New Zealand Journal of Psychology, 1987, 16, 28–36

WISC-R Item Characteristics: A Study with 11-year-old New Zealand Children*

Ross St. George
and
James W. Chapman
Department of Education, Massey University

The WISC-R is a widely used individual ability measure, and in New Zealand some 4,500 test forms are purchased annually. However, to date only one suitability study has been undertaken in New Zealand (Tuck, Hanson, & Zimmerman, 1975) and this was not a study of item characteristics. This study reports the WISC-R item characteristics of a sample of 11-year-old New Zealand children. Particular attention is given to the data on local item substitutions and to generally discrepant items. Gender differences on item performance are also noted. It is evident that most misplaced and unsuitable items occur in the WISC-R Verbal subtests. It is suggested that formal suitability studies should be conducted prior to widespread use of individual intelligence tests.

Educational and psychological test use in New Zealand has generally been based on limited adaptations of overseas scales, or on their presumed suitability without adaptation. This practice has been particularly evident in the case of individually administered ability tests. With these tests, clinical judgements appear to be made in relation to content suitability, with users making a variety of changes to test content, and administration procedures, to suit local conditions. Little, if any, local empirical data have been available to support content modifications or administration changes. This situation is certainly true of the Wechsler Scales, despite the fact that in their various forms these tests are widely used in New Zealand. For example, from 1974 to 1984, 559 WISC-R kits and 49,225 Record Forms were sold in New Zealand.[1]

Shortly after the publication of the WISC-R an initial investigation of its suitability for use in New Zealand was made by Tuck, Hanson, and Zimmerman (1975) who tested a sample of 100 Form 1 pupils with the WISC-R. Standard administration and scoring procedures were followed, but some item content changes were made to the Information, Similarities, Arithmetic, and Comprehension subtests to suit the New Zealand context. From the sample statistics, intercorrelational, and factorial data, the authors concluded that there was evidence in support of the WISC-R's validity as an ability measure with New Zealand children. With the content changes the WISC-R was judged to meet content validity requirements. Psychometrically the WISC-R was also viewed as satisfactory with reference made to the subtest means and standard deviations being "close to the expected values of 10 and 3" (Tuck et al., p.56). However, no New Zealand sample reliability estimates for the WISC-R Full Scale, Verbal Scale, Performance Scale, or subtest scores were reported. In terms of item characteristics, Tuck et al. mentioned that ". . . some of the Information items were inappropriate for New Zealand Form 1 children" (p.56). Unfortunately difficulty indices were not reported and are not now available.[2]

To date no other research has specifically addressed the issue of WISC-R suitability with New Zealand children, but as a consequence of WISC-R use in research contexts some data bearing on the suitability issue have been reported. For example, Silva (1982) and Silva, McGee and Williams (n.d.) have noted a small but apparently consistently higher mean WISC-R Full Scale IQ with New Zealand (Dunedin) seven-year-olds in comparison with

*This study was funded by a Research Contract (#55-2-119) between the authors, on behalf of Massey University and the Department of Education, Wellington. The authors are grateful to Judith Loveridge and Wendy Edmondston for their assistance in the data collection. Address correspondence to Dr R. St. George, Department of Education, Massey University, Palmerston North, New Zealand.

1 Sales figures from January 1974 to September 1984, kindly supplied by NZCER, Cedric Croft to Ross St. George, 8 October 1984.

2 Bryan Tuck to Ross St. George, 1 March 1984. The information subtest item analysis data was mislaid during subsequent shifts by the authors.

U.S. age peers. Silva et al. suggest a 5-point negative adjustment when using U.S. WISC-R norms in New Zealand.

The present authors (Chapman & St. George, 1984) have also reported WISC-R summary statistics. Full Scale, Verbal Scale and Performance Scale mean scores were comparable to those reported in the Manual for similar aged U.S. children. However, subtest mean score differences indicate that WISC-R item characteristics need to be considered more closely. This accords with views expressed by Tuck et al. (1975), Silva (1982), and Ballard (1984), who all suggest that more should be known about the WISC-R's characteristics and technical properties in the New Zealand context.

Psychometrically, an important step when investigating test suitability is the use of item analysis methods. Item analysis approaches can be employed to investigate item suitability, issues of item order, and the effect of item substitutions or modifications. These procedures also lead into reliability estimation using internal consistency methods. A knowledge of reliability estimates in turn bears upon validity claims. The WISC-R Manual (Wechsler, 1974) does not report any item analysis statistics in the form of difficulty or discrimination indices. It is reported that items were largely transferred from the 1949 edition of the WISC. Item modifications, deletions, or additions were made to reduce ambiguity, to replace obsolete items or those deemed to be differentially unfair to some subgroups of children, and to enhance reliability. Presumably, item orders were established empirically. Sattler's (1982) review and analysis of the WISC-R in his major volume on the assessment of children's intelligence also does not mention in any substantive manner item development and analysis.

The dearth of basic WISC-R item information is further reflected in the research literature. A search located only one paper which specifically focused on a WISC-R item analysis. Vance, Gaynor, and Coleman (1977) reported an item analysis based on 142 U.S. children aged 6 to 15 years. The sample consisted of children referred in the course of diagnostic assessment and was markedly atypical (mean Full Scale IQ 78.2, range 44 to 95). This small study was not mentioned in Quattrocchi and Sherrets' (1980) subsequent review of WISC-R research, nor is there any other information on WISC-R item studies. These authors do however, note non-United States suitability studies including that of Tuck et al. in New Zealand and reiterate remarks made about the WISC-R as a "useful" measure of intelligence in these suitability studies.

Clearly, there is a paucity of New Zealand-based empirical information on the WISC-R even though it must be regarded as a high use test. The present study reports on the item characteristics of the WISC-R with a sample of 11-year-old New Zealand children. Gender item differences were also investigated.

Method

Sample

The data in the present study were obtained from the 1982 cohort of Form 1 children attending five Palmerston North and Fielding intermediate schools. These children were taking part in a larger three-year study of school-related affective development and achievement (Chapman, 1985).

Initially, a sample was obtained by randomly selecting, from within each of the five participating schools, 10% of the 1,220 pupils who constituted the Form 1 cohort. This sampling procedure resulted in 125 children being tested with the full form of the WISC-R. For this study, only those in the 11-year-old range had their WISC-R data considered for analysis. This was because of the importance of chronological age grouping with WISC-R normative data. Therefore, the sample comprised 95 11-year-old children (56 boys, 39 girls), with mean age at the time of testing of 11.54 years (SD=0.24). Although random sampling procedures were used, the ratio of boys to girls is disproportionate in terms of the actual boy:girl ratio of 50:50 in the 1982 Form 1 cohort.

WISC-R Administration

The WISC-R was administered by trained testers during June, July, and August 1982. Children were randomly assigned to the testers. Following Tuck et al. (1975), changes were made to some items in the Information, Similarities, Arithmetic, and Comprehension subtests. In the Information subtest, Q12 was changed from "America" to "New Zealand", Q19 from "two countries that border the United States" to "two Australian states", Q20 from "pounds" and "ton" to "kilograms" and "tonne", Q24 "American man" to "New Zealand man", and Q27 "New York to Los Angeles" was changed to "Auckland to Sydney". In the similarities subtest Q10 was changed from "pound-yard" to "kilogram-metre". In the Arithmetic subtest, Q7 was changed from "pennies" to "cents", and in Q12 "bottles" was substituted for "cartons". Finally, changes in the Comprehension

subtest included Q11 "meat packing plants" to "freezing works", Q12 "street beggar" to "beggar", and Q17 "senators and congressmen" to "members of parliament". Otherwise, the tests were administered and scored according to the directions in the manual (Wechsler, 1974), with scale scores being based on the United States normative data.

Results

General WISC-R results

Summary IQ and subtest data are presented in Table 1. Overall, IQ and subtest scores tend to be slightly higher than the mean values indicated in the manual (Wechsler, 1974). These slightly higher scores may be due to the restriction of range in the lower ability levels. Children in the "Borderline" and "Mentally Deficient" range of intelligence (theoretically about 9% of the population, Wechsler, 1974, p.26), would not normally be placed in regular classrooms. However, scores on four subtests (Information, Digit Span, Picture Arrangement, and Coding) were slightly below the normative mean of 10 scaled score points. Silva (1982) has reported a similar slightly higher age related mean IQ with a sample of 950 7-year-old Dunedin children.

Table 1 also reports the internal consistency reliability estimates and standard error of measurement (SE_m) estimates for the WISC-R Full Scale, Verbal Scale and Performance Scale IQs, and WISC-R subtests. Internal consistency reliability estimates were computed by Hoyt's (1941) analysis of variance procedure, which, as Mehrens and Lehmann (1978) observe, yields exactly the same result as the Kuder-Richardson 20 or coefficient alpha. Wechsler (1974) estimated subtest reliabilities using a corrected split-half procedure. Although not favouring the split-half method Nunnally (1967) notes that the corrected correlation between any two halves of a test is an estimate of coefficient alpha. The Verbal Scale, Performance Scale and Full Scale IQ reliability estimates for the 11-year-old New Zealand sample are lower than estimates reported for similar aged children in the United States WISC-R norming sample (Wechsler, 1974).

Mean score differences between boys and girls were evident on most subtests, and on the Verbal, Performance and Full scales. Only one of these differences however, was statistically significant: girls obtained higher scores than boys on the coding subtest (t = 2.32, df = 93, $p < .05$).

WISC-R Subtest Item Analyses and Reliability Data

Item analysis data for the 11-year-old sample of 95 students are reported in Table 2 (Verbal Subtests) and Table 3 (Performance Subtests). Item reponses for each pupil were coded dichotomously as either "1" for right or "2" for wrong, in accordance with instructions detailed in the WISC-R manual. Following Vance, Gaynor and Coleman (1977), partial or qualitative scores were not separately coded. Items were considered to have been answered cor-

Table 1: *WISC-R IQ and subtest data for 11-year-old New Zealand sample*

			Total (N = 95)			Boys (n = 56)		Girls (n = 39)	
	r_{xx}[1]	r_s[2]	SEm	M	SD	M	SD	M	SD
Full Scale IQ	.90	—	4.22	102.89	13.35	104.32	14.49	100.85	11.38
Verbal IQ	.90	—	3.93	102.41	12.44	104.13	13.66	99.95	10.09
Performance IQ	.79	—	7.04	103.18	15.36	104.14	16.10	101.79	14.32
Information	.78	.97	1.11	9.57	2.36	9.92	2.54	9.05	2.00
Similarities	.63	.92	1.74	10.61	2.86	10.95	2.94	10.13	2.70
Arithmetic	.64	.97	1.59	10.38	2.65	10.64	2.96	10.00	2.09
Vocabulary	.71	.87	1.11	10.49	2.07	10.68	2.20	10.23	1.87
Comprehension	.59	.91	1.70	10.99	2.65	11.04	2.64	10.92	2.70
Digit Span	.70	.99	1.47	9.72	2.69	9.66	2.84	9.79	2.49
Picture Completion	.70	.97	1.52	10.48	2.77	10.93	2.98	9.85	2.32
Picture Arrangement	.58	.99	1.63	9.52	2.52	9.73	2.22	9.21	2.90
Block Design	.75	1.00	1.88	11.46	3.76	11.86	4.02	10.90	3.32
Object Assembly	.08	.80	3.42	11.08	3.57	11.23	3.64	10.87	3.50
Coding	—	—	—	9.82	2.94	9.25	2.91	10.64	2.83

1 Internal consistency reliabilty estimate coefficients, Hoyt's ANOVA.

2 Rank order correlations of WISC-R items based on standard test administration item orders and item difficulty orders for the NZ 11-year old sample. Corrected for tied ranks.

rectly if any scoring credit could be allowed. As Vance et al. (1977) point out, this procedure may reduce individual differences in item responses, but does increase the ordinal consistency of items in the test because the effect of partial credits are removed.

Item statistics were estimated by computing the Index of Difficulty (proportion passing, P),

Table 2: *WISC-R Verbal subtest item analysis statistics (n=95). Item Difficulty (P) and Discrimination (r_{pb}).*

Information Subtest

Items	P	r_{pb}
1	1.00	.00
2	1.00	.00
3	1.00	.00
4	1.00	.00
5	.99	.21
6	.98	-.02
7	1.00	.00
8	.96	.23
9	.99	.18
10	.91	.45
11	.95	.27
12	.95	.21
13	.74	.43
14	.38	.37
15	.62	.57
16	.19	.55
17	.28	.51
18	.35	.43
19	.20	.58
20	.32	.63
21	.11	.55
22	.20	.48
23	.09	.55
24	.04	.20
25	.05	.44
26	.05	.36
27	.04	.46
28	.03	.48
29	.06	.50
30	0.00	.00

Comprehension Subtest

Items	P	r_{pb}
1	1.00	.00
2	1.00	.00
3	.99	.00
4	1.00	.00
5	.92	.20
6	.93	.10
7	.93	.25
8	.99	.20
9	.96	.20
10	.54	.56
11	.90	.42
12	.85	.37
13	.38	.50
14	.45	.67
15	.51	.51
16	.43	.58
17	.26	.49

Vocabulary Subtest

Items	P	r_{pb}
1	1.00	.00
2	1.00	.00
3	1.00	.00
4	1.00	.00
5	1.00	.00
6	1.00	.00
7	1.00	.00
8	1.00	.00
9	.98	.11
10	.95	.33
11	.88	.39
12	.97	.22
13	.86	.27
14	.88	.28
15	.85	.41
16	.62	.46
17	.95	.31
18	.68	.51
19	.58	.58
20	.42	.49
21	0.00	.00
22	.04	.28
23	.46	.46
24	.04	.28
25	.10	.45
26	.06	.59
27	.04	.37
28	.04	.47
29	.01	.14
30	.05	.56
31	.03	.49
32	0.00	.00

Arithmetic Subtest

Items	P	r_{pb}
1	1.00	.00
2	1.00	.00
3	1.00	.00
4	1.00	.00
5	1.00	.00
6	1.00	.00
7	.99	.10
8	.98	.22
9	.94	.37
10	.93	.20
11	.78	.61
12	.66	.53
13	.56	.58
14	.54	.54
15	.17	.53
16	.26	.57
17	.15	.65
18	.03	.34

Similarities Subtest

Items	P	r_{pb}
1	1.00	.00
2	.96	.18
3	.96	.31
4	.97	.15
5	.97	.21
6	.99	.20
7	.98	.18
8	.94	.24
9	.88	.18
10	.53	.60
11	.71	.60
12	.34	.57
13	.31	.70
14	.17	.42
15	.21	.60
16	.06	.19
17	.05	.37

Digit Span Subtest

Items	P	r_{pb}
	Digits Forward	
1	1.00	.00
2	1.00	.00
3	.86	.66
4	.62	.85
5	.30	.79
6	.05	.46
7	0.00	.00
	Digits Backward	
1	1.00	.00
2	.98	.31
3	.72	.75
4	.37	.83
5	.08	.64
6	.03	.49
7	0.00	.00

Table 3: *WISC-R Performance subtest item analysis statistics (n=95). Item Difficulty (P) and Discrimination (r_{ph}).*

Picture Completion Subtest

Items	P	r_{pb}
1	1.00	.00
2	1.00	.00
3	1.00	.00
4	1.00	.00
5	.99	.19
6	.99	-.03
7	.95	.11
8	.99	.27
9	.94	.38
10	.94	.32
11	.86	.43
12	.96	.49
13	.95	.36
14	.93	.37
15	.82	.50
16	.86	.37
17	.84	.53
18	.67	.32
19	.47	.45
20	.75	.59
21	.36	.42
22	.26	.49
23	.28	.51
24	.29	.45
25	.16	.28
26	.02	.22

Object Assembly Subtest

Items	P	r_{pb}
1	1.00	.00
2	.95	.65
3	.99	.23
4	.92	.77

Picture Arrangement Subtest

Items	P	r_{pb}
1	.99	.34
2	1.00	.00
3	.98	.26
4	.91	.44
5	.81	.26
6	.75	.62
7	.78	.42
8	.65	.55
9	.58	.58
10	.38	.46
11	.41	.49
12	.18	.42

Block Design Subtest

Items	P	r_{pb}
1	1.00	.00
2	1.00	.00
3	.99	.17
4	.90	.32
5	.97	.44
6	.88	.55
7	.71	.67
8	.70	.75
9	.56	.80
10	.34	.71
11	.28	.69

and the Index of Discrimination (point-biserial correlation, r_{pb}). Differences in the proportion of boys and girls passing each item were computed, and differences greater than .20 are noted in the text[3] as major gender differences.

Information

Item analysis data for the Information Subtest are presented in Table 2. The first 12 items appear to have been fairly easy for most pupils, although Q10 (dozen) might have been better placed after Q12. A rather dramatic drop in correct responses occurred for Q13 (stomach) and Q14 (sunset), whereas a relative increase in correct responses occurred for Q15 (leap year). Q15 might have been better placed before Q14. Similarly, Qs 18, 20, and 22 appear easier than Qs 16, 19, and 21. From Q23 however, there is a more uniform fall-off in correct responses. Clearly, most of the discriminative power in the Information subtest for this sample of 11-year-olds occurred between Qs 12 and 23. This is further supported by the Index of Discrimination data which show that for Qs 13 to 22,, the point-biserial correlations are in the mid-range of .37 to .63. No marked departure from the pattern appears to have been caused by the use of the New Zealand substitute items (Qs 12, 19, 20, 24, 27), except for Q19 (Australian states), which appears inordinately difficult.

In terms of differences as a function of gender, Qs 20 (tonne) and 22 (glass) showed a greater difficulty for girls than for boys. In both cases the difference in the difficulty index between boys and girls was .25. Q19 (Australian

[3] A full set of item data by gender can be obtained from the authors.

states) also approached a difference of .20 in the difficulty index, with girls again finding the question more difficult than boys. The internal consistency reliability estimate was .78. This estimate is lower than the coefficient of .88 reported in the manual (Wechsler, 1974, p.28). Note that in this report, reliability estimates from the present data are compared in each case with the 11.5 year old children in Wechsler's standardization sample.

Similarities

Data for the Similarities subtest are shown in Table 2. The first nine items appear to be relatively easy at this age level, with the fall-off in correct responses occuring from Q9 (telephone-radio). The substitute item, Q10 (kilogram-metre), shows a more dramatic fall-off, and in terms of difficulty, would have been better placed after Q11 (anger-joy). Similarly, Q14 (liberty-justice) is slightly more difficult than Q15 (first-last). No major differences between boys and girls were evident. In terms of discriminative power, Qs 10 to 15 offer the most value, with point-biserial coefficients in the range of .42 to .70. The internal consistency reliability estimate of .63 is considerably lower than the value of .81 reported by Wechsler (1974).

Arithmetic

Data for Arithmetic are shown in Table 2. The main decrease in items answered correctly occurs from Q11 (supermarket), with a more dramatic decrease between Q14 (pencils) and Q15 (four boys). Q16 (bubble gum) was generally easier than Q15 and Q17 (bicycle), and is somewhat misplaced for this sample. Index of Discrimination coefficients reveal that Qs 11 to 17 have the most power, with r_{pb} values in the range of .53 to .65. In terms of the internal consistency reliability estimate, the coefficient of .64 for the present sample is considerably lower than the .81 value reported by Wechsler (1974). The only major gender difference occurred on Q16 (bubble gum) where only 15% of the girls answered the item correctly compared to 34% of the boys.

Vocabulary

Item data for Vocabulary items are presented in Table 2. These data suggest that a number of items appear to be misplaced in terms of their difficulty levels. Q12 (diamond), Q17 (nuisance), and Q23 (mantis) were somewhat easier relative to their placements, whereas Q21 (stanza) and Q22 (seclude) were relatively difficult. Indeed, no one answered Q21 correctly, and it appeared considerably out of place for this sample in its position of some 11 items before the end. Most of the discriminative power of the Vocabulary subtest occurred between Qs 15 and 31, although Table 2 shows that six items (Qs 17, 21, 22, 24, 27, 29) in that cluster have discrimination indices of less than .40. The internal consistency reliability estimate was .71, compared to .86 for the standardization sample (Wechsler, 1974). Only one major gender difference was apparent on this subtest. On Q19 (hazardous), 22% fewer girls than boys answered the item correctly.

Comprehension

Data for Comprehension subtest items are presented in Table 2. More than 90% of the sample answered Qs 1 to 9 correctly, with a fall-off in correct responses occurring after Q9. Indeed, Q10 (stamps) appears to have been misplaced, as a dramatic drop in the proportion of correct responses occurred from Q9 (.96) to Q10 (.54). Q13 (elections) also seems difficult relative to its placement. Relatively easy items in terms of their placement appear to be Qs 8 (number plates), 9 (criminals), and 15 (promise). With the exception of Q12 (beggar), Qs 10 to 17 provide the most discriminative power, with coefficients in the range of .42 to .67.

The reliability estimate for the comprehension subtest was .59, substantially lower than the .81 value reported by Wechsler (1974). No marked gender differences were observed, although on Q13 (elections), 19% fewer girls than boys answered the item correctly.

Digit Span

Data for Digit Span are presented in Table 3. The gradients of item difficulty for both Digits Forward and Digits Backward indicate a fairly even increase in item difficulty. Qs 3 to 6 in both parts of the subtest provide the most powerful discrimination, with coefficients in the range of .46 to .83. No important gender differences were evident. The reliability estimates for Digits Forward and Digits Backward were .58 and .56 respectively, and .70 overall. The .70 reliability estimate is lower than the .75 estimate obtained by Wechsler (1974).

Picture Completion

The Picture Completion item analysis data are presented in Table 3. These data suggest a relatively gradual increase in item difficulty from Q5 to Q17, with a more rapid increase in difficulty from Q18. Some items however,

appear to be misplaced. Q11 (belt) and Q15 (girl running) were difficult relative to their placement, whereas Q20 (screw) was relatively easy for its placement. In terms of discriminative power, Qs 9 to 17, and Qs 20, 23 and 24 are the most effective items. The reliability estimate was .70, which was lower than the .80 value reported by Wechsler (1974). In terms of gender differences, Q18 (scissors) was answered correctly by 21% more boys than girls. No other items showed major gender differences.

Picture Arrangement

Data for Picture Arrangement items are reported in Table 3. A fairly consistent increase in item difficulty is evident for items in this subtest, with no major misplacements occurring. With the exception of Q5, Qs 4 to 12 all have discrimination indices greater than .40. The internal consistency reliability estimate of .58 for this sample is lower than the .71 value reported by Wechsler (1974). No major gender differences were evident on any Picture Arrangement items.

Block Design

Data for Block Design items are presented in Table 3. The gradient of difficulty levels suggest that no major item misplacements are apparent for this subtest, although Q4 may have been better placed after Q5. The point-biserial correlations show that Qs 5 to 11 have good discriminative power, with coefficients in the range of .44 to .80. The reliability estimate for Block Design with this sample is .75, compared to .89 for Wechsler's (1974) sample. In terms of gender differences, there was a tendency for girls to experience greater difficulty from Q7 onwards. The last two items revealed marked differences between boys and girls, with 25% fewer girls completing Q10 correctly, and 21% fewer completing Q11.

Object Assembly

Table 3 also presents the Object Assembly item data. In that all four puzzles were correctly completed by at least 90% of the sample, this subtest appears to offer relatively limited discriminative power. Most of the variance on this subtest however, is due to the bonus points awarded for speed of completion. This element is not reflected in the item analysis data. In this instance the small number of items and restriction of range in correct vs incorrect responses has resulted in the extremely low internal consistency reliability coefficient of .08.

Item ordinality

Because some items in the subtests appeared to be misplaced, it was decided to investigate item order correspondence empirically.[4] Following Jensen (1980), the approach adopted was to derive the Spearman rank order correlation coefficient between the WISC-R item administration order within subtests and the order based on the ranking of item difficulties. The rank order correlation coefficients are reported in Table 1.

Inspection of these coefficients indicates a generally high degree of item order correspondence between the standard administration sequence and the ranked difficulty order based on the New Zealand sample data. In relative terms however, item order discrepancy could be an issue with the Similarities, Vocabulary and Comprehension subtests of the Verbal Scale. The value for the Object Assembly subtest is a function of there being only four items and the order of Items 2 and 3 were reversed by a difficulty difference of $p = .04$ with the New Zealand sample.

Discussion

This study compared the performance of a sample of 11-year-old New Zealand children on the WISC-R, with the American 11.5-year-old standardization sample. The focus was on the operating characteristics of individual items.

The means for the Verbal, Performance and Full Scale IQs suggest that the New Zealand sample is performing at a slightly higher level than the United States standardization sample. This result probably arose from restriction of range at the lower IQ end in the present sample. Wechsler (1974, p.26) reported that 8.2% of the standardization sample had Full Scale IQs below 79; here only one person in the sample (about 1%) obtained an IQ below that level. The relatively lower standard deviation for the Verbal and Full Scale are also indicative of a restricted range.

In that the means for the subscales and the Verbal, Performance and Full Scale scores are within one SE_m unit of the standardization means, the slight fluctuations noted for this New Zealand sample probably hold no practical significance. Although there was a tend-

[4] The method adopted here is but a simple first step to the range of analysis techniques outlined (see Jensen, 1980, pp.432-446).

ency for boys to obtain higher scores than girls on the Verbal, Performance and Full Scales, and on most subscales, the only statistically significant gender difference was on the Coding subtest, where girls obtained higher scores overall than boys.

Composite reliability estimates for the Verbal, Performance, and Full Scale IQs were all lower than those reported by Wechsler (1974), as were the individual subtest reliability estimates, calculated using Hoyt's analysis of variance procedure. Lower reliabilities increase the magnitude of SE_m estimates, however, larger SE_m values were not found for subtests with the present sample because of the restriction of range, as shown in the comparatively lower standard deviations. Were a full range sample to be tested, on the basis of results from this study, lower reliability estimates and larger SE_m values would be expected. In essence, there is some indication that the WISC-R may not be as reliable for New Zealand children as it is for children in the United States. Users making important placement decisions on the basis of WISC-R scores should bear this in mind. A cautious approach would involve employing a confidence interval based on a SE_m value at the .01 level (i.e., $SE_m \times 2.58$). This would result in an 11 IQ point confidence interval around an estimated true Full Scale IQ score.

In terms of subtest item statistics, it is clear that some items are misplaced. Assuming that a fairly even increase in the difficulty of items is desirable, then it may be advantageous to reorder some items. The item order problem is most evident in the Verbal Scale subtests of Information, Similarities, Vocabulary, and Comprehension. These subtests are probably most influenced by educational differences between New Zealand and the United States. With the possible exception of a few items in those four subtests which might represent a North American knowledge bias (e.g., Information Q16 and Q21), there is no clear evidence that item misplacements are due to sociocultural differences between New Zealand and the United States. The apparently misplaced items in these four subtests may just be due to different teaching emphases and learning opportunities in and out of schools. For example, the sudden drop in correct responses for Information Q14 (sunset) and Q16 (light bulb) cannot easily be attributed to sociocultural differences. Similarly, the content of Vocabulary Qs 16 (contagious), 21 (stanza) and 22 (seclude) and Comprehension Q10 (stamps) are not obviously slanted in favour of U.S. children.

As far as the substitute New Zealand items recommended by Tuck et al. (1975) are concerned, only one may have given rise to misplacement, Q19 of the Information subtest. Wechsler's original item referred to two countries that border the United States. The substitution required that any two states in Australia be named. Considering the number of correct responses for that item (20%) relative to the preceding and succeeding items (38% each), the substitution may not be an appropriate one for New Zealand children. Two additional substitutions however, could also be considered for the Information subtest. These are Qs 16 and 21. The inventor of the electric light bulb (Q16) and the continent in which Chile is located (Q21) are possibly more frequently taught to and relevant for United States children than New Zealand children. For the Performance subtests, no major item misplacements occurred and the rank-order correlations are very high for all subtests except Object Assembly.

In terms of gender some marked differences in item responses were noted. About 20% or more boys answered eight items more correctly than girls over five of the subtests. While it would be inappropriate to suggest that the WISC-R is in general systematically biased against girls, some specific items do appear to discriminate inappropriately on the basis of gender. However, sampling restrictions must be borne in mind here.

At a more general level, the discrepancies in the item order positions, particularly in their Verbal Scale subtests, points to the need for adaptation studies to be done speedily following the introduction of such a test. Furthermore, the reliability data that can be generated simultaneously are equally important. Availability of such data seem imperative to us if tests results are to form a part of educational, occupational, or legal decision-making. It hardly needs reiterating that such confirmatory or checking analyses are clearly referred to in the *Standards for educational and psychological tests* (American Psychological Association, 1974). An important issue, relating to this concern, is the effect that markedly misplaced items could have on a child's score on the

WISC-R. The application of standard subtest discontinuation rules could effectively prevent the gaining of credit on subsequent passable items and failed misplaced difficult items could well have negative motivational consequences for the child and inferential consequences for the test administrator if making "clinical" interpretations of the test protocol.

Given the rate of WISC-R test form consumption, on average 4,000 plus forms per annum, we are surprised that prior analyses and checks of the nature illustrated in this report have not been made. Even a 50% data collection rate would return over 2000 protocols per annum and atypicality in sampling could be accommodated for by restricted random sampling from relevant age/class populations. Procedures for test modification and local data provision that go beyond "eye-ball" adaptation are clearly needed. Early work of this nature could well prevent unsuitable, unreliable, and hence invalid psychological and educational tests doing more harm than good in New Zealand.

References

American Psychological Association (1974). *Standards for educational and psychological tests.* Washington: Author.

Ballard, K. D. (1984). Interpreting Stanford-Binet and WISC-R IQs in New Zealand: The need for more than caution. *New Zealand Journal of Psychology, 13,* 25–31.

Chapman, J. W. (1985). *Self-perceptions of ability, learned helplessness and achievement expectations of children with learning disabilities.* Report to the Director, Research and Statistics, Department of Education, Wellington.

Chapman, J. W., & St. George, R. (1984). The WISC-R: Results with a New Zealand sample and relationships with TOSCA and PATs. *New Zealand Journal of Educational Studies, 19,* 184–188.

Hoyt, C. J. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6,* 153–160.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Mehrens, W. A., & Lehmann, I. J. (1978). *Measurement and evaluation in education and psychology,* 2nd Edition. New York: Holt, Rinehart & Winston.

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Quattrocchi, M., & Sherrets, S. (1980). WISC-R: The first five years. *Psychology in the Schools, 17,* 297–312.

Sattler, J. M. (1982). *Assessment of children's intelligence and special abilities,* (2nd edition). Boston: Allyn & Bacon.

Silva, P. A. (1982). Interpreting Stanford-Binet and WISC-R IQs in New Zealand. *New Zealand Journal of Educational Studies, 17,* 195.

Silva, P. A., McGee, R., & Williams, S. (no date). *Results from use of the Stanford-Binet Intelligence Scale at Ages Four, Five and Six and the Wechsler Intelligence Scale for children at Age Seven.* Report from the Dunedin Multidisciplinary Health and Development Research Unit. Dunedin: University of Otago Medical School.

Tuck, B. F., Hanson, A. L., & Zimmerman, M. (1975). The WISC-R: A New Zealand Study of norms and validity. *New Zealand Journal of Educational Studies, 10,* 52–58.

Vance, H. B., Gaynor, P., & Coleman, M. (1977). Item analysis of the Wechsler Intelligence Scale for Children — Revised. *Psychology in the Schools, 14,* 132–139.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children — Revised WISC-R, Manual.* New York: Psychological Corporation.