Wechsler Intelligence Tests: Do We Really Have a Criterion of Mental Retardation?

JAMES R. FLYNN

The University of Otago (Dunedin, New Zealand)

The definition of mental retardation offered by the American Association on Mental Deficiency refers to an IQ of approximately 70 or below. This is identical to the Wechsler criterion of a test performance two standard deviations (*SD*s) below the population mean. In fact, Wechsler tests have not supplied such a criterion, rather they have deviated from it by anything from .27 to a full *SD*. Having done without such a criterion for 40 years, we should consider exchanging it for one that is fixed in time and whose external validity is attested to by an accumulated body of evidence.

The American Association on Mental Deficiency defines mental retardation in terms of both impaired adaptive behavior and "significantly subaverage" general intellectual functioning, the latter term being defined as IQ of 70 or below (Grossman, 1983, p. 1). The fact that 70 represents the Wechsler cutting line for two standard deviations (*SD*s) below the mean is presumably not coincidental; and presumably the Association would not cite Wechsler tests so frequently if they did not believe that they give us a reasonably accurate picture of a subject's standing with reference to the general population of America.

In this paper I have attempted to show that Wechsler scales can be grossly unrepresentative of the general population and that psychologists who use them to classify subjects as 2 *SD*s below the mean may err by anything from .27 to a full *SD*. More specifically: (a) The Wechsler Intelligence Scale for Children (WISC) may have been about .55 *SD*s too lenient in its criterion of mental retardation by the time it was supplanted; (b) The WISC-R (revised) may be as much as .27 *SD*s too lenient at the present time; (c) The Wechsler Adult Intelligence Scale (WAIS) may have been a full

SD too lenient by the time it was supplanted and the WAIS-R equally misleading the day it was published. Finally, I have argued that these conclusions follow when the sampling procedures of the Psychological Corporation, the publisher of the Wechsler scales, are given the maximum benefit of the doubt.

Method

The basic method was to analyze all studies in which two Wechsler tests were administered to the same subjects in counterbalanced order, although there were certain complications and departures as will be discussed later. The rationale behind this method can be best explained by focusing on the WISC and WISC-R. Assume for the moment that subjects consistently get much higher scores on the WISC and that this remains true even after the standardization samples of the two tests are rendered comparable in character. Two alternative explanations are possible.

First, both the WISC and WISC-R have norms based on standardization samples that were grossly unrepresentative of their time, or at least one does, and there is no way of establishing which; or second, both the WISC and WISC-R have norms representative of their times, in which case Americans in general have been making IQ

gains at a rate of over .300 IQ points per year. After all, if subjects find the WISC normed in 1947–1948 (Wechsler, 1949) easier than the WISC-R normed in 1971–1973 (Wechsler, 1974) and the principal reason is that a representative WISC sample set a standard 7.68 points below a comparable WISC-R sample, then Americans' performance improved by that amount over 24.5 years. This would mean that WISC norms rapidly became obsolescent and were not representative of Americans in general during the larger part of the time it was used.

Results for WISC and WISC-R

Table 1 presents results from 29 studies and 1,607 subjects; it includes 33 samples whose mean WISC IQs ranged from 56 to 136. These samples were grouped by IQ level. As the weighted average shows, subjects appear to have found the WISC 8.54 IQ points easier than the WISC-R. This result cannot be accepted at face value, and Table 1 requires some words of clarification.

First, recall the possibility that the WISC and WISC-R standardization samples were

TABLE 1
WISC AND WISC-R: DIFFERENTIAL DIFFICULTY AS REVEALED BY UNIFORM SCORING

Source	N of subjects	WISC	WISC-R	WISC-R(T)[a]	Difference[b]
Solly (1977)	12	136.08	123.67	122.94	13.14
Larrabee & Holroyd (1976)	38	132.00	122.60	121.79	10.21
Tuma et al. (1978)	18	127.60	124.70	124.04	3.56
Wheaton et al. (1980)	50	127.24	122.84	122.05	5.19
Appelbaum & Tuma (1977)	20	125.21	123.42	122.67	2.54
Schwarting (1976)	58	113.40	105.91	103.91	9.49
Rowe (1976)	128	109.92	103.30	101.11	8.81
Stokes et al. (1978)	59	109.83	107.32	105.42	4.41
Davis (1977)	54	107.45	99.80	97.36	10.09
Klinge et al. (1976)	32	98.72	95.33	92.57	6.15
Appelbaum & Tuma (1977)	20	98.25	94.30	91.47	6.78
Brooks (1977)	30	96.40	89.17	85.48	10.42
Tuma et al. (1978)	18	96.20	91.00	87.94	8.26
Munford (1978)	20	92.50	87.50	84.19	8.31
Swerdlik (1978)	164	91.33	85.86	82.43	8.90
Hartlage & Boone (1977)	42	90.93	85.86	82.43	8.50
Weiner & Kaufman (1979)	46	89.70	81.70	77.97	11.73
Covin (1977)	30	89.33	89.40	86.22	3.11
Solway et al. (1976)	180	84.74	79.57	75.69	9.05
Reynolds & Hartlage (1979)	66	84.06	79.41	75.52	8.54
Thomas (1980)	93	80.97*	75.85	71.70	9.27
Sherrets & Quattrocchi (1979)	28	78.43	75.00	70.79	7.64
Reschly & Davis (1977)	48	76.65	73.04	68.24*	8.41
Covin (1976)	101	76.63	74.00	69.12*	7.51
Rowe (1977)	22	76.39	74.85	70.63	5.76
Solly (1977)	12	76.25	65.42	60.53	15.72
Hamm et al. (1976)	48	70.41	62.85	59.65*	10.76
McGinley (1981)	21	68.05	61.57	57.04*	11.01
Udziela & Barclay (1983)	45	66.79	64.80	59.86	6.93
Catron & Catron (1977)	29	65.30	59.66	54.36	10.94
Gironda (1977)	20	63.90	64.00	58.11*	5.79
Spitz (1983)	33	61.42	56.30	50.07*	11.35
Rowe (1977)	22	56.23	48.79	42.71	13.52
Total	1607				8.54[c]

[a] WISC-R(T) is the result of translating WISC-R scores into a uniform scoring convention described in the text; scores marked with an asterisk have had a slight additional adjustment also described in the text.
[b] Difference equals WISC minus WISC-R(T) and represents the greater difficulty of the WISC-R as expressed in IQ points.
[c] Weighted average.

representative of their time. The WISC was standardized on a sample of white Americans only, with the mean and *SD* set at 100 and 15, respectively; therefore, the rationale of a WISC score of 70 as a criterion of mental retardation is that this score puts a subject at two *SD*s below the mean IQ of white Americans. In order to calculate what a WISC score of 70 really meant by the early 1970s, how far below the mean of white Americans that score really was by then, I used the WISC-R (normed 1971–1973) to make such a calculation in Table 1; however, the WISC-R was normed on an all races' standardization sample inclusive of groups that score below the white American average. That is why within the WISC-R sample, the white subjects' mean is 102.26 and the white subjects' *SD* is 14 (Kaufman & Doppelt, 1976). This implies that WISC-R scores do not mean the same thing as WISC scores; e.g., a WISC-R score of 70 is clearly more than two *SD*s below the mean of white Americans.

In Table 1, this problem was solved by using a uniform scoring convention: The WISC-R scores were equated with WISC-R(T), where the (T) stands for translated; these are what WISC-R scores become when they are put on a white-subjects'-only scale to make them comparable to WISC scores. As for the mechanics, WISC-R 70 becomes WISC-R(T) 65.44: WISC-R 70 is 2.304 white subjects' *SD*s below the white subjects' mean (102.26 – 70 = 32.26, which divided by 14 = 2.304); the score that is 2.304 *SD*s below the WISC mean for white Americans is 65.44 (100 – 65.44 = 34.56, which divided by 15 = 2.304). Once again, this translation is essential: The whole purpose of using the WISC-R is to measure how deceptive WISC scores had become in the 1970s if interpreted in terms of the WISC manual, in terms of its own scoring convention and its own criterion of mental retardation. The translated scores reveal that the WISC-R standardization sample outperformed the WISC standardization sample by 8.54 IQ points.

The WISC sample, however, suffered from a handicap; it included 2.5% institutionalized mentally retarded children, a group absent from the WISC-R sample. Fortunately, Seashore, Wesman, and Doppelt (1950) allowed determination of the percentage of such subjects at each level towards the bottom of a normal curve. It is then possible to calculate how much the mean of that curve would be elevated by their elimination, namely, .057 of an *SD* or .86 IQ points. If that amount is deducted from the weighted average of 8.54 points from Table 1, 7.68 IQ points are left as the adjusted estimate of how much the WISC-R sample of 1972 outperformed the WISC sample of 1947–1948.

This conclusion assumes that counterbalanced administration is sufficient to control for practice effects and that assumption has been questioned in the case of these two tests (Thomas, 1980). The WISC-R differs from the WISC in that the examiner is instructed to present the correct solution or procedure if the subject fails the first item of a subtest. The hypothesis is that this makes for differential practice effects: When the WISC is administered first, the subject goes to the WISC-R with the sole benefit of having taken a test very similar in content; but when the WISC-R is administered first, the subject goes to the WISC with that benefit plus having received "coaching" by the examiner as well, an extra advantage not envisaged when the WISC was normed. Assume that subjects get higher scores on the WISC; a differential practice effect would inflate half of the WISC scores, and, therefore, at least part of the lesser difficulty of the WISC would be an artifact of the research design.

Because of this, I included four studies that departed from the usual research design in a way that would test the hypothesis of differential practice effects. Catron and Catron (1977) used subjects who were randomly assigned to take either the WISC or the WISC-R. Udziela and Barclay (1983) randomly selected students from the records of the Department for the Mentally Retarded in St. Louis for the years 1974 to 1975 and found that about half had taken the WISC only, the other half the WISC-R only. Solway et al. (1976) used subjects who took either the WISC or WISC-R and

were considered equivalent because they were all referred for testing from the same pool of juvenile delinquents. Thomas (1980) compared subjects who took the WISC 3 years after an initial test on the WISC (Group 1) with subjects who took the WISC-R 3 years after an initial test on the WISC (Group 2), that period being deemed sufficient to nullify practice effects. The difference between the two groups on their initial test when all took the WISC (Group 1 was 3.33 points inferior) can be used to control for quality (add 3.33 points on to Group 1's later performance) and afford a fair comparison.

A second reason for departure from the criterion of counterbalanced administration had to do with the dearth of studies of subjects with WISC IQs below 80, the very level that concerns mental retardation workers most. It seemed best to include some additional studies of low-IQ subjects whose research design posed problems, but relatively manageable problems. Hamm et al. (1976) used a counterbalanced research design, but 34 subjects took the WISC-R first and only 14 subjects the WISC. The average test–retest interval was 39 days, and even if practice effects did not differ with test order, they would in this case inflate the WISC mean. To compensate, I increased the WISC-R mean by 1.87 points, a value that assumes a practice effect of 4.485 points; that is the average practice effect for these two tests based on data from six studies (Davis, 1977; Klinge, Rodziewicz, & Schwartz, 1976; Larrabee & Holroyd, 1976; Swerdlik, 1978; Tuma, Appelbaum, & Bee, 1978; Wheaton, Vandergriff, & Nelson, 1980). For exactly the same sort of reasons, the WISC-R mean found by McGinley (1981) was increased by .64 points. There are also four studies in which subjects simply took the WISC and anywhere from 1.5 to 3 years later, took the WISC-R (Covin, 1976; Gironda, 1977; Reschly & Davis, 1977; Spitz, 1983). These studies possess both an advantage and a disadvantage over counterbalanced administration: The test–retest interval was long enough so that practice effects, whether differential or not, should pose little problem; but it was also long enough so that the

subjects may not have maintained a constant IQ from one test to the other. The best I could do was to apply the conclusion suggested by the possibility that the WISC and WISC-R standardization samples were each representative of its time, namely, that Americans have been making IQ gains at a rate of about .300 points per year. This meant deducting from .45 to .90 points from the WISC-R scores. Whether or not the conclusion is true, these adjustments are clearly too small to have a significant effect.

I have accounted for all of the studies and samples in Table 1. To test the hypothesis of differential practice effects (i.e., determine whether counterbalanced administration is sufficient to offer a fair comparison of the WISC and WISC-R), I divided the 33 samples into three kinds: 25 that had counterbalanced administration, 4 designed to circumvent the possibility of differential practice effects (Catron & Catron, 1977; Solway et al., 1976; Thomas, 1980; Udziela & Barclay, 1983), and 4 that mitigate that possibility, although posing other problems (Covin, 1976; Gironda, 1977; Reschley & Davis, 1977; Spitz, 1983). These three kinds of samples show that the WISC-R has a greater difficulty of 8.47 IQ points, 8.99 points, and 8.18 points, respectively. The consistency of these results means that the hypothesis of differential practice effects requires some hard evidence in its favor to gain credibility.

These 33 samples provide sufficient data for estimates of the differential difficulty of the WISC and WISC-R at virtually every level of the IQ scale. In Table 2, weighted averages were calculated for six levels, and these revealed a uniformity that is most striking: For those with WISC IQs from 70 to 115, the WISC-R is 8.5 points more difficult with minor fluctuations; the difference for gifted subjects was a bit less—just under 7 points, and the difference for those with a WISC IQ below 70 was a bit more—just over 9.5 points. A good estimate at the WISC cutting line for mental retardation would be 9.11 points; i.e., a performance that earned a WISC score of 70 would earn a WISC-R translated score of 60.89. If the .86 points needed to make the WISC-R a

fully accurate check on the WISC is deducted, Table 2 suggests that by 1972 the WISC cutting line for mental retardation was obsolescent by 8.25 points, or .55 SDs.

TABLE 2

WISC AND WISC-R: DIFFERENTIAL DIFFICULTY AT VARIOUS LEVELS OF PERFORMANCE

WISC range	WISC minus WISC-R(T)[a]	N subjects
125–140	6.67	138
115–125	—	—
105–115	8.30	299
100–105	—	—
90–100	8.52	326
80–90	8.89	415
70–80	8.52	259
55–70	9.70	170
Weighted average	8.54	1607[b]

[a] The greater difficulty of the WISC-R is expressed in IQ points based on a uniform scoring convention with a mean and SD of 100 and 15, respectively; the values are weighted averages for those samples whose subjects had a mean IQ within each WISC range (see Table 1 for sources and data).
[b] Total.

Discussion

Psychologists using the WISC from 1971 to 1974 thought that they were isolating subjects two SDs below the mean IQ of white Americans: This would be the bottom 2.27% of white persons and the bottom 16.90% of black persons. In reality, by that time the WISC cutting line of 70 was far too lenient in terms of current norms: They were actually isolating the bottom .54% of white persons and the bottom 5.89% of black persons. In terms of all races combined, the difference is between 4.32% and 1.29%, which means that of every 10 subjects who should have been classified as within the range of mental retardation only the bottom 3 were! In terms of numbers, of 8.8 million Americans eligible for classification, only 2.6 million would have been deemed eligible. These figures are based on the WISC-R standardization data as compiled by Kaufman and Doppelt (1976). They cannot be taken literally because pathological cases distort the distribution at the very bottom of the IQ curve, but they do indicate the magnitude of the problem.

With the publication of the WISC-R in 1974 with its all races standardization sample, the Psychological Corporation introduced a new criterion of mental retardation: no longer two SDs below the mean IQ of white Americans, but two SDs below the mean for all races. The criterion remained at a score of 70, but this masked a significant difference. As previously discussed, a WISC-R score of 70 is actually 2.304 SDs below the mean of white Americans when put on the WISC scale, and this amounts to 4.56 IQ points (.304 × 15 = 4.56). In the WISC-R manual, Wechsler (1974) stated that his scale provides "a time-tested classification of IQ equivalents for diagnostic terms in common use" (p. 26). Clearly, the old criterion and a new one 4.56 IQ points more lenient cannot both be attested to by the same body of accumulated evidence.

How well does the WISC-R today do what it may have done in 1972, namely, isolate Americans of all races two SDs below the all races' mean? Assuming the WISC and WISC-R standardization samples were each representative of their time, American IQ gains over the 24.5 years prior to 1972 relaxed the WISC criterion of mental retardation by .55 SDs; therefore, if IQ gains have continued since 1972 at the same rate, the WISC-R criterion may have relaxed by .27 SDs, or over 4 IQ points during the 12 years from 1972 to 1984. I do not know whether IQ gains have continued or not, but that is the whole point. Assuming the radical malleability of IQ in the past, the Psychological Corporation would have to restandardize all of their tests every 7 years in order to be certain the norms were current. Recall that the alternative to the radical malleability of IQ is that their samples are grossly unrepresentative even at the time they are selected, in which case their norms should never be accepted as reliable.

Results for Other Wechsler Tests

The method already described, analyzing all studies in which two Wechsler tests were administered to the same subjects, was applied to other combinations of tests. There are 4 such studies for the Wechsler

Adult Intelligence Scale (WAIS) and WISC-R (Craft & Kronenburger, 1979; Nagle & Lazarus, 1979; Spitz, 1983; Wechsler, 1974), 8 for the WAIS and WAIS-R (Lippold & Claiborn, 1983; Mishra & Brown, 1983; Prout & Schwartz, 1984; Simon & Clopton, in press; Smith, 1983; Spitz, 1983; Urbina, Golden, & Ariel, 1982; Wechsler, 1981). Altogether, this afforded three combinations of paired Wechsler tests, and within each pair the tests were compared for differential difficulty. In addition, nonretarded and retarded subjects were separated so that I could establish whether the differential difficulty was similar for both average and below-average subjects.

Table 3 shows that the pair of WISC and WISC-R, with little separating their differential difficulty for nonretarded and retarded subjects, stands out as the exception. The WISC-R was about 6 points more difficult than the WAIS for most subjects, but on the level of retardation, its greater difficulty rose to almost 13 points. On the other hand, the greater difficulty of the WAIS-R over the WAIS, which was also about 6 points for most subjects, fell away to nothing at the level of retardation. The most interesting thing about Table 3 is a prediction that can be generated from the comparisons shown. If the WAIS-R and WAIS are equivalent at the level of retardation, and if the WAIS is 13 points easier than the WISC-R, then the WAIS-R should

be 13 points easier than the WISC-R. If that is true—if the WAIS-R is really radically more lenient at the level of retardation than is the WISC-R—the most serious consequences follow. First, however, I must try to explain why Wechsler tests exhibit differential difficulty for nonretarded and retarded subjects.

I believe that the procedures of the Psychological Corporation are basically sound, taking their samples as a whole and focusing on nonretarded subjects. As I have shown elsewhere (Flynn, 1984), when the improved performance of their standardization samples is put in the context of a wider array of data, it is far more compatible with the hypothesis that Americans have made massive IQ gains over time than with sampling error; but there is a special problem at the level of retardation due to the small number of subjects. Take two standardization samples both numbering 2,200 and reasonably matched for quality. If you have bad luck, the bottom 50 (2.27%) of the first sample will match the bottom 22 (1.00%) of the second; and if the mathematics of a normal curve are applied, a performance that scores 70 on the first test will receive only 65 on the second; i.e., even given sound procedures overall, sampling error at the bottom of the curve can account for a differential difficulty between nonretarded and retarded subjects amounting to 5 or 6 points. Further, at a particular age, for example, age 16, where sample numbers in the range of retardation are only four or five, chance may dictate even greater discrepancies.

Therefore, in addition to obsolescence as an unpredictable factor, for the rate of IQ gains is known only after the event, sampling error, which seems endemic at the level of mental retardation, must be added. From this perspective, Table 3 shows that by 1972 the WAIS was 6.20 points too lenient because of obsolescence and perhaps another 6.59 points too lenient because of sampling error to make up the total of 12.79 points. I say "perhaps" because, for all we know, the sampling error at the level of retardation may have come from the WISC-R, the other test in this pair. The WISC and WISC-R do seem to add credi-

TABLE 3

WECHSLER TESTS: DIFFERENTIAL DIFFICULTY FOR NONRETARDED AND RETARDED SUBJECTS

Test	Non-retarded	N	Retarded	N
WISC and WISC-R(T)	8.40	1437	9.70	170
WAIS and WISC-R	6.20	40	12.79	86
WAIS and WAIS-R	5.55	328	−.06	81

Note. For each pair of tests, the values refer to the greater difficulty of the second test expressed in IQ points. The WISC and WISC-R(T) pair have their differential difficulty expressed in our uniform scoring convention with white subjects' $SD = 15$, the remainder in a shared all races' convention with white subjects' $SD = 14$. Multiply by 1.07 to convert the latter into the former. For sources see Table 2 and text.

bility to each other's sampling at the level of retardation, but this may be due only to errors that happened to be in the same direction. If the WAIS was as defective as it seems to be, psychologists using it as a criterion of mental retardation in 1972 were using a criterion too lenient by over .85 *SD*s. By 1980, if IQ gains persisted, they were off by a full *SD*: Rather than isolating those 2 *SD*s below the mean, they were isolating subjects 3 *SD*s below; setting aside pathological cases, only one-seventh of 1% of the population was eligible to be classified as mentally retarded! Moreover, when the WAIS-R was published in 1981, it brought no improvement; the WAIS-R and the WAIS appear to be equivalent at the level of mental retardation.

This led to the prediction that the WAIS-R will prove to be 13 points more lenient than the WISC-R for retarded subjects. Imagine that the Psychological Corporation took two samples at the same time and that although well matched for overall quality, they showed great discrepancies at the level of retardation. The fact that they were taken at the same time rules out obsolescence as a factor; the match for overall quality indicates basically sound sampling procedures; but the great discrepancies for retarded subjects would signal beyond doubt the existence of a special problem at the bottom of the curve. Now the pair of WISC-R and WAIS-R come close to meeting the above specifications. The samples were tested only 6 years apart; one study at least shows that they were well-matched for quality (Wechsler, 1981); and yet preliminary reports from the field indicate that the WAIS-R is at least 13 points more lenient for retarded subjects (Ring, 1983). In my opinion, the pair of WAIS-R and WISC-R constitutes an *experimentum crucis* for the reliability of Wechsler criteria of mental retardation.

The pair of WAIS-R and WISC-R are of great value for another reason: They overlap only at age 16. As previously mentioned, sampling error should make a greater difference for retarded subjects at a particular age than for subjects of all ages as a group because the number of relevant persons in the standardization sample falls off to four or five. Note the significance of this for the pair WISC and WISC-R, the only Wechsler pair that seems to suggest no special problem at the level of retardation; for any given age the WISC and WISC-R sampling could be as defective as the WAIS-R, this fact being concealed only by merging all ages together. Every subject tested has a particular age, and it is no comfort to know that they would not have been wrongly classified if only they had been fictitious subjects each aged 6 to 15.

A final point about sampling error and Wechsler IQs. For IQs of 55 or below, the relationship between standardization samples and norms is at its weakest: These scores are 3 or more *SD*s below the mean and include only .14 of 1% of the sample, which means these subjects are virtually nonexistent in a sample of 2,200. At these levels, different scoring practices rather than sampling error cause IQ discrepancies from test to test. For example, a 16-year-old subject must actually earn the minimum WISC-R IQ of 40 by giving a fair number of correct answers; that same subject can then take the WAIS-R and get its minimum IQ of 46 (at ages 16 to 17) without giving any correct answers whatsoever (Wechsler, 1974, 1981). Indeed, the different methods of going from raw scores to scaled scores to IQs on Wechsler tests may be having a significant impact throughout the whole range of mental retardation, although little above it. On another level, however, this problem reflects the deeper problem of sampling. When the device of deviation IQs is used, large and accurate samples, such as we have for nonretarded subjects, tend to nullify the effects of different scoring practices because deviation IQs make sample distribution the primary determinent of IQs. When the sampling factor, however, is too weak to provide a corrective, the scoring practices are free to produce their discrepancies without interference.

Conclusion

I am uncertain that anyone will ever know just who is 2 *SD*s below the population mean, at least for a given age at a given time, but assume that the sampling prob-

lems are solved and that day arrives. For 40 years, without realizing it, psychologists have applied criteria far more lenient than 2 *SD*s below the mean and presumably found them reasonably satisfactory in terms of external validity. If accurate tests become available, rather than suddenly strengthening these criteria for no good reason, psychologists must consider setting the cutting-line for mental retardation at something below 2 *SD*s below the mean; e.g., adopt a score of 60 rather than 70 on the ideal test of the future. If IQ gains over time continue, psychologists will have to continue to be flexible. The next test published would be based on the higher quality of a new standardization sample, and its norms would force a performance that previously earned 60 down to 50. If the old level of performance continues to have external validity on its side (e.g., if 60 on the old test signals an ability to read at a functional level), it makes no sense to classify subjects as retarded merely because they get 50 on the new test. A score of 50 would have to become our cutting line.

Is it really important to determine who is so many *SD*s below the mean? If a score of 70 on the old WISC really was "time-tested" in terms of a body of evidence and evidence for its external validity continues to accumulate, what difference would it make how far below the population mean WISC 70 lies? Rather than revising our criterion of mental retardation with each new test, psychologists should fix it at a particular point in time, perhaps at the standards set by the WISC sample of 1947–1948 and the WAIS sample of 1953–1954. In practice, this would mean keeping a WAIS-R score of 70, because of its equivalence to the old WAIS, but lowering the cutting line for the WISC-R down to 65 or even 60 for older subjects, so as to establish equivalence with WISC and WAIS. Whenever a new test is published, it should be administered to a large sample of retarded subjects along with the appropriate reference test from the past, so as to derive a new score equivalent to the old cutting-line.

The soundness of these reflections de-

pends on future research, studies of the WISC-R and WAIS-R and, far more important, studies that determine whether or not we can freeze our criterion of mental retardation at a given point in time. This much is certain: Because of obsolescence and sampling problems, there is at present no coherent criterion of mental retardation.

References

Appelbaum, A. S., & Tuma, J. M. (1977). Social class and test performance: Comparative validity of the Peabody with the WISC and WISC-R for two socio-economic groups. *Psychological Reports, 40*, 139–145.

Brooks, C. R. (1977). WISC, WISC-R, S-B L & M, WRAT: Relationships and trends among children ages six to ten referred for psychological evaluation. *Psychology in the Schools, 14*, 30–33.

Catron, D. W., & Catron, S. S. (1977). WISC vs. WISC-R: A comparison with educable mentally retarded children. *Journal of School Psychology, 15*, 264–266.

Covin, T. M. (1976). Comparability of WISC and WISC-R full scale IQs for elementary school children with learning difficulties. *Psychological Reports, 39*, 280.

Covin, T. M. (1977). Comparability of WISC and WISC-R scores for 30 8- and 9-year-old institutionalized Caucasian children. *Psychological Reports, 40*, 382.

Craft, N. P., & Kronenburger, E. J. (1979). Comparability of WISC-R and WAIS IQ scores in educable mentally handicapped adolescents. *Psychology in the Schools, 16*, 502–504.

Davis, E. E. (1977). Matched pair comparison of WISC and WISC-R scores. *Psychology in the Schools, 14*, 161–166.

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

Gironda, R. J. (1977). A comparison of WISC and WISC-R results of urban educable mentally retarded students. *Psychology in the Schools, 14*, 271–275.

Grossman, H. J. (Ed.). (1983). *Classification in mental retardation*. Wahsington, DC: American Association on Mental Deficiency.

Hamm, H., Wheeler, J., McCallum, S., Herrin, M., Hunter, D., & Catoe, C. (1976). A comparison between the WISC and WISC-R among educable mentally retarded students. *Psychology in the Schools, 13*, 4–8.

Hartlage, L. C., & Boone, K. E. (1977). Achievement test correlates of Wechsler Intelligence Scale for Children and Wechsler Intelligence Scale for Children—Revised. *Perceptual and Motor Skills, 45*, 1283–1286.

Kaufman, A. S., & Doppelt, J. E. (1976). Analysis of WISC-R standardization data in terms of the stratification variables. *Child Development, 47*, 165–171.

Klinge, V., Rodziewicz, T., & Schwartz, L. (1976).

Comparison of the WISC and WISC-R on a psychiatric adolescent inpatient sample. *Journal of Abnormal Child Psychology, 4,* 73–81.

Larrabee, G. J., & Holroyd, R. G. (1976). Comparison of WISC and WISC-R using a sample of highly intelligent students. *Psychological Reports, 38,* 1071–1074.

Lippold, S., & Claiborn, J. M. (1983). Comparison of the Wechsler Adult Intelligence Scale and the Wechsler Adult Intelligence Scale—Revised. *Journal of Consulting and Clinical Psychology, 51,* 315.

McGinley, P. (1981). A comparison of WISC and WISC-R test results. *The Irish Journal of Psychology, 1,* 23–24.

Mishra, S. P., & Brown, K. H. (1983). The comparability of WAIS and WAIS-R IQs and subtest scores. *Journal of Clinical Psychology, 39,* 754–757.

Munford, P. R. (1978). A comparison of the WISC and WISC-R on black child psychiatric outpatients. *Journal of Clinical Psychology, 34,* 938–943.

Nagle, R. J., & Lazarus, S. C. (1979). The comparability of the WISC-R and WAIS among 16-year-old EMR children. *Journal of School Psychology, 17,* 362–367.

Prout, H. T., & Schwartz, J. F. (1984). Validity of the Peabody Picture Vocabulary Test-Revised with mentally retarded adults. *Journal of Clinical Psychology, 40,* 584–587.

Reschly, D. J., & Davis, R. A. (1977). Comparability of WISC and WISC-R scores among borderline and mildly retarded children. *Journal of Clinical Psychology, 33,* 1045–1048.

Reynolds, C. R., & Hartlage, L. (1979). Comparison of WISC and WISC-R regression lines for academic prediction with black and white referred children. *Journal of Consulting and Clinical Psychology, 47,* 589–591.

Ring, B. C. (1983, Sept.). WISC-R/WAIS-R score differences: Are we WAIS-ting our time? *Oregon School Psychologists' Association Bulletin,* 4.

Rowe, H. A. H. (1976). *The comparability of the WISC and WISC-R.* Hawthorn, Victoria: Australian Council for Educational Research.

Rowe, H. A. H. (1977). 'Borderline' versus 'mentally deficient'. *The Australian Journal of Mental Retardation, 4,* 11–14.

Schwarting, F. G. (1976). A comparison of the WISC and WISC-R. *Psychology in the Schools, 13,* 139–141.

Seashore, H., Wesman, A., & Doppelt, J. (1950). The standardization of the Wechsler Intelligence Scale for Children. *Journal of Consulting Psychology, 14,* 99–110.

Sherrets, S., & Quattrocchi, M. (1979). WISC—WISC-R differences—Fact or artifact? *Journal of Pediatric Psychology, 4,* 119–127.

Simon, C. L., & Clopton, J. R. (in press). A comparison of WAIS and WAIS-R scores among mildly and moderately retarded adults. *American Journal of Mental Deficiency.*

Smith, R. S. (1983). A comparison study of the Wechsler Adult Intelligence Scale and the Wechsler Adult Intelligence Scale—Revised in a college population. *Journal of Consulting and Clinical Psychology, 51,* 414–419.

Solly, D. C. (1977). Comparison of WISC and WISC-R scores of mentally retarded and gifted children. *Journal of School Psychology, 15,* 255–258.

Solway, K. S., Fruge, E., Hays, J. R., Cody, J., & Gryll, S. (1976). A comparison of the WISC and WISC-R in a juvenile delinquent population. *Journal of Psychology, 94,* 101–106.

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. *The Journal of Special Education, 17,* 69–80.

Stokes, E. H., Brent, D., Huddleston, N. J., Rozier, J. S., & Marrero, B. (1978). A comparison of WISC and WISC-R scores of sixth grade students: Implications for validity. *Educational and Psychological Measurement, 38,* 469–473.

Swerdlik, M. E. (1978). Comparison of WISC and WISC-R scores of referred black, white and Latino children. *Journal of School Psychology, 16,* 110–125.

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

Tuma, J. M., Appelbaum, A. S., & Bee, D. E. (1978). Comparability of the WISC and the WISC-R in normal children of divergent socioeconomic backgrounds. *Psychology in the Schools, 15,* 339–346.

Udziela, A. D., & Barclay, A. G. (1983). A note on the WISC-WISC-R differences with mentally retarded and borderline functioning adolescents. *Psychology in the Schools, 20,* 27–28.

Urbina, S. P., Golden, C. J., & Ariel, R. N. (1982). WAIS/WAIS-R: Initial comparisons. *Clinical Neuropsychology, 4,* 145–146.

Wechsler, D. (1949). *WISC manual.* New York: Psychological Corp.

Wechsler, D. (1974). *WISC-R manual.* New York: Psychological Corp.

Wechsler, D. (1981). *WAIS-R manual.* New York: Psychological Corp.

Weiner, S. G., & Kaufman, A. S. (1979). WISC-R versus WISC for black children suspected of learning or behavioral disorders. *Journal of Learning Disabilities, 12,* 100–107.

Wheaton, P. J., Vandergriff, A. F., & Nelson, W. H. (1980). Comparability of the WISC and WISC-R with bright elementary school students. *Journal of School Psychology, 18,* 271–275.