding together the error due to lack of stability and the error due to lack of internal consistency. The author is aware of only one study that reported split-half reliability on a group of children with low IQs. Davis (1966) found split-half reliabilities of .90 for children with moderate ID (mean IQ 48) and .97 for those with borderline ID (mean IQ 76), the weighted mean reliability being .92, suggesting that error due to a lack of internal consistency was .08. If this figure is then added to error due to a lack of stability of .18 the total error comes to .26. Subtracting .26 from one it will give an affective reliability figure of .74 accounting for all chance error. When this reliability figure is used to calculate the SEM and 95% confidence interval it gives an SEM of 7.65 and a 95% confidence interval of 15 points either side of the obtained IQ.

Systematic Error in the Measurement of True intellectual ability

There are at least three major sources of systematic error when measuring low IQ. The first is due to an unacknowledged floor effect. The second derives from the Flynn Effect, whereby the intellectual ability of the population as a whole is increasing systematically over the years, resulting in IQ tests gradually becoming out of date. The third is apparent from different IQ tests systematically scoring either higher or lower than other tests.

A floor effect

In order to calculate IQ, the raw scores on subtest are converted into normalised scaled scores with a mean of 10, an SD of 3 and a range between 1 and 19. Whitaker (2005) and Whitaker and Wood (2008), have suggested that as this conversion can allocate a scaled score of one to low raw scores and raw scores of zero, it will result in some clients being credited with a greater ability than they have. This can be illustrated by looking at the example from the raw score to scaled score conversion tables in the WISC-IV (UK) Administrative Manual (Wechsler 2004) of the Digit Span subtest for age groups 16:00 to 16:30.

Scaled Score:	10	9	8	7	6	5	4	3	2	1
Raw Score:	18	17	16	15	14	13	12	11	10	0-9

There is a linear relationship between raw scores and scaled scores from raw score 18 to raw score 10, with a reduction in a raw score by one corresponding to a reduction in a scaled score by one. All raw scores of nine and less are then given a scaled score of one. However, there is no empirical reason to suppose that all raw scores below nine are equivalent to a scaled score of one, and logic suggests that the linear relationship

between scaled scores and raw scores should continue for some way below raw score nine. This means that there should be scaled scores of zero and less. It is therefore likely that some clients who gain low raw scores will have their ability overestimated by the allocation of a scaled score of one. This will obviously affect IQs in the 40s where scaled scores of one are inevitable. The degree to which it also affects IQs in the 50s, 60s and 70s was investigated by Whitaker and Wood (2008), who plotted the distribution of scaled scores in all the WISC-III (UK) and WAIS-III (UK) that had been given as part of clinical practice by the psychology services provided for people with low intellectual ability. It was found that the distribution of scaled scores for the WAIS-III (UK) appeared approximately normal with very few scaled scores of one, suggesting that the floor effect would only be a potential problem for IQs in the 40s and 50s. However, with the WISC-III (UK) there was a skewed distribution of scaled scores with more scaled scores of one than any other scaled score. Scaled scores of one were found at all IQ levels up to those in the 70s where they accounted for 10% of the scaled scores. There is therefore a distinct possibility that IQ scores are increased at low ability level due to a floor effect. The current evidence suggests that this is far more of a problem for the WISC-III (UK) than the WAIS-III (UK).

The Flynn Effect

Flynn (1984) found that the longer it was since an IQ test was standardized the higher the IQ, the rate of increase being about three points a decade. The implication is that as tests go out of date they will over-score an individual's true IQ by about 0.3 of a point for every year since they were standardized. There are two sources of evidence suggesting the effect also applies to people with low intellectual ability. First, there are studies in which people with low IQs have been given both a late and earlier edition of the same IQ test. Flynn (1985) looked at comparisons of the Wechsler Intelligence Scale for Children (WISC) and its revised version (WISC-R) standardized 25 years later. He found that the gains appeared to be higher at the low levels, 0.396 per year for IQs 55 to 70 as compared to 0.272 per year for IQs in the range 125-140. In a more up to date review (Flynn 2006), he suggests that low IQs are still increasing by about .3 of a point per year in the US.

Secondly, there is evidence from the assessment of military conscripts that the intellectual ability of the bottom of the ability range has been increasing at a higher rate than in the top of the range. In Norway, military service is compulsory for every able young man. As part of their induction process they are given an IQ test. This provides an opportunity to study what amounts to half the population of 18-year-olds. Sundet, Barlaug and Torjussen (2004) used this data to compare the gains made for conscripts scoring above and below the median for pooled data from 1957 to 1959 with data from 1993 to 2002. For those scoring below the median there was an 11 point IQ point gain, which compared to a 4.4 point gain for those above the median. Teasdale and Owen (1989) used similar data from Denmark and found average gains in IQ over the 30 years up to the late 1980s of about 7.5 IQ points. The gains were greatest in the lower ability range. The maximum gains were near the 11th percentile, at which point they were 41% greater than those at the median. At the 90th percentile there had been very little gain over the years. However, when Teasdale and Owen (2005) looked at the new data up to

2004 they found that there had been a peak in intellectual ability for the population as a whole in 1998 followed by a decline. They also report that after 1995 there was an increased number of people scoring at the lower end of the tests showing a decline in the intellectual ability for conscripts with lower IQ.

It therefore seems that the Flynn Effect occurred at a higher rate for people with low IQs than for the rest of the population in the past. However, there is evidence suggesting that these gains may have now stopped or even gone into reverse in some parts of the world. It would therefore be an act of faith to assume that low IQ will continue to increase at 0.3 of a point per year or that subtracting 0.3 of a point for each year since the test was standardized would compensate for the effect.

Differences between IQ scales

If different IQ tests systematically measured either higher or lower than other tests it would raise the question as to which IQ test was providing the best estimate of an individual's true intellectual ability. In the absence of a test that clearly is an accurate measure of true intellectual ability, the best that could be done would be to decide which of the many IQ tests is likely to be the most accurate and take that as the "gold standard" assessment against which other assessments should be compared. The Wechsler assessments should have a good claim to be regarded as the gold standard assessments. They have evolved over 70 years since the Wechsler Bellevue was first published in 1939 (Wechsler 1939), are apparently well standardized and are probably the most widely used tests of child and adult intelligence. However, there may be a major lack of agreement between the WISC and the WAIS in the lower IQ ranges. Both Flynn (1985) and Spitz (1986; 1989) reported the Wechsler Intelligence Scale for Children – Revised (WISC-R Wechsler 1974) gives IQ scores up to 15 points lower than the Wechsler Adult Intelligence Scale – Revised (WAIS-R Wechsler 1981) for IQs of 70 and below. Recent work (Gordon 2007, Gordon, Duff, Davison and Whitaker in press) has compared the latest UK standardizations of these assessments, the WISC-IV (UK) and the WAIS-III (UK), on a group of 16 year-olds receiving special education. It was found that, although there was a high correlation between the two assessments (r=.93), in each case the FS IQ on the WISC-IV (UK) was less than that on the WAIS-III (UK); a mean FS IQ of 53.00 was found on the WISC-IV (UK) which compared to a mean of 64.82 on the WAIS-III (UK), a difference of just less than 12 points. It is therefore clear that either one or both of these assessments are failing to produce an accurate measure of an individual's true IQ. As the degree to which either assessment is in error is not known, it is possible that either the WISC-IV (UK) is systematically underestimating true IQ by up to 12 points, or the WAIS-III (UK) is systematically overestimating true IQ by 12 points or both assessments are making systematic errors of less than 12 points.

Discussion

The purpose of this paper was to quantify the various sources of error, both chance and systematic, in the measurement of low IQ in order to get an estimate of the degree of accuracy to which true intellectual ability can be measured in the low range. Some of these errors are more easily quantified than others and combining error from various sources can only be done by making assumptions. Therefore any estimate of the overall

degree of accuracy with which true intellectual ability can be measured must be regarded as tentative, nonetheless it is hoped that doing this will be informative.

The overall chance error was estimated at 0.26, corresponding to an effective reliability figure of .74 which corresponds to a SEM of 7.65 and a 95% confidence interval 15 points either side of the obtained IQ. Added to this there are various sources of systematic error, if they cannot be corrected for.

The degree to which the floor effect could increase an IQ or index score will be dependent upon the number of subtests gaining a scaled score of one and the degree to which that overestimates the client's ability. An increase in the sum of scaled scores by one results in an average increase in FS IQ on the WISC-IV of 0.6 points, so if there were a floor effect on several subtests there could be an increased FS IQ of several points. Given that scaled scores of one are far more common for IQs in the 40s and 50s, the floor effect is going to mainly boost measured IQs in the 40s and 50s. However, as Whitaker and Wood (2008) report that 10% of the scaled scores are one on the WISC-III for IQs in the 70s, it is likely that IQs in the 60s and 70s on the WISC-IV will be subject to a floor effect of about one IQ point.

If the raw scores are known it may be possible to correct for the floor effect by extrapolating the relationship between raw scores and scaled scores down below scaled score one and deriving a lower scaled score which could be then used to calculate the sum of scaled scores. However, this is assuming the relationship between raw scores and scaled scores in the table continues below scaled score one; if this is not the case an extrapolation would lead to further error.

If it is assumed, as Flynn (2006) suggests, that low intellectual ability is increasing at .3 of a point per year, then the Flynn effect could be corrected by subtracting .3 of an IQ point from a FS IQ for every year that has elapsed since the test was standardized. However, this assumption may not be true. There is evidence that, particularly in Scandinavia, the rate of gain in intellectual ability may have slowed or gone into reverse. Therefore although it is likely that tests are getting less accurate as they go out of date it is not clear if this then leads them to either under or overestimate IQ. If it is assumed that the error is of the order of 0.3 of a point per year then currently the WISC-IV, which was standardized seven years ago, could be in error by one or two points and the WAIS-III, standardized 13 years ago, by about three points. Although this is a systematic error, as we do not know how large the gain is or even if it is positive or negative, for this analysis it will be treated as an additional chance error. It will therefore add an additional one point to the effective 95% confidence interval of the WISC-IV bringing it up to 16 points, and an additional three points to the effective 95% confidence interval of the WAIS-III bringing it up to 18 points.

Although the current evidence suggests that the WISC-IV scores 12 points less than the WAIS-III, some of this difference may be accounted for by the Flynn Effect although, as argued above, it is not known to what degree. However, if Flynn's estimate of 0.3 of a point increase in IQ is assumed to be correct, given that the WISC-IV was standardized 6

years after the WAIS-III it would be expected to score two points less than the WAIS-III. Therefore there is a possible 10 point difference between the two assessments in the low IQ range due to factors other than the Flynn Effect. So any FS IQ on the WISC-IV may be up to 10 points too low and any FS IQ on the WAIS-III up to 10 points too high, due to this systematic error. As there is no more information as to which test is in error or by how much, as far as this error is concerned, all that can be assumed is that the true IQ lies within this 10 point range. Therefore a client's true IQ on the WISC-IV may be between zero and 10 points higher than the measured IQ and on the WAIS-III between zero and 10 points less than the measured IQ. There is therefore an additional 10 point uncertainty with each test which could be added to the 95% confidence intervals for the two tests, so that the interval would be extended up by 10 points for the WISC-IV and down by 10 points for the WAIS-III.

When these sources of error are combined it results in different effective confidence intervals for the WISC-IV and the WAIS-III. The WISC-IV will have a chance error of 15 points, to which must be added one point due to uncertainty as to the Flynn Effect, which gives an effective 95% confidence interval of 16 points. The WISC-IV may also measure up to 10 points too low due to other systematic error demonstrated by differences between it and the WAIS-III; however, it possibly measures one point too high due to the floor effect, therefore overall it may be measuring 9 points too low. Combining sources of error suggests an effective confidence interval extends 16 points below the measured IQ and 25 points above it. Therefore when using the WISC-IV one could not be confident that a child had an IQ less than 70 unless they got a measured IQ of less than 54 and could not be certain that they had an IQ above 70 unless they obtained a measured IQ of 95.

The WAIS-III will have a chance error of 15 points, to which must be added three points due to uncertainty as to the Flynn Effect, giving an effective 95% confidence interval of 18 points. It may also measure up to 10 points too high due to other systematic error demonstrated by the difference with WISC-IV. When these sources of error are combined the effective confidence interval extends 18 points above the measured IQ and 28 points below. Therefore on the WAIS-III one could not be confident that a client had an IQ less than 70 unless they got a measured IQ of less than 42 or that he/she had an IQ above 70 unless they obtained a measured IQ of 88.

These margins of error seem reasonable in the light of the evidence reviewed above, however, they are considerably greater than has previously been assumed and so are likely to be examined very critically and the assumptions made questioned. It may therefore be helpful to recalculate the margin of error making different assumptions. It could be argued that the 95% confidence interval of 13 points for FS IQ stability reported by Whitaker (2008b) is an overestimate. This is because the stability coefficients on which it was based were not corrected for the restricted range as is often done, and the SD of 15 that was used is the SD of IQ for the population as a whole rather than the SDs of the samples reported in the studies which were usually smaller. Although Whitaker (2008b) argues that the 13 point 95% confidence interval gives a good estimate of the percentages of IQs that actually do change by specific amounts, he re-calculated the

95% confidence interval using both stability coefficients corrected for restricted range and the mean SD reported in the studies, which resulted in a 95% confidence interval of 8 points. If it is also assumed that (a) it is appropriate to base the reliability figure on only the lack of stability of the test and (b) that the floor effect and Flynn effect can be corrected for, then the only error that would need to be added is the 10 points apparent from the systematic difference between the WISC-IV and WAIS-III. Adding these 10 points to the 8 point 95% confidence interval, based on the corrected stability figures, gives an effective 95% confidence interval for the WISC-IV, 18 points above the obtained score and 10 points below it, and for the WAIS-III 8 points above the obtained score and 18 points below it, which is still considerably greater than the five points suggested in the manuals.

Although the manuals of both the WISC-IV and WAIS-III do not suggest that there is more error at the low IQ level than in the mid range it is perhaps not surprising that there is. For example, if one considers how the tests are standardized it is apparent that it is not as rigorous at the low range as it is in the mid range. IQ tests are standardized using samples of the population as a whole; the bulk of the sample will therefore be made up of people in the average range and relatively few at the extremes. The standardization samples for both the WISC-IV and WAIS-III were split into groups of 200 people at different age levels. Therefore with IQ having a mean of 100 and an SD of 15 one would expect the samples to have only five people with IQ of less than 70, and none with IQs less than 55. This would have several effects. First, it would make sampling error much more likely. Second, it would mean that the relationship between raw scores, scaled scores and IQ scores at low levels would be based on extrapolation from what occurs in the average range rather than what is observed to occur. Thirdly, as the test items that effectively measure low IQ are items that are passed by all or the vast majority of subjects in the standardization sample, the psychometric properties of items will not have been properly assessed.

Clearly further research needs to be done. The current analysis is based on a small number of studies that have relatively small sample sizes and were possibly done under less than optimal conditions. It would help to clarify the issue of the degree of error in the measurement of low IQ if a large sample of several hundred children and adults with low IQ was used and the assessments were given under as near optimal conditions as possible. More reliable estimates could therefore be obtained for the split half reliability figure, the test re-test reliability figure and the degree to which the WISC-IV differs from the WAIS-III. In addition the data could be analysed using more up to date theoretical models of Generalizability Theory to specify the proportion of the variance accounted for by the various sources of error and Item Response Theory which would demonstrate how individual items function at the low IQ range.

Although the analysis presented in this paper may be questioned it is still difficult to escape the conclusion that the margin of error in the assessment of low IQ is much greater than the five points suggested in test manuals. Because of this, it seems unreasonable to base the definition of ID on a specified low IQ score such as 70, and a new definition should be sought such as that based on clinical judgment suggested by

Whitaker (2006, 2008a).

References

American Association on Mental Retardation (2002). *Mental Retardation: Definition, Classification, and System of Supports (10th Edition).* Washington DC: American Association on Mental Retardation.

American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders (4th edn), Text Revision.* Washington DC: American Psychiatric Association.

Anastasi, A. and Urbina,S. (1997). *Psychological Testing (seventh edition)*. Upper Saddler River: Prentice-Hall Inc

Baroff, G.S. (1999).General learning disorder: A new designation for mental retardation. *Mental Retardation*, 37, 68-70

British Psychological Society (2001). *Learning Disability: Definitions and Contexts.* Leicester: The British Psychological Society.

Cronbach, L.J., Gleser, G.C., Nanda, H. and Rajaratnam, N. (1972). The Dependability of behavioral measurement: Theory of Generalizability of scores and profiles. New York Wiley.

Davis, L.J. (1966).The internal consistency of the WISC with the mentally retarded. *American Journal of Mental Deficiency*, 70, 714-716

Department of Health (2001). *Valuing People: A New Strategy for Learning Disability for the 21st Century*. London: HMSO.

Flynn, J.R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, 95, 29-51

Flynn, J.R. (1985). Wechsler intelligence tests: Do we really have a criterion of mental retardation? *American Journal of Mental Deficiency*, 90, 236-244.

Flynn, J.R. (2006).Tethering the elephant capital cases, IQ and the Flynn Effect. *Psychology, Public Policy and Law*, 12, 170-189

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

Gordon, S. (2007). *Comparison of the WAIS-III and WISC-IV in 16 year olds who receive special education*. Unpublished Doctorate in Clinical Psychology at the University of Liverpool.

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

Sundet, J.M., Barlaug, D.G. and Torjussen, T.M. (2004). The end of the Flynn Effect? A study of secular trends in the mean intelligence test scores of Norwegian conscripts during the half a century. *Intelligence*, 32, 249-262

Schalock, R.L., Luckasson, R.A., Shogren, K.A., Borthwick-Duffy, S., Bradley, V., Buntinx, W.H.E., Coulter, D.L., Craig, E. P. M., Gomez, S.C., Lachapelle, Y., Reeve, A., Snell, M.E., Speat, S., Tasse' M.J., Thompson, J.R., Verdugo, M.A., Wehmeyer, M.L. and Yeager, M.H. (2007). The renaming of mental retardation: Understanding the change to the term intellectual disabilities. *Intellectual and Developmental Disabilities*, 45, 116-124

Shavelson, R. J. and Webb, N.M. (1991). Generalizabilty Theory: A Primer. Newbury Park: Sage Publications. Inc.
Spitz, H.H. (1986). Disparity in mental retarded persons' IQs derived from different intelligence tests. *American Journal of Mental Deficiency*, 90, 588-591.

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

Teasdale, T.W. and Owen, D.R. (1989). Continuing secular increases in intelligence and stable prevalence of high intelligence levels. *Intelligence*, 13, 255-262

Teasdale, T.W. and Owen, D.R. (2005). A long-term rise and recent decline in intellectual test performance: the Flynn Effect in reverse.
Personality and Individual Differences, 39, 837-843

Wechsler, D. (1939). *Wechsler-Bellevue Intelligence Scale*. New York: The Psychological Corporation.

Wechsler, D (1974). *Wechsler Intelligence Scale for Children – Revised: Manual.* New York: Psychological Corporation.

Wechsler, D (1981). *Wechsler Adult Intelligence Scale – Revised: Manual*.
New York: Psychological Corporation.

Wechsler, D. (1997). *WAIS-III, WMS-III: Technical and Interactive Manual*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children – Fourth Edition: Technical and Interactive Manual. San Antonio, TX*: The Psychological Corporation.

Wechsler, D. (2004). *Wechsler Intelligence Scale for Children – Fourth UK Edition:*

Administrative and Scoring Manual. London: The Psychological Corporation.

Whitaker, S. (2003). Should we abandon the concept of mild learning disability? *Clinical Psychology*, 29, 16-19

Whitaker, S. (2005). The use of the WISC-III and the WAIS-III with people with a learning disability: Three concerns. *Clinical Psychology*, 50, 39-40

Whitaker, S. (2006) What's in a name? Alternatives to learning disability. Mental *Health and Learning Disabilities Research and Practice*. 3, 177-191

Whitaker, S. (2008a). Intellectual disability: a concept in need of revision. *British Journal of Developmental Disabilities*, 54, 3-9.

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

Whitaker, S. & Wood, C. (2008). The distribution of scale score and possible floor effects on the WISC-III and WAIS-III. *Journal of Applied Research in Intellectual Disabilities*. 21, 136-141.