Empirical Versus Random Item Selection in the Design of Intelligence Test Short Forms—The WISC-R Example

David S. Goh
Central Michigan University

This study demonstrated that the design of current intelligence test short forms could be improved by employing a more effective method of item selection based on psychometric theory. Two short forms of the recently published WISC-R were developed, one employing a design determined by empirical item analysis results of the standard test battery and the other employing the well-known Yudin scheme determined by systematic random selection of test items. In all analyses the item analysis method of item selection was shown to yield more accurate results than the Yudin procedure. Practical usefulness as well as limitations of the present WISC-R Short form are discussed.

Satz and Mogel (1962) published the first "split-half" abbreviation procedure of an intelligence test in which all subtests of the original test battery were retained, but the number of items within each subtest was greatly reduced. The major advantage of this approach over the "subtest-combination" approach is that it allows the abbreviated instrument to sample the complete variety of psychological facets tapped by the original test while still effecting a considerable saving in administration time. Following this proposed procedure by Satz and Mogel, Yudin (1966) constructed a short form of the WISC employing selected items on 9 of the 11 subtests, while leaving the Digit Span and Coding subtests unchanged. He reported that correlations ranged from .76 to .94 between the subtests of the short form and the standard WISC. The correlations between the two instruments for the Verbal, Performance, and Full-Scale IQs were .96, .93, and .97, respectively. As Yudin's sample comprised only emotionally disturbed children, researchers have reported similar correlations on other populations (Erikson, 1967; Gayton, Wilson, & Bernstein, 1970; Reid, Moore, & Alexander, 1968). However, in spite of these high correlations, significant differences between obtained scores on the short form and the standard test battery have also been found. These significant differences between an individual's scores on the two instruments have greatly reduced the practical value of the short forms in clinical usage (Erikson, 1967; Lombardi & Cohen, 1976). Therefore, the current design of intelligence test short forms remains to be improved.

A major problem of the well-known Yudin procedure, as well as other similar procedures (e.g., Silverstein, 1968), seems to lie in their rather unwarranted method of item selection. These authors had selected, either arbitrarily or randomly, items within each subtest to be included in the short forms. In their procedures

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 3, No. 1 Winter 1979 pp. 75-82

none had taken into consideration the psychometric properties of individual test items of the original test battery. From a psychometric viewpoint, it obviously would be more desirable to select test items for the short form on an empirical rather than on a random basis.

Accordingly, a highly recommendable approach would be to design the short form of a test based on item analysis results of the original instrument. Item difficulty and reliability or discrimination values should be considered in the selection of items to be used in the short form. In other words, while all subtests of the standard test battery are retained, only test items with high reliability or discrimination values within each subtest should be selected for the short form. In the meantime a reasonable continuity of item difficulty levels within each subtest should be maintained. Presumably, by eliminating unreliable or non-discriminatory items, the original test can be shortened to produce a short form which would yield more valid and comparable results.

The present study was conducted to test this proposed empirical method of item selection in the design of intelligence test short forms. The instrument employed was the recently published Wechsler Intelligence Scale for Children—Revised (Wechsler, 1974). The WISC-R is a revised and updated version of the WISC. A total of 72% of the WISC items were retained in the WISC-R, with a small portion of the items slightly modified. New items were added to 7 of the 12 subtests. In view of some of the changes in item content and organization over the two test batteries, existing WISC short forms would not apply to the WISC-R, and new short forms are clearly needed for the latter. The present study attempted to construct two WISC-R short forms, one employing the item analysis approach proposed above and the other employing the Yudin method. The two methods then were compared to determine whether the former yielded more valid results than the latter.

Method

Subjects

Subjects for the present study were randomly sampled from a several-county area in central Michigan. During the past three years, 142 children (65 males and 77 females) were administered the WISC-R individually under the standard testing procedure in a midwestern state university psychology center. The mean age of the sample was 10 years and 11 months (range 6–0 to 16–11). The mean Full-Scale IQ of the sample on the standard WISC-R was 110.05, with a standard deviation of 15.64.

Procedures

The WISC-R protocols were first scored in the standard form, and then in the two short forms constructed in the present study. The first short form was designed based on results of an item analysis study of the WISC-R (Goh, 1977) which reported indices of difficulty and discrimination for all items. In selecting items for this short form, the following procedures were used: (1) in each subtest, the number and/or percentage of test items used were determined according to the Yudin scheme to insure comparability between the lengths of the two short forms; (2) in each subtest except Coding, successive dyads or triads of items were examined, and items with the highest index of discrimination and item difficulty closest to .50 were selected for the short form; (3) the Coding subtest was not shortened; and (4) the two supplementary subtests—Digit Span and Mazes—were not included in the construction of the short form, since they were not used regularly in the computation of WISC-R IQs.

The test items finally selected for the WISC-R short form were as follows: Information—Items 2, 6, 8, 11, 14, 17, 19, 22, 26, 28; Similarities—Items 2, 3, 8, 10, 11, 13, 15, 17; Arithmetic—Items 2, 4, 5, 8, 10, 12, 13, 15, 17; Vocabulary—Items 1, 5, 9, 12, 15, 18, 19, 22, 25,

28, 31; Comprehension—Items 2, 3, 6, 8, 10, 12, 14, 16, 17; Picture Completion—Items 2, 5, 8, 11, 15, 18, 20, 23, 25; Picture Arrangement—Items 1, 4, 6, 8, 10, 11; Block Design—Items 2, 5, 8, 10; Object Assembly—Items 2, 4; and Coding—all items. In the scoring of this short form, each subtest was multiplied by the appropriate constant—2 or 3—depending on whether 1/2 or 1/3 of the items were used. Raw scores on the subtests were then transformed to scaled scores, and IQs were obtained from the manual in the usual manner.

The WISC-R protocols were also rescored on the second short form which was developed according to the Yudin (1966) method. In this procedure, every third item was scored for Information, Vocabulary, and Picture Completion; every even item for Arithmetic; every odd item for Similarities, Comprehension, Picture Arrangement, Block Design, and Object Assembly; and all of the items for Coding. IQs were obtained in the same manner as mentioned above, except that the Yudin method applied different correction factors to 6 of the 10 subtests in the computation of their raw scores.

Results

Table 1 presents the means and standard deviations of the Verbal IQ(VIQ), Performance IQ(PIQ), and Full-Scale IQ(FSIQ) for the standard WISC-R and the two short forms. The data in Table 1 were subjected to two-tailed t-tests ($\alpha = .05$) to determine whether there were significant differences between the IQs obtained from the standard and two short forms. The results indicated no significant IQ differences between the standard WISC-R and the empirical short form, while the Yudin short form significantly underestimated the standard WISC-R VIQ ($d = 4.21$, $t = 10.58$, $p < .0001$), PIQ ($d = 7.16$, $t = 14.25$, $p < .0001$), and FSIQ ($d = 6.29$, $t = 17.43$, $p < .0001$).

Table 2 shows the Pearson product-moment correlation coefficients between the standard WISC-R and the short forms for subtest scaled scores and IQs. As can be seen from Table 2, both short forms produced significantly high correlations with the standard WISC-R. These correlations compared favorably with those reported on the WISC short forms (Gayton, Wilson, & Bernstein, 1970; Yudin, 1966). Tests of difference between correlations of the two short forms and the standard WISC-R revealed no significant differences between the two short forms for the Verbal, Performance, or Full-Scale IQs. As for the subtest scaled scores, no significant differences between the two short forms were found for the Verbal subtests; however, the empirical short form had significantly higher

Table 1
Means and Standard Deviations of IQs
for the Standard WISC-R and Two Short Forms (N = 142)

IQ		Standard	Goh	Yudin
VIQ	Mean	107.86	106.87	103.99
	S.D.	15.78	19.97	15.44
PIQ	Mean	110.37	110.82	102.96
	S.D.	15.10	17.23	14.94
FSIQ	Mean	110.05	109.66	103.82
	S.D.	15.64	17.38	15.64

Table 2
Correlations[a] Between the Standard and Two Short Forms for Scaled Scores and IQs (N = 142)

Measure	Goh	Yudin	Measure	Goh	Yudin
Information	.87	.89	Picture Completion	.79	.71
Similarities	.85	.81	Picture Arrangement	.86	.77
Arithmetic	.85	.87	Block Design	.92	.86
Vocabulary	.85	.83	Object Assembly	.88	.73
Comprehension	.84	.85	Coding	1.00	1.00
VIQ	.93	.95	PIQ	.94	.91
FSIQ	.94	.95			

[a] all $p < .001$

correlations with the standard WISC-R than the Yudin abbreviation in three of the five Performance subtests—Picture Arrangement ($p < .05$), Block Design ($p < .05$), and Object Assembly ($p < .01$).

Table 3 shows the results of comparison between the standard WISC-R and the two short form FSIQs for both age and intelligence groups. Four age groups were established in this analysis, with resultant *N*s of 28 (10 males and 18 females), 53 (30 males and 23 females), 38 (16 males and 22 females), and 23 (9 males and 14 females) for the 6.0–7.11, 8.0–10.11, 11.0–13.11, and 14.0–16.11 age levels, respectively. The intelligence groups were established based on the cutoff IQ scores for the various intelligence classifications reported in the WISC-R manual (Wechsler, 1974). However, the 11 subjects who scored at or below the "Low Average" level were grouped together into one category, 60–89 IQs. As can be seen from Table 3, the Yudin short form again consistently underestimated the standard WISC-R IQ for all age and intelligence groups by a range of 4.35 to 10.17 points. There were no significant IQ differences between the standard WISC-R and the empirical short form for the intelligence groups, although significant differences were found for three of the age groups.

One of the most frequent uses of IQ scores in a school or clinical psychological setting is to classify intellectual functioning level of the subject for diagnostic purposes. Table 4 presents the results of intellectual classification of the present subjects based on their FSIQs on the standard WISC-R and the two short forms. As can be seen from Table 4, the empirical short form correctly classified 95 (67%) of the 142 subjects, while the Yudin abbreviation correctly classified 78 (55%) of the subjects against the standard WISC-R classification. An examination of Table 4 indicated that when the empirical short form did misclassify a subject, it classified the subject into either the next higher or lower category. However, when the Yudin short form misclassified a case, it consistently classified the subject into the lower categories. The two short forms only agreed entirely in the IQ 80–89 and 60–79 categories.

The data were finally analyzed by examining the magnitude of score deviation between the subject's obtained IQs on the standard and the

Table 3
Comparisons Between Standard and Two Short Form FSIQs by Age and Intellectual Level

Group	N	Standard	Short Forms		$\bar{d}$	t	p
Age							
6.0-7.11	28	111.25	Goh	106.61	5.64	5.33	.001
			Yudin	106.50	5.75	6.45	.0001
8.0-10.11	53	114.26	Goh	113.77	.49	.71	N.S.
			Yudin	107.02	7.24	9.53	.0001
11.0-13.11	38	107.87	Goh	109.53	-1.66	-2.16	.05
			Yudin	101.55	6.32	13.40	.0001
14.0-16.11	23	101.26	Goh	104.13	-2.87	-2.62	.05
			Yudin	96.91	4.35	4.32	.0001
Intellectual Level							
60-89	11	85.00	Goh	83.30	1.70	1.58	N.S.
			Yudin	80.00	5.00	5.67	.0001
90-109	61	100.46	Goh	99.18	1.28	1.93	N.S.
			Yudin	95.22	5.24	10.59	.0001
110-119	30	114.30	Goh	115.87	-1.57	-1.30	N.S.
			Yudin	107.47	6.83	12.30	.001
120-129	22	123.82	Goh	125.45	.37	.25	N.S.
			Yudin	118.09	5.73	5.15	.0001
130-149	18	135.72	Goh	135.83	-.11	.08	N.S.
			Yudin	125.55	10.17	5.73	.0001

Table 4
Frequency Distributions of Intellectual Classifications
Based on the Standard and Two Short Form FSIQs

Standard Classification	Method	Short Form Classification					
		130-145	120-129	110-119	90-109	80-89	60-79
130-145	Goh	16	2				
(N=18)	Yudin	5	12	1			
120-129	Goh	5	10	7			
(N=22)	Yudin		9	10	3		
110-119	Goh		11	14	5		
(N=30)	Yudin			7	23		
90 -109	Goh			8	47	6	
(N=61)	Yudin				49	12	
80 - 89	Goh					6	3
(N= 9)	Yudin					6	3
60 - 79	Goh						2
(N= 2)	Yudin						2

two short forms. Table 5 displays the frequencies and percentages of the short form protocols deviating from the standard WISC-R IQs. As can be seen from Table 5, there was a clear difference between the distributions of subjects on the two short forms in terms of score deviations from the standard WISC-R. Approximately 43% of the subjects showed a difference within 3 IQ points (less than 1 standard error of measurement of the WISC-R FSIQ) between the empirical short form and the standard WISC-R, while only 22% had the same difference for the Yudin short form. Furthermore, the two short forms included 81% and 60% of the subjects, respectively, within 7 IQ points (about 2 standard errors of measurement) of the standard WISC-R IQs. On the other hand, the Yudin short form classified twice as many subjects as the empirical short form within 10 or more IQ points from the standard test battery.

Discussion

The main purpose of the study was to point out that the current design of intelligence test short forms could be improved by employing a more effective item selection method based on psychometric theory. The WISC-R example provided in the study has demonstrated that the proposed empirical item selection method is indeed more desirable and meaningful than the systematic random selection method typically used by previous studies. In all analyses, the present empirical method yielded more satisfactory results than the well-known Yudin (1966) procedure. Both methods showed very high correlations with scores on the original test battery. However, the empirical short form produced more accurate estimation of the standard WISC-R IQs for the subjects as a group. When the subjects were divided into subgroups at different intellectual levels, the empirical short

Table 5
Frequencies and Percentages of FSIQ Deviations of the Two Short Forms from the Standard WISC-R

Deviation in IQ Points[a]	Goh		Yudin	
	f	%	f	%
0	11	7.7	6	4.2
1	15	10.5	5	3.5
2	18	12.6	13	9.1
3	17	11.9	7	4.9
4	7	4.9	14	9.8
5	20	14.0	15	10.5
6	14	9.8	13	9.1
7	12	8.4	12	8.4
8	8	5.6	21	14.7
9	8	5.6	12	8.4
10	12	8.4	24	16.8

[a] In absolute value.

form also yielded similar mean scores in comparison to the standard WISC-R IQs. Nevertheless, when the subjects were collapsed into age groups, with resultant smaller sample size and larger dispersion of scores, the empirical short form IQs differed from the standard test battery in three of the four groups. This result indicated the importance of the age factor in estimating standard WISC-R IQs from the empirical short form measure. Future research should use larger samples for these three age groups (6.0 to 7.11, 11.0 to 13.11, and 14.0 to 16.11) to further determine the comparability between the standard and short form IQs.

Despite the high correlations and comparability of mean IQ scores between the empirical short form and the standard WISC-R, the data suggested somewhat limited value of both short forms in psychological settings where *individual* test scores are to be used for diagnostic or classification purposes. The present empirical and Yudin short forms correctly classified only 67% and 55%, respectively, of the total sample into the proper standard WISC-R intelligence groups. These classification rates fall within the range of findings of previous short form studies. Psychologists are strongly cautioned not to use short form IQs for diagnostic classification or labeling purposes in school or clinical settings. Instead, for the sake of time saving, it is best to use intelligence test short forms as screening devices in identifying subjects with potential difficulties for further and more complete psychological evaluation. The present empirical short form appears to be particularly useful in this regard, as it scored 81% of the subjects within 7 IQ points of the standard WISC-R results. In addition, the empirical short form is recommended as a desirable group measure for research and other similar purposes, such as program development and evaluation.

References

Erikson, R. V. Abbreviated form of the WISC: A re-evaluation. *Journal of Consulting Psychology*, 1967, *3*, 641.

Gayton, W. F., Wilson, W. T., & Bernstein, S. An evaluation of an abbreviated form of the WISC. *Journal of Clinical Psychology*, 1970, *26*, 466–468.

Goh, S. S. An item analysis of the WISC-R. Unpublished manuscript, Department of Psychology, Central Michigan University, Mt. Pleasant, MI, 1977.

Lombardi, D. A., & Cohen, S. H. Differential reliability and validity of two selected WISC short forms. *Catalog of Selected Documents in Psychology*, 1976, *6*, 1–15.

Reid, W. B., Moore, D., & Alexander, D. Abbreviated form of the WISC for use with brain-damaged and mentally retarded children. *Journal of Consulting and Clinical Psychology*, 1968, *32*, 236.

Satz, P., & Mogel, S. An abbreviation of the WAIS for clinical use. *Journal of Clinical Psychology*, 1962, *18*, 77–79.

Silverstein, A. B. Evaluation of a split-half short form of the WAIS. *American Journal of Mental Deficiency*, 1968, *72*, 839–840.

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

Yudin, L. W. An abbreviated form of the WISC for use with emotionally disturbed children. *Journal of Consulting Psychology*, 1966, *30*, 272–275.

Acknowledgments

This study was supported by Faculty Research and Creative Endeavors Grant 4-22136 from Central Michigan University.

Author's Address

David S. Goh, Department of Psychology, Sloan Hall, Central Michigan University, Mt. Pleasant, MI 48859.