The Stability of IQ in People With Low Intellectual Ability: An Analysis of the Literature

Simon Whitaker

Abstract

A meta-analysis of the stability of low IQ (IQ < 80) was performed on IQ tests that have been commonly used—tests that were derived by D. Wechsler (1949, 1955, 1974, 1981, 1991, 1997) and those based on the Binet scales (L. M. Terman, 1960; L. M. Terman & Merrill, 1972). Weighted-mean stability coefficients of .77 and .78 were found for Verbal IQ (V IQ) and Performance IQ (P IQ) on the Wechsler tests and .82 for Full-Scale IQ (FS IQ) on both Wechsler and Binet tests, for a mean test–retest interval of 2.8 years. Although the majority of FS IQs changed by less than 6 points, 14% changed by 10 points or more. The author suggests that the results of IQ assessment should be treated with more caution than previously thought.

IQ tests have been used to measure low intellectual ability ever since Binet and Simon produced their original test in 1905. The diagnosis of *mental retardation* still requires a measured IQ below a specified figure, usually IQ 70 (American Association on Intellectual and Developmental Disabilities [formerly the American Association on Mental Retardation], 2002; American Psychiatric Association, 2000; British Psychological Society, 2001). Consequently, individuals' measured IQ can have a significant effect on their diagnosis and the services they receive. Knowledge of the degree to which individuals' measured IQ is likely to change if they are reassessed would seem to be important, as it would have clear implications for any diagnosis of mental retardation.

Any psychometric assessment is subject to error. An estimate of this error is given by the *standard error of measurement* (*SEM*), which is the *standard deviation* (*SD*) of test scores that would be expected to occur if the tests were repeatedly given to the same client. *SEM* is a function of the reliability of the test: the higher the reliability coefficient, the lower the *SEM*. For the Wechsler IQ assessments (Wechsler, 1949, 1955, 1974, 1981, 1991), *SEM*s of approximately 2.5 points have been reported. However, the reliability coefficient used to calculate *SEM* on the Wechsler tests is based on the split-half reliability of the subtests, obtained by correlating participants' performance on alternative items in subtests. The resulting correlation gives an estimate of the error that is due to a lack of internal consistency (Anastasi & Urbina, 1997). However, as split-half reliability only requires a client to take the test on one occasion, it does not account for any error that is due to change between the two assessments. These could include changes such as in the situation in which the test was given, the state of the client when he or she was assessed, or any genuine change in intellectual ability.

The degree to which measured IQ changes over time is indicated by the stability coefficient, the correlation between assessments done some time apart with the same individuals, using the same test. The stability coefficient accounts for error that is due to changes both in test situations and in the state of the individuals between assessments, as well as any genuine change in ability. Stability coefficients for Full-Scale IQ (FS IQ) of .89 and .91 have been reported for the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV; Wechsler, 2003), and the Wechsler Adult Intelligence Scale, Third Edition (WAIS-III; Wechsler, 1997), respectively, for assessments done about 1 month apart. Although these stability coefficients seem reasonable, they result in greater *SEM*s than those based

on the split-half reliability. Anastasi and Urbina (1997) gave the following formula to calculate *SEM*:

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

where *SD* is the standard deviation of the test and *r* is a reliability coefficient. When this formula is applied to above stability coefficients for FS IQ and an *SD* of 15, it results in *SEM*s of 5.0 and 4.5 for the WISC-IV and the WAIS-III, respectively. This compares with the *SEM*s cited in the manuals that are based on split-half reliability of 2.7 and 2.3 for the WISC-IV and the WAIS-III, respectively.

Because the *SEM* corresponds to the *SD* of IQ scores that would be expected if an assessment were repeatedly given to a single client, it can be used to calculate the proportion of scores that would be expected to change by given amounts. If IQ were a continuous variable, one would expect 68% of scores to change by less than 1 *SEM* and 4.5% by more than 2 *SEM*. However, because IQ is given in whole numbers, allowance has to be made for this in the calculations. Taking this into account, one would expect that, for the stability *SEM* of 5 points, 73% of WISC-IV (Wechsler, 2003) FS IQs would change by less than 6 points and 6% would change by 10 points or more. However, it would be a mistake to use these *SEM*s to predict the stability of IQs for clients suspected of having intellectual disability. The stability coefficients given in the Wechsler test manuals were obtained using participants in the normal range of intellectual ability and, therefore, may not necessarily apply to individuals with low intellectual ability (cf. Anastasi & Urbina, 1997). To ascertain the stability of tests of intelligence in the low-ability range, one must look at studies that specifically test this. The purpose of this study was to conduct a meta-analysis of these studies.

Method

A search of the literature was made for studies reporting on the stability of low IQ. The following inclusion criteria were used: The mean FS IQ on at least one assessment was less than 80, the sample of clients included more than one diagnostic group, and the assessment used was well standardized, with good concurrent validity. Studies were not included if the participants were exclusively from a single diagnostic group where a general decline in intellectual ability over time may be expected: for example, Down syndrome (Carr, 1988), fragile X (Hodapp et al., 1990) syndrome, fetal alcohol syndrome (Streissguth, Herman, & Smith, 1978), or Lesch-Nyham syndrome (Matthews, Solan, Barabas, & Robey, 1999). Studies were also excluded if the assessment had been shown to have poor concurrent validity when used with participants with low intellectual ability, such as the Peabody Picture Vocabulary Test (Burnett, 1965; Pasewark, Fitzgerald, & Gloechler, 1971). An initial search was undertaken using the *Web of Science* and *Cambridge Scientific* databases, with the following search words: *intelligence*, *mental retardation*, and *learning disabilities*. Identified articles were read and any relevant references cited in them followed up on. This process was continued until no more new articles were found.

Results

In all, 18 studies were located involving 2,026 participants. Five of these studies (Goodman, 1976; Holowinsky, 1962; Reger, 1962; Thomas, 1980; Whatley & Plant, 1957) reported only on changes in mean IQ over time, whereas the others reported stability coefficients.

Change in Mean IQ

The average change in mean IQ (weighted for the number of participants in the studies) for all studies was −0.56 for the Verbal IQ (V IQ), 2.60 for the Performance IQ (P IQ) on the Wechsler tests (Wechsler, 1949, 1955, 1974, 1981, 1991) and 0.41 for FS IQ, which also included the Binet tests (Terman, 1960; Terman & Merrill, 1972). The changes in V IQ and FS IQ of less than one point were not clinically or, applying central limit theorem (cf. Howell, 1992), statistically significant. The 2.60 increase in P IQ was statistically significant ($p < .0001$), suggesting a systematic increase in P IQ and possibly indicating a systematic variance in the measurement of P IQ associated with repeated testing (possibly due to a practice effect and/or a genuine increase in P IQ over time).

Stability Coefficients

Table 1 shows the studies that reported stability coefficients in addition to the following: the number of people in the sample, the assessment used, the average interval between assessments, the mean age of the participants when first assessed, the mean FS IQ and *SD* when first assessed, and the mean

Table 1 Demographics and Stability Coefficients of Meta-Analytic Studies Reviewed

Study	*N*	Tests	Interval	Age	Initial IQ	IQ *SD*	Vr	Pr	FSr
Throne et al. (1962)	39	WISC	3.5	12	51.80	12.00	.92	.89	.95
Friedman (1970)	44	WISC	17	8	77.80	NG	.48	.78	.68
Spitz (1983)	69	WISC	25	12	58.71	8.74	NG	NG	.75
Wesner (1973)	51	WISC	19	NG	63.71	13.59	.88	.89	.89
Walker & Gross (1970)	49	WISC	35	11	68.00	6.50	.70	.73	.76
Elliott et al. (1985)	382	WISC-R	36	12	77.10	15.1	.81	.78	.85
Naglieri & Pfeiffer (1983)	53	WISC-R	34	9	74.60	10.59	.54	.54	.56
Vance et al. (1981)	75	WISC-R	26	10	75.91	12.72	.80	.91	.88
Spitz (1983)	24	WISC-R	21	13	54.96	9.33	NG	NG	.84
Bolen (1998)	70	WISC-III	35	10	61.30	6.34	.68	.62	.73
Canivez & Watkins (2001)	66	WISC-III	34	10	63.00	9.88	.85	.90	.93
Spitz (1983)	42	WAIS	41	17	61.33	6.69	NG	NG	.75
Spitz (1983)	23	WAIS	42	21	61.48	8.56	NG	NG	.88
Watkins & Campbell (1992)	50	WAIS-R	34	30	55.48	6.03	.80	.76	.86
Keogh et al. (1997)	82	Binet	60	6	69.60	16.89	NG	NG	.85
Walker & Gross (1970)	29	Binet	33	7	62.00	5.80	NG	NG	.79
Silverstein (1982)	101	Binet	36	NG	65.73	9.76	NG	NG	.81
M	**73.47**		**31.26**	**12.53**	**64.85**	**10.15**	**.75**	**.78**	**.81**
Weighted *M*			**33.54**		**68.76**	**11.67**	**.77**	**.78**	**.82**

Note. These studies reported an intertest correlation (stability coefficient) for IQs together with the number of participants in the study (*N*), the test that was used (WISC = Wechsler Intelligence Scale for Children [Wechsler, 2003]; WAIS = Wechsler Adult Intelligence Scale [Wechsler, 1997]; WISC-R = Wechsler Intelligence Scale for Children-Revised; WAIS-R = Wechsler Adult Intelligence Scale-Revised; WISC-III = Wechsler Intelligence Scale for Children, 3rd Edition; Binet = Stanford-Binet Test), the interassessment interval in months, the mean age of the participants at the time of the first assessment in years, the initial mean Full-Scale IQ, standard deviation of the initial IQ (IQ *SD*), the correlation for Verbal IQ (Vr), the correlation for Performance IQ (Pr), and the correlation for Full-Scale IQ (FSr). The weighted means are the means weighted for the number of participants in the study. NG = not given. Note that the *N* of 66 given for Canivez and Watkins (2001) differs from the *N* of 60 given for the same study in Table 2.

stability coefficient (weighted for number of participants in the sample), for V IQ, P IQ, and FS IQ across the studies. To identify factors that may have influenced the stability of low IQ, these coefficients were correlated with the following variables: group size, the test–retest interval, the initial average age of the participants, and the initial IQ of the participants. None of these correlations approached statistical significance.

Anastasi and Urbina (1997) suggested that reliability correlations should be .80 or above. Although many of the stability coefficients fell below this level, the weighted-mean correlations were .77 (range = .48–.92), .78 (range = .54–.91), and .82 (range = .56–.95) for V IQ, P IQ, and FS IQ (including the studies using Binet scales; Terman & Merrill, 1972), respectively. This suggests reasonable stability for FS IQs and acceptable stability for V IQ and P IQ.

Changes in Subtest Scaled Scores

The six studies reporting stability coefficients on individual subtests in the Wechsler scales (Wechsler, 1949, 1974, 1981, 1991) are shown in Table 2. These coefficients were generally lower than those for IQ, with weighted-mean correlations ranging from .79 for the Digit Span subtest to .47 for the Comprehension subtest.

Table 2 Studies in Which Stability Confidences Were Reported for Subtests, Together With Studies' Demographics

Variable	Throne et al. (1962)	Naglieri & Pfeiffer (1983)	Vance et al. (1981)	Watkins & Campbell (1992)	Bolen (1998)	Canivez & Watkins (2001)	Weighted *M*
Test	WISC	WISC-R	WISC-R	WAIS-R	WISC-III	WISC-III	Stability
N	39	53	75	50	70	60	
Age	12 years	9 years	10 years	30 years	10 years	10 years	
Interval	3.5 months	34 months	26 months	34 months	35 months	34 months	
Vr (IQ)	.92 (57.7)	.54 (73.32)	.80 (76.63)	.80 (57.82)	.68 (64.38)	.85 (65.89)	
Pr (IQ)	.89 (54.3)	.54 (79.19)	.91 (78.72)	.76 (59.56)	.62 (64.49)	.90 (65.77)	
FSr (IQ)	.95 (51.8)	.56 (74.60)	.88 (75.91)	.86 (55.48)	.73 (61.30)	.93 (63.00)	
I	.72	.47	.69	.56	.57	.69	.62
S	.67	.41	.53	.40	.41	.48	.48
A	.80	.39	.56	.73	.50	.60	.58
V	.79	.55	.72	.54	.52	.57	.61
C	.83	.35	.59	.17	.26	.66	.47
DS	.76		.80	.78		.82	.79
PC	.84	.15	.59	.37	.47	.59	.49
PA	.67	.63	.67	.28	.42	.65	.55
BD	.82	.23	.73	.82	.44	.74	.62
OA	.74	.49	.80	.61	.52	.59	.63
CD	.83	.53	.77	.79	.55	.61	.67
VCI						.84 (67.93)	
POI						.87 (65.61)	
FDI						.81 (65.78)	
PSI						NG	

Note. Study demographics included the number of participants in the study (*N*), the test that was used (WISC = Wechsler Intelligence Scale for Children [Wechsler, 1949]; WAIS = Wechsler Adult Intelligence Scale [Wechsler, 1955]; WISC-R = Wechsler Intelligence Scale for Children- Revised [Wechsler,1955]; WAIS-R = Wechsler Adult Intelligence Scale-Revised [Wechsler, 1981]; WISC-III = Wechsler Intelligence Scale for Children 3rd Edition [Wechsler, 1991]), the interassessment interval in months, the mean age of the participants at the time of the first assessment in years, the initial mean Full-Scale IQ, the correlation for Verbal IQ (Vr), the correlation for Performance IQ (Pr), and the correlation for Full-Scale IQ (FSr), with the mean IQ when initially tested in brackets. The correlations, weighted for number of participants, of each subtest: I = Information, S = Similarities, A = Arithmetic, V = Vocabulary, C = Comprehension, DS = Digit Span, PC = Picture Completion, PA = Picture Arrangement, BD = Block Design, OA = Object Assembly, CD = Coding/ Digit Symbol. NG = not given. Note that the *N* of 60 given for Canivez and Watkins (2001) differs from the *N* of 66 given for the same study in Tables 1 and 3, and is the highest *N* for any subtest applying to S and PA, only; *N* for the other subtests were as follows: 59 for PC, I, A, BD, and V; 58 for CD, OA, and C; and 41 for DS.

The Proportion of Clients Whose IQ Scores Changed

Table 3 shows the 11 studies that gave information on the proportion of clients whose IQ changed by specific amounts between assessments. From these data, it is apparent that most IQs changed relatively little, with a weighted mean of 57% of IQs changing by less than 6 points. However, a weighted mean of 14% of IQs changed by 10 points or more.

Table 3 Studies That Showed the Percentage of Clients Who Changed by Specified Number of IQ Points, Together With Study Demographics

Study	*N*	Interval	Test	Range	% change
Spitz (1983)	69	25	WISC	<6	59
				6–10	28
				11–15	13
Walker & Gross (1970)	49	35	WISC	<5	57
				5–9	30
				10–14	10
				15–20	2
Whatley & Plant (1957)	70	17	WISC	<6	50
				6–10	26
				11–15	14
				16–20	7
				21–25	3
Elliott et al. (1985)	382	36	WISC-R	<6	54
				6–10	26
				>10	20
Spitz (1983)	24	21	WISC-R	<6	54
				6–9	29
				11–15	8
				16	8
Canivez & Watkins (2001)	66	35	WISC-III	<3.2	47
				3.2–6.4	26
				6.4–9.6	17
				>9.6	11
Spitz (1983): Young group	42	41	WAIS	<6	67
				6–10	31
				14	2
Spitz (1983): Older group	23	42	WAIS	<6	74
				6–10	26
Watkins & Campbell (1992)	50	34	WAIS-R	<6	88
				>5	12
Walker & Gross (1970)	29	33	Binet	<5	55
				5–9	31
				10–14	14
Silverstein (1982)	101	24	Binet	<3	33
				3–4	23
				5–6	18
				7–8	11
				9–10	7
				11–12	5

Note. Study demographics include the number of participants in the study (*N*) and the interassessment interval in months. The *N* of 66 given for Canivez and Watkins (2001) differs from the *N* of 60 given for the same study in Table 2. WISC = Wechsler Intelligence Scale for Children (Wechsler, 1949); WAIS = Wechsler Adult Intelligence Scale (Wechsler, 1955); WISC-R = Wechsler Intelligence Scale for Children- Revised (Wechsler, 1974); WAIS-R = Wechsler Adult Intelligence Scale-Revised (Wechsler, 1981); WISC-III = Wechsler Intelligence Scale for Children 3rd Edition (Wechsler, 1991); Binet = Stanford-Binet Test (Terman, 1960; Terman & Merrill, 1972).

Discussion

The purpose of this meta-analysis was to find a mean stability coefficient of low IQ that could be used to estimate likely changes in measured IQ between assessments. Weighted-mean stability coefficients of .77 for V IQ, .78 for P IQ, and .82 for FS IQ were found that corresponded to *SEM*s of 7.2,

7.0, and 6.4 for V IQ, P IQ, and FS IQ, respectively. For the mean *SEM* for FS IQ of 6.4, one would expect that 61% of FS IQs would change by less than 6 points and 13% by 10 points or more. This estimate is similar to that found by those studies that reported on the proportion of clients whose IQ changed by specific amounts, where a weighted average of 57% of IQs changed by less than 6 points and 14% of IQs changed by 10 points or more. This similarity suggests that mean *SEM* for stability provides a reasonable estimate of probability of IQ change by specific amounts between assessments.

The subtest scores were less reliable than the IQs. None of the subtests had a weighted-mean stability coefficient of more than .8, the figure suggested by Anastasi and Urbina (1997) as the minimum acceptable reliability coefficient. Similarities, Comprehension and Picture Completion subtests were below .5. It is not surprising that the subtests were less reliable than the IQs, given that IQ scores are a function of the client's performance on several subtests. Nonetheless a subtest score obtained some time ago cannot be regarded as a very accurate predictor of the score the client would obtain today. Therefore, subtest scores should only be interpreted with caution.

In broad terms, there are probably three basic reasons why an individual's tested IQ may change over several years. First, random error due to small changes in factors, such as test administration and scoring, the level of distraction present, and the state of the person being assessed on the day, may result in variations in IQ around an unchanged mean. Second, there may be some systematic error, whereby scores change consistently in a particular direction, resulting in a change in mean IQ. Although a significant change in mean IQ was not found for V IQ and FS IQ, there was a significant increase in P IQ of 2.60 points. The most obvious source of this error is a practice effect, whereby the client does better on the test the second time, having practiced it on the first occasion. However, an additional factor may be the Flynn effect (Flynn 1985): the tendency for measured IQ to increase by approximately 0.3 points a year, though more so in the P IQ than V IQ. A third possible reason for a change in measured IQ over time is an actual change in a client's intellectual ability. The studies considered here had an average test–retest interval of just under 3 years, which may be sufficient time for factors such as changes in quality of education, change in diet, or intellectual stimulation to affect an individual's intellectual ability. If the change in IQ was largely due to a genuine change in ability, more recent tests must be considered more accurate estimates of an individual's true IQ. However, if the change is mainly due to error, more recent tests cannot be considered significantly more accurate than those carried out some time ago. Unfortunately, the data that were available here did not allow a full analysis of the degree to which change was due to error or to genuine change in ability, though the failures to get significant correlations between the test–retest interval and stability coefficient of the studies suggest that change was mainly due to error.

These findings have implications for our interpretation of the results of IQ assessment and any diagnosis made on the basis of them. It is typical in presenting the results of IQ tests to give a 95% confidence interval, that is, the range of scores in which the client's true IQ has a 95% chance of falling. 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. The 95% confidence intervals for Wechsler (1949, 1955, 1974, 1981, 1991) assessments are about 5 points. However, these are based on the *SEM* for internal consistency. If the 95% confidence interval is calculated using the stability *SEM* for FS IQ of 6.4, it is 25.08 points, or 12.54 points on either side of the obtained IQ. This would mean that an individual would have to obtain an IQ above 82 before one could be 95% confident that he/she had an IQ above 70, or an IQ below 57 before one could be 95% confident that he/she had an IQ below 70. It follows from this that if one is simply going to use information from a single IQ assessment done some years ago, one cannot be 95% certain of a diagnosis of intellectual disability unless the IQ is less than 57. Similarly, one could not say with 95% confidence that a client should not have a diagnosis of intellectual disability unless his/her IQ was above 82. However, IQ results do not always come in isolation. Information may be available that could suggest that an obtained IQ is likely to be a low or high estimate of an individual's true IQ. A brief discussion of this may be helpful to the reader. As noted above, the lack of a significant correlation between the test–retest interval and stability coefficient suggests that changes in IQ scores are mainly due to error, rather than genuine change in intellectual ability. Therefore, the change in IQ scores is due to changes in the environment and nonin-

tellectual factors in the client between the two assessments. These changes are likely to be in variables such as degree of distraction in the environment, the level of cooperation of the client, minor illness or fatigue on the part of the client, or variation in how the test was administered and scored. Most of these factors will have the effect of reducing IQ. It is also likely that they would only have occurred to a minimum extent when the test was standardized, because the participants would be expected to be well motivated and in good health and efforts would be made to give the test under optimal conditions. However, these IQ-reducing factors may well have occurred in many of the assessments in the studies used in this meta-analysis as well as those done in normal clinical practice. It follows, therefore, that assessments done as part of normal clinical practice may tend to underestimate a client's true IQ. Therefore, any information with regard to the conditions under which the test was given should be examined. If these circumstances were suboptimal, an obtained IQ may be a low estimate of the individual's true IQ. An additional factor that could support the accuracy of an IQ is a second IQ score. This may be available or could be obtained by reassessing an individual. Assuming the second IQ was obtained using the same test within a few years of the first, two similar IQs could be regarded as confirming each other. If there was a large difference between the two IQ scores, it would be likely that the higher result was the more accurate, as the nonintellectual factors that affect IQ scores will tend to reduce them; therefore, lower scores will on average have been subject to more error and be less accurate. Other sources of information regarding the likely accuracy of an IQ result come from how the individual functions in his/her environment. If he/she is coping with intellectual demands in everyday life—for example, if he/she is able to budget or read and understand complex information and yet has an IQ of 55—it is likely that this IQ is a low estimate of the true IQ. If the person is failing to function with such tasks and has an IQ of 80, it may be a high estimate of the person's true intellectual ability. However, none of this information can be used quantitatively to reduce margin of error in an IQ score by a specific amount. Therefore, an IQ figure can only be regarded as a guide to a client's true IQ. Because of this, it seems unreasonable to have a specific IQ figure below which somebody could be regarded as having intellectual disability.

The analysis in this article may be subject to a number of sources of error. First, there were very few studies, which meant the meta-analysis had to be done by combining the data from different assessments, not only different versions of the Wechsler (1949, 1955, 1974, 1981, 1991, 1997) assessments but also Binet assessments (Terman, 1960; Terman & Merrill, 1972). However, as the stability coefficients for these different assessments do not differ greatly, this does not seem unreasonable. Second, means for the client's initial age, IQ, and the test–retest interval were reported and used in calculation. Although no relationships were found between the test–retest interval, age of client, or initial IQ and the correlation between tests, it is possible that using meaned data obscured effects that were happening at the extremes of the data set. Third, stability coefficients reported in Table 1 and used to calculate *SEM*s were not corrected for the restricted range of IQ scores in the studies. In addition, when calculating *SEM*s, an *SD* of 15, which was the *SD* for the population as a whole, was used rather than the *SD* of the sample. Both these could be argued to have resulted in the lower stability coefficients and the relatively large SEMs. However, the concern here was with predicting change in IQ of people with low IQs, which is a restricted sample of the whole population and will almost inevitably have a smaller range of IQs. The mean *SEM* of 6.4, which was calculated using uncorrected stability coefficients, and an *SD* of 15 gave an estimate of the percentages of IQs that would change by specific amounts, similar to that found in the studies that reported on this. Nonetheless, it may be of interest to some readers to see what the corrected stability coefficient and *SEM* would be. Guilford and Fruchter (1978) provided a formula for correcting correlations for restricted range based on the *SD* found in the population as a whole (in this case, 15) and the *SD* of the restricted sample (in this case, the weighted-mean *SD* of 11.63), as shown in Table 1. This gives a corrected, weighted-mean stability coefficient of .88, which corresponds to an *SEM* of 4.03, which is still substantially higher than the reliability and *SEM* reported in the test manuals.

It is clear that additional work needs to be done on the stability of the assessment of low intellectual ability. However, until this is done, the above analysis of the studies that are available is the best description of the stability of low IQ.

References

American Association on Mental Retardation. (2002). *Mental retardation: Definition, classification, and system of supports* (10th ed.). Washington, DC: Author.

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders—Text revision*. Washington, DC: Author.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Binet, A., & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique, 11,* 191–244.

Bolen, L. M. (1998). WISC-III score changes for EMH students. *Psychology in the Schools, 35,* 327–332.

British Psychological Society. (2001). *Learning disability: Definitions and contexts*. Leicester, United Kingdom: Author.

Burnett, A. (1965). Comparison of the PPVT, Wechsler-Bellevue, and Stanford-Binet on educable mentally retarded. *American Journal of Mental Deficiency, 69,* 712–715.

Canivez, G. L., & Watkins, M. W. (2001). Long-term stability of the Wechsler Intelligences Scale for Children—Third Edition among students and disabilities. *School Psychology Review, 30,* 438–453.

Carr, J. (1988). Six weeks to twenty-one years old: A longitudinal study of children with Down's syndrome, and their families. *Journal of Child Psychology and Child Psychiatry, 29,* 407–431.

Elliott, S. N., Piersel, W. C., Witt, J. C., Argulewicz, E. N., Gutkin, T. B., & Galvin, G. A. (1985). Three-year stability of the WISC-R IQs for handicapped children from three racial groups. *Journal of Psychoeducational Assessment, 3,* 233–244.

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

Friedman, R. (1970). Reliability of the Wechsler Intelligence Scale for Children in a group of mentally retarded children. *Journal of Clinical Psychology, 26,* 181–182.

Goodman, J. F. (1976). Ageing and IQ changes in institutionalized mentally retarded. *Psychological Reports, 39,* 999–1006.

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (6th ed.). New York: McGraw-Hill.

Hodapp, R. M., Dykens, E. M., Hagerman, R. J., Schreiner, R., Lachiewicz, A. M., & Leckman, J. F. (1990). Developmental implications of changing trajectories of IQ in males with fragile X syndrome. *Journal of the American Academy or Child and Adolescent Psychiatry, 29,* 214–219.

Holowinsky, I. (1962). IQ consistency in a group of institutionalized mental defectives over a period of 3 decades. *Training School Bulletin, 59,* 15–17.

Howell, D. C. (1992). *Statistical methods for psychologists* (3rd ed.). Belmont, CA: Duxbury Press.

Keogh, B. K., Bernheimer, L. P., & Guthrie, D. (1997). Stability and change over time in cognitive level of children with delays. *American Journal on Mental Retardation, 101,* 365–373.

Matthews, W. S., Solan, A., Barabas, G., & Robey, K. (1999). Cognitive functioning in Lesch-Nyhan syndrome: A 4-year follow-up study. *Developmental Medicine and Child Neurology, 41,* 260–262.

Naglieri, J. A., & Pfeiffer, S. I. (1983). Reliability and stability of the WISC-R for children with below average IQs. *Educational and Psychological Research, 3,* 203–208.

Pasewark, R. A., Fitzgerald, B. J., & Gloechler, T. (1971). Relationship of Peabody Picture Vocabulary Test and the Wechsler Intelligence Scale for Children in an educable retarded group: A cautionary note. *Psychological Reports, 28,* 405–406.

Reger, R. (1962). Repeated measurement with the WISC. *Psychological Reports, 11,* 418.

Silverstein, A. B. (1982). Note on the constancy of the IQ. *American Journal of Mental Deficiency, 87,* 227–228.

Spitz, H. H. (1983). Intratest and intertest reliability and stability of the WISC, WISC-R and WAIS full scale IQs in a mentally retarded population. *Journal of Special Education, 17,* 69–80.

Streissguth, A. P., Herman, C. S., & Smith, D. W. (1978). Stability of intelligence in the fetal alcohol syndrome: A preliminary report. *Alcoholism: Clinical Experimental Research, 2,* 165–170.

Terman, L. M. (1960). *Stanford-Binet Intelligence Scales*. Circle Pines, MN: American Guidance Service.

Terman, L. M., & Merrill, M. A. (1972). *Standford-*

Binet Intelligence Scale: Manual for the third revision: Form L-M. Boston: Houghton-Mifflin.

Thomas, P. J. (1980). A longitudinal comparison of the WISC and WISC-R with special education pupils. *Psychology in the Schools, 17,* 437–441.

Throne, F. M., Schulman, J. L., & Kaspar, J. C. (1962). Reliability and stability of the Wechsler Intelligence Scale for Children for a group of mentally retarded boys. *American Journal of Mental Deficiency, 67,* 455–457.

Vance, H. B., Blixt, S., Ellis, R., & Debell, S. (1981). Stability of the WISC-R for a sample of exceptional children. *Journal of Clinical Psychology, 37,* 397–399.

Walker, K. P., & Gross, F. L. (1970). I.Q. stability among educable mentally retarded children. *Training School Bulletin, 66,* 181–187.

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

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children*. New York: The Psychological Corporation.

Wechsler, D. (1955). *Wechsler Adult Intelligence Scale*. New York: The Psychological Corporation.

Wechsler, D. (1974). *Wechsler Adult Intelligence Scale—Revised*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale—Revised*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children—Third edition*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *WAIS-III, WMS-III, technical manual*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003). *WISC-IV, technical and interpretive manual*. San Antonio, TX: Harcourt Associates.

Wesner, C. E. (1973). The relationship between WISC and WAIS IQs with educable mentally retarded adolescents. *Educational and Psychological Measurement, 33,* 465–467.

Whatley, R. G., & Plant, W. T. (1957). The stability of W.I.S.C. IQs for selected children. *Journal of Psychology, 44,* 165–167.

Received 4/11/06, first decision 5/24/07, accepted 7/10/07.
Editor-in-Charge: Steven J. Taylor

Author:
Simon Whitaker, BSc, PhD, Dip Clin Psych, Consultant Clinical Psychologist and Senior Visiting Research Fellow, Univerity of Huddersfield, Learning Disability Research Unit, Queensgate, Huddersfield, W. Yorkshire HD1 3DH, United Kingdom. E-mail: s.whitaker@hud.ac.uk

