Error in the measurement of low IQ: Implications for the diagnosis of Intellectual Disability in court cases

By

Simon Whitaker

Introduction

A diagnosis of intellectual disability (ID) or what used to be called mental retardation (MR), could always have a major effect on people's lives. On the positive side it could provide services, finance, and help in schools. On the negative side it can be a stigmatizing label that an individual may seek to avoid (Baroff 1999). However, since the Supreme Court, in the case of Atkins vs. Virginia, prohibited the execution of individuals with MR, it can have life and death implications (Flynn 2006; Flynn 2007; Schalock et al 2007).

The Supreme Court did not provide a definition of MR, leaving it up to individual states to develop their own. This has resulted in a variety of definitions, which require different information to establish whether an individual has ID (c.f. Duvall and Morris 2006). A lack of intellectual ability has been explicitly part of most definitions since at least 1959, when the American Association on Mental Retardation (AAMR) (Heber 1959) defined it as a "Subaverage general intellectual functioning which originates during the developmental period and is associated with impairment in one or more of the following: (1) maturation, (2) learning, (3) social adjustment." (Cited by AAMR 2002, page 21). However, the different definitions used by individual states (Duvall and Morris 2006, Death

Penalty Information Center (DPIC) http://www.deathpenaltyinfo.org/state-statutes-prohibiting-death-penalty-people-mental-retardation) differ in whether an IQ cut-off point is specified. For example, according to DPIC's website, Maryland defines MR as: "An individual who has significantly subaverage intellectual functioning as evidenced by an IQ of 70 or below on an individually administered IQ test, and impairment in adaptive behavior. The age of onset is before the age of 22." As this definition states an IQ cutoff point of 70, it implies that unless an individual has a measured IQ of 70 or below he/she cannot be considered to have MR. On the other hand, the definition used in California is: "Significantly subaverage general intellectual functioning existing concurrently with deficits in adaptive behavior and manifested before the age of 18." Here no IQ cutoff point is specified so a measured IQ above 70 would not automatically rule out a diagnosis of MR.

It is the aim of this chapter to outline recent research findings on the accuracy to which low IQ can be measured and to consider the implications of this for definitions of ID/MR and its diagnosis in capital cases.

Error in the measurement of low IQ

It has always been accepted that IQ tests are subject to some error due to non-intellectual variables affecting the IQ score. The manuals of most modern intellectual assessment, such as the Wechsler Adult Intelligence Scale – fourth edition (WAIS-IV) (Wechsler 2008) and Wechsler Intelligence Scale for Children (WISC-IV) (Wechsler 2003a), provide information as to the accuracy of the assessments and examiners are encouraged to include this information in their reports. In the case of the WISC-IV and WAIS-IV it is claimed in the manuals that the measured IQ will be within five points of the true IQ 95% of the time. However, this claimed five point accuracy for modern tests may be very optimistic and misleading to those who are not familiar with how it is calculated. I therefore intend to give a brief description of how error can affect test scores and how test accuracy is calculated.

These errors are of two broad types, chance and systematic (Anastasi and Urbina 1997).

Chance errors are due to a large number of relatively small factors that may or may not occur during an assessment. These errors have the effect of diminishing the accuracy of an individual assessment but have a much smaller

effect on the mean of several assessments. According to Anastasi and Urbina (1997) there are three types of chance error in the measurement of IQ. First, a lack of internal consistency in the test due to test items measuring factors other than the psychological trait being assessed cause this. Secondly, temporal error, which is due to variation in the conditions under which assessments are administered, for example, the level of distraction in the room or the level of motivation of the client. Thirdly, there is scorer error, which is due to inconsistency in scoring the assessment and can be assessed by correlating the different scorers.

The degree to which each of these errors affect measured IQ can be represented statistically by the 95% confidence interval, which is the range of scores, either side of the measured IQ, in which the notional true IQ has a 95% chance of falling. Anastasi and Urbina (1997) provide the following formula for calculating it:

$$95\% \text{ Confidence interval} = 1.96 \times SD \times \sqrt{(1-r)}$$

where SD is the standard deviation of the test, usually set at 15 and r is the reliability coefficient of the test. This confidence interval can be calculated separately for each of the above sources of chance error.

In the past the reliability score for lack of internal consistency have been calculated by the split-half reliability method, where the items on a subtest are split in two halves and a correlation is found between them. However, it is now more commonly specified in terms of coefficient alpha, which gives a more sophisticated measure of this error. Temporal error reliability is calculated by correlating the scores when the same test is administered on two occasions to the same people. Scorer error is indicated by the correlation between scores when two separate scorers score the same test. It has been argued by Whitaker (2008, 2010) that an estimate of error due to a lack of internal consistency does not take into account temporal error or scorer error, and that an estimate of temporal error does not take into account error due to a lack of internal consistency, though it may take into account scorer error. Therefore to gain an estimate of the total chance error affecting a test score, one must take into account both error due to a lack of internal consistency, temporal error and scorer error. It is notable that this is not done when the 95% confidence interval is calculated for the most commonly used IQ tests such as the WAIS-

III, WISC-IV, and WAIS-IV (Wechsler 1997a, Wechsler 2003a, Wechsler 2008), where it is calculated taking into account only one source of error per subtest, usually the lack of internal consistency. This failure to use all the chance error affecting IQ scores in calculating the 95% confidence interval will have the effect of producing a much smaller confidence interval than if all sources of error were taken into account. If all sources of chance error were used then it would result in a 95% confidence interval much greater than the 5 points suggested in the test manuals.

Systematic error. These errors cause one IQ test to systematically score either higher or lower than other IQ tests, so that one test will on average score a fixed number of points higher or lower than another IQ test. Some of these errors are now well understood, such as the Flynn effect: the tendency for the intellectual capacity of the population as a whole to go up from one generation to the next, causing tests to overestimate IQ as they go out of date. Other sources of systematic error are not yet clear.

Error in the low IQ range

When one considers how tests are developed it seems likely that both chance and systematic error will be greater in the low IQ range than in the average

range. IQ tests are standardized using a representative sample of the population as a whole; most people in the sample will therefore be in the average intellectual range and relatively few will be at the high and low extremes. The standardization samples for both the WISC-IV and WAIS-IV (Wechsler 2003a, Wechsler 2008) were split into groups of 200 people at different age levels and the test standardization was essentially done on these sub-samples. Therefore, as IQ is set to have a mean of 100 and an SD of 15 and is normally distributed, one would expect these sub-samples to have only five people with IQ of less than 70, and none with IQs less than 55 in any sample of 200 people. Having only a small number of people with low intellectual ability in the sample could have a number of influences on the accuracy of the test in the low ability range. First, sampling error would have a much greater effect. For example, if, rather than having five individuals with true IQs less than 70, the sample had only two, which is perfectly possible when taking a random sample of 200, then the bottom 2% of the sample would perform better than if it was a truly representative sample. This would mean that test criteria for getting an IQ of about 70 would be set too high. Second, it would mean that the relationship between performance on the test and obtained IQ would have to be based on an extrapolation from the relationship found in the mid range, rather than being empirically derived. This is because the IQ an individual obtains is based on how

well they perform compared to the standardization sample. If, however, an individual performs less well than anybody in the standardization sample, the only way an IQ can be allocated is to assume that the relationship between performance on the test, found in the average range, continues into the low range and this may not be valid. Thirdly, as the vast majority of subjects in the standardization sample pass the test items that effectively measure low IQ, the psychometric properties of items may not have been properly assessed. Recent evidence seems to support the notion that both chance and systematic error may be greater in the low range.

Chance error

Temporal error. The current evidence suggests that temporal error may be greater in the low range than it is in the average range. A meta-analysis (Whitaker 2008) of the test re-test reliability for Full Scale IQ (FSIQ) for assessments in the low IQ range (IQ<80) found a weighted mean test re-test reliability of .82. To explain what this means in terms of quantifying the error: All the studies used in the meta analysis gave the same test to the same individuals twice. As one would expect the individual's "true IQ" to remain the same between assessments, it would be expected that if the tests were not

subject to error the same score would be obtained on both assessments so the correlation would be 1.00. Therefore, the difference between the actual obtained correlation and 1.00 represents the amount of error due to a lack of stability of the test. The temporal error is therefore one minus .82, which is .18 and the corresponding 95% confidence interval is 12.5 points. As several of the studies in the meta analysis reported the proportion of IQs that changed by specific amounts, Whitaker was able to check how accurate this 95% confidence interval was in predicting IQ change when it was re assessed. It would be expected that, for a 95% confidence interval 12.5 points for stability, 61% of IQs would change by less than 6 points and 13% would change by 10 points or more. In the studies, 57% of IQs changed by less than six points and 14% changed by 10 points or more, which is a good estimate of what was predicted, suggesting that the 95% confidence interval of 12.5 points is accurate for clients in the low ability range. However, it is considerably greater than the corresponding 95% confidence intervals for temporal error of 9.75 and 5.88 for the WISC-IV and WAIS-IV respectively for people in the average ability range (Wechsler 2003a, Wechsler 1997a). There is therefore more fluctuation in individuals' tested IQs at the low level than in the average range. This may be due to greater error in measurement, actual change in intellectual ability between tests or both. Whitaker (2008) found that there was no significant

relationship between the test re-test reliability and the interval between testing. This suggests the change was not due to a systematic change in intellectual ability over time and the most likely explanation for the change in scores is measurement error, however, there is a possibility that intellectual ability fluctuates from day to day.

Internal consistency. There is now an up to date estimate of error due to a lack of internal consistency in the low range. As part of the validation of the WAIS-IV (Wechsler 2008), it was given to a group of 75 adults with mild ID and 35 with moderate ID and the internal consistency reliability calculated for those sub-tests where this could be done. On average this reliability was approximately the same as the internal consistency reliability found for the standardization sample, which means that the overall reliability for internal consistency in the low range is approximately .98, the reliability found in the average range. This would mean the error due to a lack of internal consistency in the low range is about .02 and the 95% confidence interval about 4.2 points.

Total chance error. If the error due to lack of internal consistency of .02 is added to the error due to a lack of temporal stability, found by Whitaker (2008), to be .18, it gives an estimate of the total chance error in the low range

of .20. It was argued by Whitaker (2010) that an estimate of the effective total reliability could be obtained by subtracting the estimate of total error from one. If this is done for the total chance error of .20, it gives an effective total reliability figure of .80, which corresponds to a 95% confidence interval of 13 points.

Systematic error

A floor effect. All the Wechsler intellectual assessments measure IQ by giving the client a number of subtests measuring different aspects of intellectual ability. The maximum possible raw score on the different subtests is different on each of them. As part of the calculation of FS IQ, the raw scores on these subtests are converted to scaled scores with a mean of 10, an SD of three, and a range between one and 19. Whitaker (2005) has suggested that allocating a scaled score of one to low raw scores or a raw score of zero could result in an overestimate of intellectual ability. This can be illustrated by looking at the relationship between raw scores and scaled scores for the Digit Span subtest for age groups 15:8 to 15:11 taken from WISC-IV Administrative Manual (Wechsler 2003b):

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

The relationship is linear between raw score 18 and raw score 10, a reduction in a raw score by one corresponding to a reduction in a scaled score by one. However, all raw scores from nine down to zero are then given a scaled score of one. 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. Therefore a scaled score of one given for a raw score of less than nine is likely to be an overestimate of the client's ability. So generally when a scaled score of one is given there is a distinct possibility that the client's ability is being overestimated. This will clearly 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 from WISC-III (UK) (Wechsler 1992), and WAIS-III (UK) (Wechsler 1997b), assessments that had been given

as part of clinical practice for people with ID. The distribution of scaled scores for the WAIS-III (UK) was 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 with more scaled scores of one than any other scaled score. Scaled scores of one were found at all IQ levels up to and including those in the 70s, where they accounted for 10% of scaled scores. A similar distribution of scaled scores to that found on the WISC-III (UK) has now been found on the WISC-IV (UK) (Whitaker and Gordon 2012). There is therefore a distinct possibility that, particularly on the WISC-III and WISC-IV, IQ scores are increased at low ability level due to this floor effect.

The Flynn Effect. There is good evidence that the intellectual ability of the population as a whole has increased over at least the last 100 years at about 0.3 of an IQ point per year on average (Flynn 1984; 1987; 2007). This is known as the Flynn effect, after James Flynn, who initially researched it. The evidence also shows that the effect occurs at the low IQ levels (Flynn 1985; 2006) and is still doing so today in the US (Flynn 2009). The implication of the Flynn effect for the assessment of IQ is that tests will become less accurate and overestimate intellectual ability as they become more out of date. On average a

test will overestimate an individual IQ by about 0.3 IQ for each year since it was standardized. It has therefore been argued by Flynn (2007; 2009) that it is possible to compensate for this error by subtracting .3 of an IQ point from the measured IQ for each year between the test being standardized and given.

Error apparent from the differences between IQ scales. It is accepted that different IQ tests will give slightly different results (Floyd et al 2008). In the absence of a test that is clearly an accurate measure of true intellectual ability the best that can be done is 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, it has been reported that early versions of the WISC scored systematically lower than the equivalent WAIS when used in the low intellectual range (Flynn 1985, Spitz 1986, 1989). There is also a recent study by Gordon et al (2010) that compared the WISC-IV (UK) and the WAIS-III (UK) in the low range. Both assessments were given in counterbalanced order to a group of 16 year-olds receiving special

education. In each case the FS-IQ on the WISC-IV (UK) was less than that on the WAIS-III (UK); the mean FS-IQ on the WISC-IV (UK) was 53.00, which compared to a mean of 64.82 on the WAIS-III (UK), a difference of just less than 12 points. The correlation between the two assessments was relatively high (r=.93), suggesting that the tests were both measuring the same trait and were given consistently. Of this 12 point difference between the two tests, it is likely that two points are due to the Flynn effect as the WAIS-III was standardized six years before the WISC-IV, while the remaining 10 points are due to yet unknown factors. As the degree to which either assessment is in error is not known, it is clearly possible that either the WISC-IV (UK) is systematically underestimating true IQ by up to 10 points, or the WAIS-III (UK) is systematically overestimating true IQ by 10 points or both assessments are making systematic errors of less than 10 points. Although this work was done with the UK versions of the WISC-IV and the WAIS-III, there is unpublished evidence (abstract is available Bresnahan 2008) that the same effect occurs on the US versions of the tests.

Combined error. Whitaker (2010) combined the error in the low range from lack of internal consistency, temporal error, the floor effect, the Flynn effect and the error apparent from the difference between the WISC-IV and WAIS-III

and estimated the 95% confidence intervals for both the WISC-IV and WAIS-III. For the WISC-IV there was an effective confidence interval, which extends 16 points below the measured IQ and 25 points above it. For the WAIS-III the effective confidence interval extended 18 points above the measured IQ and 28 points below. This analysis is based on a number of assumptions with regard to combining measurement error and makes use of data from the UK versions of the WISC and WAIS. It also based the error due to a lack of internal consistency on a study by Davis (1966) in which a mean reliability figure of .92 was found for the low range on the WISC, which is significantly lower than was found in the standardization of the WAIS-IV referred to above. Because of this, the effective 95% confidence for total chance error was 15 points rather than the 13 points suggested above, using the internal consistency of .98 taken from the WAIS-IV standardization. It therefore will be argued that this very large margin of error will not apply in the US today. However, it would be very difficult to escape the conclusion that the degree of error in the measurement of low IQ is much greater and the tests far less accurate than had previously been supposed.

Other error specific to legal assessments

In addition to the above sources of error that Whitaker (2010) combined, the literature on the measurement of IQ in forensic cases highlights two additional sources of error: malingering and the practice effect.

Malingering. This is deliberately underperforming on an assessment in order to get a low score. An individual appealing the death penalty on the grounds of having an ID may be motivated to do this. It is clear from the literature (c.f. Salekin and Doane 2009) that courts are well aware of the possibility of this error in assessment, however, currently there does not seem to be a reliable way of detecting when it is occurring when assessing an individual with low intellectual ability.

Practice effect. This occurs when an individual does better on a test the second time he/she is given it due to having had the opportunity to practice when they were first given it. The degree to which it affects the second test score can be estimated from studies that have given the same test to the same clients on two occasions, such as test re-test reliability studies. Both the WISC-IV and WAIS-IV manuals (Wechsler 2003, Wechsler 2008) describe such studies, with mean FSIQ increasing by 5.6 points on the WISC-IV over an

average test re-test interval of 32 days and by 4.3 points on the WAIS-IV over a mean interval of 22 days. However, practice effect is likely to decrease as the interval between testing increases and may well also be a function of ability level. In this respect it is notable that Whitaker' s (2008) meta analysis of test re-test reliability in the low range found a mean increase in FSIQ of only .41 of an IQ point for a mean test re-test interval of 2.3 years.

Correcting for errors in the assessment of low IQ

It may be possible to reduce and even eliminate some of these errors in the assessment of an individual's intellectual ability. However, it is the contention of the current author that it is not currently possible to reduce the errors to such an extent that the accuracy of the tests approaches the five points suggested in the manuals, nor is it possible to quantify how accurate the tests are after most corrections have been applied.

The error due to lack of internal consistency reported in the WAIS-IV manual, even in the low IQ range, is very small. The error is due to inconsistent test items and the experience of the client, possibly over many years. Neither of these can be changed for an individual assessment and so there is no score of reducing this error for an individual assessment. If, however, it is assumed that

these errors will be the same when tests are given in a clinical or forensic setting, and there is no reason to suppose they would not be, then the accuracy of the test can be quantified with regard to lack of internal consistency by the 95% confidence interval.

Temporal error is due to changes in the state of the individual, such as alertness and motivation, changes in the environment in which the test is given, such as level of distraction, or changes in the way the test is given, such as how the accurately the test instructions are given, between assessments. Here being very strict about only giving the test in optimum conditions may well reduce error. It is likely that many of the assessments that are done in clinical and criminal justice settings are not done under optimal conditions and this may be one reason why the test re-test reliability found by Whitaker (2008) is lower than that found when the tests were standardized. For example in criminal justice settings such as prisons there may be distractions due to noise; the client may be depressed or anxious; he/she may not be motivated to do well, the examiner may not be used to giving assessments in such settings and/or to giving assessments in the low IQ range. However, even though these errors could be reduced, it will not be possible to tell by how much they have been reduced in numerical terms. Without further studies in which the test re-test

is assessed under various conditions, one must rely on the studies that have been done up to now which, according to Whitaker (2008), suggest a 95% confidence interval of 12.5 points for the stability of IQ over time.

The Flynn effect could be minimised by always using the latest standardization of a test, because if a test has only just been standardized there will be no Flynn effect. If a test is a few years old, Flynn (2009) has argued that the effect can be corrected by subtracting 0.3 of an IQ point for each year since the test was standardized. However, whereas it may have been the case for the US in 2009, it may not always be the case and does not seem to be the case in other areas of the world at the moment. There is evidence from Scandinavia that the effect may have gone into reverse in the low range (Teasdale and Owen 2005), resulting in tests underestimating IQ as the go out of date. It therefore cannot be assumed that this method of correction for the Flynn effect will continue to be valid, nor can we tell, without extensive studies, how valid it is at any one time.

Whitaker and Gordon (2012) have suggested that the floor effect could be corrected for by extrapolating the relationship between raw scores and scaled scores down below scaled score one, so that scaled scores of zero and less can

be allocated to low raw scores. However, although this method seems logical and may be valid for small correction where a scaled score of zero or minus one is given instead of a scaled score of one, there is no empirical evidence to support the procedure. Therefore, although applying some correction for the floor effect may result in a more accurate assessment, it is not possible to say how much more accurate an IQ score corrected for the floor effect is than one that is not corrected.

The error apparent from the differences between tests may be very difficult to correct for. It could only be done if there was firm evidence that one test was accurate and it was known by how much other tests' scores systematically differed from this gold standard test. Adding or subtracting the appropriate number of points from their obtained scores could then correct inaccurate tests. However, this is currently not possible for a number of reasons: First, although Gordon et al (2010) have produced evidence that the WISC-IV systematically measured lower than the WAIS-III, the study was based on a small sample of 16-year-olds and so it cannot be assumed that the exact difference between the tests due to factors other than the Flynn effect is 10 points. Secondly, the WAIS-III is now no longer used and there are no studies comparing the WAIS-IV and WISC-IV in the low intellectual range. Thirdly

although it could be argued that the WISC-IV is more likely to be accurate than the WAIS-IV, as getting a representative sample of children with low IQs would seem to be less subject to error than getting one of adults (c.f. Flynn and Weiss 2007), this is only speculation and in reality we do not know to what degree either of these gold standard tests is accurate or inaccurate or indeed if other less well known tests are more accurate in the low range.

The practice effect could be eliminated by not assessing an individual twice on the same assessment. However, if this has to be done, then subtracting an appropriate number of IQ points from the second assessment could reduce the error. The problem then is deciding how many points to reduce the second assessment by. The practice effect will vary depending on which test is being given, the time interval between assessments as well as factors within the individual such as his/her level of intellectual ability. Therefore one cannot be sure how accurate a corrected score is.

Clearly one way to eliminate error due to malingering is not to assess individuals who may well be motivated to do poorly in the assessment, which may very well be the case when an assessment is being done in a death penalty case. However, if this had to be done, there does not seem to be any reliable way of detecting

malingering (Salekin and Doane 2009), which means there also is no way of telling to what degree a score has been affected by it.

It is clear therefore that IQ tests are subject to more error, particularly in the low range, than has previously been accepted. It is also likely that, although some of these errors can be reduced or even eliminated, it is not possible to say how accurate the assessments are in terms of a 95% confidence interval without making assumptions for which there in a lack of evidence. One is therefore left with the relatively wide margins of error of the order of that suggested by Whitaker (2010).

Current appreciation of these errors

Judging by the current literature on the measurement of low IQ there seems to be only a partial awareness of the errors outlined in this paper. There is a clear understanding of the Flynn effect (Ceci et al 2003; Flynn 2009; Olley 2009a). There is some awareness of the poor stability in the low range (Olley 2009a, b). There is little questioning of the practice of calculating the 95% confidence interval on the basis of only one source of error, though Gresham (2009) does discuss it. Apart from Whitaker and Wood (2008) there is no acknowledgement of the floor effect, which could increase scores by the same

order of magnitude as the Flynn effect. There seems to be no realisation that the WISC and WAIS may differ by the order of 10 points. Apart from Whitaker (2010), there has been no attempt to combine all these errors and get an overall degree of confidence. There seems to be an over reliance on information on reliability given in the test manuals, which is based on data obtained using a non-clinical and non-forensic sample who, on the whole, had average intellectual abilities. If this lack of awareness extends to those advising courts and doing assessment in criminal cases then it is likely that courts are being misled.

Implications and recommendations

The greater error in the measurement of low IQ has a number of implications for the assessment of intellectual ability in criminal cases and particularly death penalty cases. These are outlined below, together with some recommendations as to how best to deal to address them.

Confusion between true intellectual ability and measured IQ

It was noted above that some state definitions of ID specify a specific IQ figure below which a defendant must score to be considered to have MR, while others simply indicate that the individual must have a significantly low

intellectual ability. It is likely that the lack of precision in the measurement of low IQ will have a greater negative impact in states where the definition of MR specifies an IQ figure. If the individual being assessed has a true intellectual ability in the mild to borderline range, whether or not he/she obtains a measured IQ below the specified figure will be to some extent a matter of luck. But more than this, it gives a score for the assessor to produce the IQ score that may suit the case of those engaging them to do the assessment. Although clearly the responsibility of an expert witness is to give impartial advice to the court, it is possible that they may unconsciously swayed to produce a particular result, or that the attorneys employing them may have chosen them for their apparent views or track record in obtaining certain results. Therefore, if the assessor is commissioned by the defence it is possible that they may want the client to have a measured IQ less than the specified figure. This could be made more likely if they do the following:

Put emphasis on WISC-IV rather than WAIS-IV scores. Although in a death penalty case that the defendant would be older than 16, so it would not be possible to reduce his/her score by assessing him/her on the WISC-IV rather than the WAIS-IV, the assessor could still argue that greater weight should be given to WISC-IV and WISC-III assessments done when the

defendant was a child rather than WAIS-IV assessments done following committing the offence. In part this is because of problem with malingering but also as it has been suggested that the WISC-IV may be a more accurate assessment at low levels as it is easier to get a representative population sample of children with low intellectual ability than it is of adults to standardize the test (c.f. Flynn and Weiss 2007).

Allow the assessment to be done under sub-optimal clinical or forensic conditions. It is likely that if an IQ assessment is done under sub-optimal conditions, such as may occur in some prison or clinic settings, the score will be reduced. In order to obtain a lower score the assessor may therefore not insist on a distraction free environment or that the defendant was in a state to give of their best.

Ensure the defendant is aware of the consequences of a high score. For example in death penalty cases the assessor could make it clear to the defendant that a high score will increase the likelihood of being executed.

Not using a test that the client has been assessed on recently. By not re-assessing the defendant on a test he/she has previously been assessed on, there will be no practice effect.

Correct for the Flynn and floor effects. It was argued above that scores can be artificially increased by both the Flynn effect and the floor effect but that these scores could be corrected to some extent.

If the assessor was working for the prosecution he/she may wish to ensure a high score, in which case he/she could:

Put emphasis on the WAIS-IV rather than the WISC-IV. The assessor could argue that emphasis should be put on more recent WAIS-IV score than older WISC-IV or WISC-III scores as the WAIS-IV results reflect how the individual is now and that higher scores are less likely to be subject to error.

Ensure that assessments are done under optimal conditions. Rather than accepting poor conditions such as may be offered in a prison setting the assessor could insist on as near to optimal conditions as possible. They could also

put off an assessment if the defendant was not in an optimal condition himself or herself to be assessed.

Ensure that the individual is motivated to do well on the assessment. Rather than emphasise the negative effects of doing well on the assessment emphasis could be put on the positive consequence, such as showing that you are smart.

Re-assess on a test that has been used recently in the past. By using an assessment that the individual has only recently been assessed on, possibly by the defence assessor, the score should be increased due to the practice effect.

Not correcting for the Flynn or floor effects. The obtained scores could simply be presented without any correction. Also, if possible, the assessor could use an older version of the test to maximise the Flynn effect.

Although it is not be possible to put an exact figure on the degree to which it would affect scores, if an individual were assessed under these two extreme sets of conditions, it is likely that the two obtained IQ scores would differ by the order of 25 points. Twelve points would be attributable to differences between the tests used due to the Flynn effect and other factors, 10 points to

temporal error due to difference in assessment conditions, one point for the floor effect, two points for the practice effect, as well as some effect due to malingering. However, both assessments would produce a measured IQ score. If a court was unaware that these manipulations could have such a large effect on measured IQ and regarded measured IQ as a good indicator of true intellectual ability, then it would be likely to accept an IQ score as a good indicator of the defendant's actual intellectual ability. The likelihood of this happening could be reduced if there was an explicit distinction drawn between measured IQ and true intellectual ability. An individual's true intellectual ability, at any one time, is the IQ score that he/she would obtain if they were assessed with a perfectly standardized, valid and reliable IQ test without any error in administration. An individual's measured IQ is the IQ score he/she would obtain if he/she were a given a current IQ test, under a particular set of conditions. Currently measured IQ is only a loose indicator of true intellectual ability.

Improving the accuracy of measured IQ

As noted above, an individual's measured IQ and his/her true intellectual ability may differ by the order of 25 points. If this is acknowledged and the assessors and the courts are motivated to get the best estimate of true intellectual ability, there are a number of things that can be done:

- It should be made explicit in any reports and evidence given to the court that measured IQ and true intellectual ability are not the same thing and that measured IQ is only a rough estimate of true intellectual ability.
- The assessor should correct for the Flynn effect and the floor effect. As these corrections may be contentious, the fact that they have been made should be indicated, together with a justification for having made them, and a statement as to how much difference it made to the IQ scores.
- If it is known that the test being used systematically measures lower or higher than another gold standard test, then this should be made very clear in the report. So, for example, if reporting on the results of a WAIS-III or WAIS-IV, it should be noted that there is evidence that it measures higher than the equivalent the WISC-IV, and that it is not clear which test is producing the best estimate of true intellectual ability. Although this will not eliminate the systematic error it will make courts aware of it.
- Although IQ tests done prior to the defendant being charged may not be adversely affected by motivation to do poorly, the information that comes with them may reduce their value. If one knows the test that was used and when it was given one can adjust for the Flynn effect and make

it comparable with other tests. If the raw scores are available the score can be corrected for any floor effect. If there is information as to the conditions under which the assessment was given one can speculate as to whether the score was higher or lower than would have occurred if the assessment was done in optimal conditions. However, simply giving an IQ score without any information about how the IQ was obtained is almost worthless. If an old IQ assessment is used in evidence the assessor should obtain as much information as possible as to how it was done and indicate in his/her report how this information would affect the score and how reliable the score is felt to be.

- If there have been several intellectual assessments given the assessor should look for consistency in the different IQ scores. In order to do this all the scores should be corrected in the same way for the Flynn effect, floor effect and the difference between tests. A key question is then: do the scores give more or less the same result? If so, it provides stronger evidence that there is consistency in measured IQ, suggesting that chance error has had only a relatively small effect on the scores. It will also provide evidence as to whether the individual's true intellectual ability is above or below the criterion IQ for a diagnosis of ID. However, some caution would have to be exercised before drawing such a conclusion

as systematic error is not corrected for. In correcting for the systematic error apparent between tests we can adjust a score on one test so that it is comparable to the score on another test. For example, if an individual scored 62 on the WISC-IV, with no floor effect, and 72 on the WAIS-III, when corrected for the Flynn effect and with no floor effect, then this 10 points, which evidence suggests is the expected difference between the two tests (Gordon et al 2010), could be subtracted from the WISC-III score, making the adjusted score on the WAIS-III 62, the same as the score on the WISC-IV. One may therefore be tempted to say that the true intellectual ability of the defendant was 62. However, this would be to assume that it is the WAIS-III that is systematically in error in measuring true intellectual ability and not the WISC-IV, whereas all we can say is that either test may be in error by up to 10 points, so the individual's true IQ could be 72. The other practical problem with this is that we do not know the degree to which most tests differ from each other in the low range so in reality adjusting test scores so they are consistent with other scores is not something that can be done.

It may well be the case that, even when test scores have been adjusted, there are still significant differences between the results of assessments done at different times and/or with different tests. In this situation it is not entirely clear what should be done, though there are a number of options, each with advantages and disadvantages. One possibility would be to take the average score. This would assume that some assessments were subject to errors that increased the scores above true intellectual ability and some to errors that decreased scores below true intellectual ability, so that averaging the scores cancelled out the errors to give a good estimate of true intellectual ability. However, this would also assume that the individual's true intellectual ability was the same on each occasions he/she was assessed and that the effect of errors was evenly distributed between those that increased scores and those that decreased scores, which are assumptions that one would not normally have evidence for.

A second approach would be to argue that most error will decrease score, so that lower scores are therefore likely to be subject to more error and therefore the highest scores are more accurate. However, whereas it does seem that high scores are more likely to be correct, this is not inevitable and

the high scores could be due to systematic error or change in intellectual ability over time.

Real world consistency, that is how the defendant has coped with intellectually loaded everyday tasks, could give support to a particular score. In part this could be indicated by an assessment of the individual level of adaptive behaviour on a scale such as the Vineland II (Sparrow et al 2005). However, the assessor should be aware that there are often only low correlations between adaptive behaviour and IQ, for example the correlation between the Vineland II composite score and FS IQ on the WISC-III is only .09. So ability to do some tasks such as socialise with others or have daily living skills are not indicative of high IQ. On the other hand IQ does correlate well with academic ability. As part of the standardization of the WAIS-IV, it was compared with the Wechsler Individual Achievement Test (WIAT-II) (Wechsler 2001). FS IQ on the WISC-IV correlated with reading .78, with Math .78, with written language .76, with oral language .75 and with total achievement .87. These relatively high correlations suggest that if an individual has a genuine IQ less than 70 it is unlikely that they would be performing in the average range academically.

It is likely that even if the above suggestions are followed it still will not be possible to give an exact figure on an individual's IQ or an exact confidence interval indicating how accurate the test result is. It should therefore be made clear in any information given to a court that any score is not exact and the confidence interval is many more points than that stated in the manual.

If the IQ score is required for a diagnosis, the issues may not be so much what the exact IQ figure is but rather whether it is below the critical figure required for a diagnosis of ID. To take an extreme example, if an individual had several measured IQs in the 90s, then, even though it would not be possible to put an exact figure on his/her IQ, one could say with a high degree of certainty that the individual's true intellectual ability was above 70. However, it is likely that the majority of defendants arguing in court that they have ID will have measured IQs in the 60s and 70s, in which case it will not be possible to say with any certainty that their true intellectual ability falls above or below 70. The best approach in such cases may be to give an estimate of the probability of the individual's true intellectual ability falling above or below the critical figure in non exact term, using such works as: possible, likely or unlikely.

Courts' understanding of the information

It has been pointed out that courts may have some difficulty understanding complex information about psychometrics and so fall back on their own preconceived ideas of what constitutes MR/ID (Greenspan and Switzky 2006). Often these lay ideas as to what MR/ID is are somewhat different from what is intended by the formal definitions of MR/ID and may require an individual to show very obvious signs of disability. The problems in measurement of low IQ outlined here are likely to add to the confusion that courts have and there may be a tendency to fall back on a lay interpretation of what IQ is. The way the courts think about the accuracy to which IQ can be measured may be influenced by the way IQ is spoken about by the general public, usually referred to as a single whole number, very much in the same way as a person's height may be reported in feet and inches. It may well be that the way measured quantities are reported implies a degree of accuracy to which they have been measured. For example, an individual's height is usually be reported as being in feet and inches implying that height can be measured to an accuracy of one inch, which it can. An individual's IQ is usually spoken about amongst lay people and reported in the media as a whole number, implying that it can be measured to the level of accuracy of one IQ point. It is therefore likely that most lay people consider that IQ can be measured to an accuracy of one point. This then may lead courts

to regard the concept of error in its measurement with suspicion. It is therefore important that it is clearly explained to courts that measured IQ is subject to error and it is only a rough estimate of true intellectual ability. Possibly a good analogy to help then grasp the point is an individual's weight which will fluctuate from day to day by a pound or so, and will vary between scales, though the overall accuracy of measuring an individual's weight is far greater than that of his/her IQ.

Burden of Proof

When the Supreme Court made its ruling prohibiting the execution of people with MR and when the states drew up their definition of MR, the prevailing opinion was that IQ could be measured to an accuracy of 5 points. It is the core argument of this paper that the margin of error is much greater than 5 points. Therefore, on the one hand it is much more difficult to prove absolutely that somebody has a true intellectual ability of less than 70, and on the other hand it raises the possibility that an individual with measured IQs in the 80s may have a true intellectual ability in the 60s.

The burden of proof for establishing that a defendant has ID usually lies with the defence. The defence will therefore have to present evidence and argument to the court to demonstrate that the individual does have ID, which may well be challenged by the prosecution. As a great deal of what was considered to be established fact with regard to the measurement of low IQ has been brought into question, the likelihood is that there will be considerable argument in court as to what a defendant's IQ is. However, in reality, with our current state of knowledge, in many cases it will not be possible to say with any degree of certainty that an individual has a true intellectual ability above or below 70. It is therefore a question of whether the benefit of the doubt is given to the defendant or the prosecution. If it is given to the prosecution then it is likely that people who genuinely do have MR will not be able to establish this and will be executed, which was not the intention of the Supreme Court.

References

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

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.

Bresnahan, J. A. (2008). A preliminary study of WISC-iv and WAIS-III IQ scores for students with extremely low cognitive functioning. *Abstract of unpublished dissertation.* Dissertation Abstracts International: Section B: The Science and Engineering Vol 68(9-B), 6383.

Ceci, S.J., Scullin, M. and Kanaya, T. (2003). The difficulty of basing death penalty eligibility on IQ cutoff scores for mentally retarded. *Ethics and Behavior*, 13, 11-17.

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

Duvall, J.C. and Morris, R.J. (2006). Assessing mental retardation in death penalty cases: Critical issues for psychologists and psychological practice. Professional Psychology: Research and Practice, 37, 658-665.

Floyd, R.G., Clark, M.H. and Shadish, W.R. (2008). The exchangeability of IQs: Implications for professional psychology. Professional Psychology: *Research and Practice*, 39, 414-423.

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. (1987). Massive gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, 101, 171-191.

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.

Flynn, J.R. (2009). The WAIS-III and WAIS-IV: Daubert motions favor the certainly false over the approximately true. *Applied Neurology*, 16, 98-104.

Flynn, R.J. and Weiss, L.G. (2007). American IQ gains from 1932 to 2002: The WISC subtests and educational progress. *International Journal of Testing*, 7, 209-224.

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.

Greenspan, S. and Switzky, H.N. (2006). Lessons from the Atkins decision for the next AAMR manual. In H.N. Switzky, and S. Greenspan (Eds) *What is Mental Retardation? Ideas for an Evolving Disability in the 21st Century: Revised and Updated Edition*.
Washington DC: American Association on Mental Retardation.

Gresham, F.M. (2009). Interpretation of intelligence test scores in Atkins cases: Conceptual and psychometric issues. *Applied Neuropsychology*, 16, 91-97.

Heber, R. (1959). *A manual on terminology and classification in mental retardation (rev. ed.)*, Washington DC: American Association of Mental Deficiency.

Olley, J.G. (2009a). Challenges in implementing the Atkins decision. *American Journal of Forensic Psychology*, 27, 63-73.

Olley, J.G. (2009b). Knowledge and experience required for experts in Atkins cases.
Applied Neuropsychology, 16, 135-140.

Salekin K.L. and Doane, B.M. (2009). Malingering intellectual disability: The value of available measurements and methods. *Applied Neuropsychology*, 16, 105-113.

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.

Sparrow, S.S., Cicchetti, D.V. and Balla, D.A. (2005). *Vineland-II Vineland Adaptive Behavior Scale Second edition.* Circle Pines: AGS Publishing.

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. (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. (1992). *WISC-III UK Administrative and scoring Manual*. London: The Psychological Corporation.

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

Wechsler, D. (1997b). *WAIS-III UK Administrative and scoring Manual*. London: The Psychological Corporation.

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

Wechsler, D. (2001). *Wechsler Individual Achievement Test – Second Edition. San Antonio, TX*: The Psychological Corporation.

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

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale – Fourth Edition: Technical and Interactive Manual. San Antonio, TX*: The Psychological Corporation.

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. (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. & Gordon, S. (2012). Floor effects on the WISC-IV. Submitted to *Journal of Applied Research in Intellectual Journal of Developmental Disabilities,* 57, 40-47.

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.