DATA EVALUATION AND METHODS RESEARCH

Series 2
Number 15

evaluation of

Psychological Measures Used in the Health Examination Survey

of children, ages 6-11

A critical review of literature pertaining to the psychological measures used in Cycle II, with recommendations concerning validity, reliability, and applicability to the Survey data.

DHEW Publication No. (HRA) 75-1295

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Public Health Service

Health Resources Administration
National Center for Health Statistics

Rockville, Maryland

Vital and Health Statistics-Series 2-No.15
First issued in the Public Health Service Publication No. 1000
March 1966

NATIONAL CENTER FOR HEALTH STATISTICS

Library of Congress Catalog Card Number 65-62272

FOREWORD

The practice of comparing one individual with another is as old as recorded history. Man's earliest writings are replete with statements indicating that he has long viewed his fellow man in terms of whether or not he measured up to an expected ideal. Similarly, the performance of a man has traditionally been described in terms of how it compares with that of another man. However, subjecting these "known" differences to the scientific method of inquiry is a recent development.

In the area of individual differences in behavior and psychological characteristics, research has progressed from the simple to the complex. The first studies dealt with the simple functions of speed of reaction time. Today, studies are aimed at measuring individual differences in the complex functions of motivation, ego-integration, and cognition.

Progress in developing a *technology* for measuring behavior has progressed in a similar manner. Instruments are available which, most scientists will agree, accurately measure the speed with which an individual taps his finger in response to a given signal. Scientists do not agree, however, on the adequacy of the equipment used to measure individual differences in intelligence. Moreover, there will even be some disagreement over the use of the word "intelligence" to describe certain aspects of behavior.

Because of the present state of the art of psychological measurement, studies such as those conducted by the Health Examination Survey encounter difficult problems in attempting to estimate the prevalence of various mental health factors in the population.

The Health Examination Survey is part of the U.S. National Health Survey, authorized by Congress in 1956 to collect information about the Nation's health. Data are collected by direct examinations of individual persons chosen to constitute a probability sample of some segment of the total population of the United States.

The first sample represented the adult population aged 18 through 79 years. Since the study was primarily concerned with the prevalence of chronic physical disease, the examination did not include psychological measurements. The second sample consisted of noninstitutionalized children ages 6 through 11, among whom the incidence of chronic disease is insignificant. The important health factors in this group are found in those functions which result in growth and development. These, then, were the factors to be studied.

Many authorities in the field of growth and development contributed to the planning phase of the Survey. Although they generally agreed on what factors should be measured, they could not agree on how the measurements should be obtained. They did conclude that present instruments were inadequate but that these were the only tools available.

The tests which are discussed in the following report were those selected for use by the Health Examination Survey. In choosing these instru-

ments, primary consideration was given to those which best met the following criteria:

1. They were capable of yielding data in those areas considered most important to the study of growth and development.
2. They would produce data in a form which would be meaningful to the individuals responsible for children's health.
3. They were suitable for use in a survey operation where examiners change frequently, where only 1 hour is available to conduct the examination, and where examining conditions are less than optimal.

The selected instruments are not ideal, but they are felt to be the best compromise offered by the present state of the art of measurement.

How *much* was compromised? What can be said about the growth and development of children from the data obtained by the use of these instruments?

Through a contractual arrangement with Dr. Sells, the first step has been taken in answering these questions.

Lois R. Chatham, Ph.D.
Psychological Advisor
Division of Health Examination Statistics

CONTENTS

CONTENTS—Con.

CONTENTS—Con.

IN THIS REPORT the psychological procedures used in the Health Examination Survey conducted between June 1963 and December 1965 for children ages 6 through 11 are critically evaluated.

In his analysis, the author combines his own professional competence with the information obtained in an extensive survey of literature pertaining to the four procedures used—the Wechsler Intelligence Scale for Children, the Wide Range Achievement Test, a modification of the Draw-A-Man Test, and the Thematic Apperception Test. The result is an evaluation of the instruments which is made in terms of their validity, reliability, and applicability for use in the Health Examination Survey.

Finally, the author points out the strengths and weaknesses of each procedure and makes recommendations concerning the eventual use of data obtained in the Survey.

SYMBOLS

Data not available	---
Category not applicable	...
Quantity zero	-
Quantity more than 0 but less than 0.05	0.0
Figure does not meet standards of reliability or precision	*

EVALUATION OF PSYCHOLOGICAL MEASURES USED IN THE HEALTH EXAMINATION SURVEY OF CHILDREN AGES 6-11

S. B. Sells, Ph.D., *Institute of Behavioral Research, Texas Christian University*

INTRODUCTION

This report is the outcome of a contract with the National Center for Health Statistics. The purpose of the contract was to obtain an objective critical evaluation of the psychological procedures chosen for use in the Health Examination Survey of children ages 6 through 11. The objectives may be summarized as follows:

1. To prepare a critical review concerning the development and use of the psychological procedures used in Cycle II based on available literature and unpublished reports (theses, dissertations, and others). These measures include the Vocabulary and Block Design subtests of the Wechsler Intelligence Scale for Children, the Oral Reading and Arithmetic subtests of the Wide Range Achievement Test (1963 edition), the Draw-A-Man Test, and cards 1, 2, 5, 8BM, and 16 of the Thematic Apperception Test.
2. To make recommendations concerning the appropriate inferences which can be made concerning individual growth and development based on scores derived from the test battery described above.
3. To recommend what research must be done if the objectives of the Health Examination Survey are to be accomplished.
4. To make original recommendations concerning the types of cross-disciplinary analyses that can be performed on data obtained in the Health Examination Survey of children.

An extensive survey of the literature was made, but only the most relevant material was included in this final report. Literature was considered relevant if it was either empirical research or a review which included or made reference to the tests used in the Survey. Empirical studies which were conducted on samples of U.S. children ages 6 to 12 years were given preference. A few important reports which did not meet these criteria were included because of their methodological features or their significant content. Unpublished master's theses and dissertations were obtained, as extensively as possible, by interlibrary loan. Information was sought and, with some success, obtained from the publishers and selected users of the reviewed tests.

One empirical study was carried out under this contract. Its results are included in the section on the Goodenough Draw-A-Man Test. The study was stimulated by a recent publication by Dale B. Harris entitled *Children's Drawings as Measures of Intellectual Maturity*. This text is basically a revision of the 1926 book by Florence L. Goodenough entitled *Measurement of Intelligence by Drawings*. In his publication, Harris includes new point-score scales and modernized norms for scoring drawings of the human figure.

The text of this report is divided into six sections. Sections I-IV present critical discussions of various tests used by the Health Examination Survey. The tests are discussed in the following order:

I. The Wechsler Intelligence Scale for Children, Vocabulary and Block Design subtests
II. The Wide Range Achievement Test, the Oral Reading and Arithmetic subtests
III. The Goodenough Draw-A-Man Test
IV. The Thematic Apperception Test

Section V briefly discusses some of the issues which arise when these tests are used as a battery. Finally, section VI considers the cross-disciplinary relationships between "psychological" and "nonpsychological" measures.

Each research study or review referred to in this report is identified by a number placed in parentheses immediately following the cited reference. Bibliographies following each of the first four sections of the report contain all references cited in the respective sections.

Research studies which were abstracted as part of the literature-review portion of this contract are also included in the four bibliographies. The actual abstracts of the reviewed literature appear as appendixes to the report. For convenience, numbers which identify the abstracts correspond to the number given when the reference is cited in the text of the report.

These abstracts have been deposited as document number 8486 with the Library of Congress. A copy may be secured by sending the document number and $28.80 for photoprints or $3.20 for 35mm. microfilm to the American Documentation Institute Auxiliary Publication Project, Photoduplication Service, Library of Congress, Washington, D.C., 20541. Advance payment is required. Checks or money orders should be made payable to Chief, Photoduplication Service, Library of Congress.

I. THE WECHSLER INTELLIGENCE SCALE FOR CHILDREN, THE VOCABULARY AND BLOCK DESIGN SUBTESTS

This section reviews the measurement characteristics of the Vocabulary (Voc.) and Block Design (BD) subtests of the Wechsler Intelligence Scale for Children (WISC), both as a separate unit and as a WISC short form. It also reviews behavioral correlates of intelligence as reported in the literature and critically evaluates the appropriateness of their use in Cycle II of the Health Examination Survey.

The selection of the Vocabulary and Block Design subtests for use as part of the psychological test battery for Cycle II, in effect, treats these subtests as a short form of the WISC. In addition to providing an estimate of the WISC score, the two subtests may be interpreted separately, in combination with other test scores, or in conjunction with other Survey data. Combinations of these measures with other data obtained in the Survey are discussed in section II.

DESCRIPTION OF THE WISC

The WISC, which was published in 1949, extended the well-known Wechsler intelligence scales for adolescents and adults into the childhood range of 5 to 15 years. During the decade and a half since its publication the WISC has been the subject of extensive investigation and has achieved wide school and clinic use where individual measures of intelligence are desired.

The WISC is patterned after the Wechsler-Bellevue Intelligence Scale both in the structure of the subtests and the scales and in the use of the deviation intelligence quotient. The test consists of 12 subtests—6 Verbal and 6 Performance—of which 2 (Digit Span of the Verbal Scale and Mazes of the Performance Scale) are supplementary and not routinely used. The 5 subtests comprising the Verbal Scale are as follows:

Information, Comprehension, Arithmetic, Similarities, and Vocabulary. The 5 Performance Scale subtests are Picture Completion, Picture Arrangement, Block Design, Object Assembly, and Coding (Digit Symbols).

An important innovation in the Wechsler intelligence tests is the use of the deviation IQ. This device supplants the *mental age* concept and evaluates the performance of each individual on the basis of the distribution of scores of a representative sample of his own chronological age. In the standardization of the WISC, Wechsler kept the standard deviation of intelligence quotients constant from year to year, with the result that "a child's obtained IQ does not vary unless his actual test performance as compared with his peers varies."

Raw scores for each subtest are converted to *scaled scores* which have a mean of 10 and standard deviation of 3 for each age level. The sum of five scaled scores for the Verbal Series constitutes the Verbal Scale score (VS), and similarly the Performance Scale score (PS) is the sum of the five Performance Series scaled scores. The Full Scale score (FS) is the sum of the Verbal Scale and the Performance Scale. Deviation intelligence quotients have been derived by a similar conversion process for VS, PS, and FS. The IQ scales at each age have a mean of 100 and standard deviation of 15.

The standardization of the WISC is reported in Wechsler's manual (101), and the standardization sample is summarized in terms of age, sex, geographic representation, urban-rural composition, and composition by socioeconomic status (reflected by occupation of fathers). The WISC was standardized on a total sample of 2,200 cases, including 100 white boys and 100 white girls at each age from 5 to 15 years. The proportion of urban children in the sample was slightly higher than in comparable United States population statistics.

Reviewers have commented very favorably on the WISC as a test of superior quality (102-104), but, as in all areas of mental measurement, imperfections have been noted and users have attempted to employ it for purposes for which it was not specifically designed. In general, the deviation IQ has been accepted as an improvement over the IQ computed by dividing mental age by chronological age. Except for a slight bias for urban and smalltown areas—as opposed to rural areas—for a native white population, the sampling basis of the WISC has been regarded as good.

Maxwell (106), and also Wilson (139), has criticized the linearity of the transformation of raw scores to scaled scores, which may be a problem when sampling extreme cases and widely varying regional, ethnic, and linguistic groups. Hite (112) reported that the WISC lacks items of middle-range difficulty at all age levels and is too difficult for young children, particularly those in the age range 5 to 6 years. In the studies reviewed, WISC Full Scale IQ's have indeed tended to be lower than comparable Stanford-Binet IQ's. This is especially true at the lower age levels. McCandless (103) noted that girls tend to test lower than boys on the WISC, but support for this generalization is equivocal in the present review.

In evaluating the utility of the Vocabulary and Block Design short form of the WISC for the Survey it is appropriate to consider shortcomings of these tests in relation to alternatives that might have been considered—given the constraints of testing time available in the Survey schedule and the general problems of a national survey. It may be noted that although the WISC norms are inappropriate in varying degrees for Negro, bilingual and foreign-born, illiterate, retarded, defective, rural, and other special groups for which the test was not designed, there is no adequate measure that can be applied to all. On the other hand, because of the extensive research on the WISC, reported below, it may be possible to estimate errors in the Vocabulary and Block Design subtests and in the scores derived from them for various components of the Survey sample. In addition, relationships of these variables to the Goodenough Draw-A-Man Test offer further opportunities for compensatory analysis.

RESEARCH ON SHORT FORMS OF THE WISC

Several investigators have combined two or more subtests in order to develop an efficient short form of the WISC that correlates well with the Full Scale and produces comparable means and standard deviations (175-179, 231, and 235). Of these, only one article, by Simpson and Bridges (177), reported favorable results with the combination of Vocabulary and Block Design. They used

a sample of 120 children over the age range of 65 to 192 months.

Finley and Thompson (231) developed for a sample of 309 mentally retarded persons a short form with five subtests, including Block Design, which correlated 0.89 with FS IQ. Significantly, their report included correlations of 0.55 and 0.45, respectively, for Voc. and BD with FS IQ, while the correlation of Voc. and BD was only 0.1. Further, estimation of mean FS IQ by proration of the sum of Voc. and BD, as reported by these authors, approximated the actual FS IQ quite closely.

Schwartz and Levitt (235) also reported a short form of the WISC for educable retarded children, consisting of six subtests including Voc. and BD which correlated 0.95 with FS IQ. However, their best combination of five subtests, which reduced the correlation to 0.92, eliminated Block Design. Osborne and Allen (239), on the other hand, cross-validated two triads of WISC subtests including Voc. and BD, one with Picture Completion and one with Picture Arrangement, using samples of 240 (initial) and 50 (validation) retarded children aged 7 to 14 years, with correlations with FS IQ of 0.88 to 0.90.

At the same time, Hite (112) has confirmed Wechsler's data (101) indicating that Vocabulary and Block Design are the most reliable subtests in the WISC battery. Hagen (109) and Cohen (111) in the United States and Gault (110) in Australia have reported that both of these subtests are highly loaded on the general factor obtained in factor analysis of the WISC over the entire age range of 5 to 15 years. Cohen found that Vocabulary was the strongest single measure of the general factor. Nevertheless, a problem exists in determining the optimal combination of these subtests to estimate the FS IQ and various parameters related to the Survey objectives.

Simpson and Bridges (177) estimated the FS IQ on the basis of a simple sum of the scaled scores of Voc. and BD and reported a conversion table for this purpose. Inasmuch as their results have not been replicated, so far as is known, cross-validation on a substantial sample should be considered before this table is adopted. The importance of this recommendation is illustrated by some computations based on the Finley and Thompson data (231). The sum of mean Voc. and BD scaled scores, 11, multiplied by 5 to prorate the FS score, gives a WISC Full Scale IQ of 70 (as compared with the actual mean of 68), while the score of 11 in the Simpson and Bridges tables yields an FS IQ of 77. Further, in view of Maxwell's criticism of the transformation of raw scores to scaled scores (106), it may be advisable also to explore empirically the alternative of predicting the FS IQ from raw scores.

In reviewing the WISC literature every effort was made to focus on the Voc. and BD subtests, and considerable data have been assembled. Nevertheless, the major portion of the information referred to in this report is based on the full test, and assumptions of equivalence of short form scores to the Full Scale must be made in generalizing the results reported. As indicated above, this assumption is not entirely inappropriate, but caution is certainly indicated.

RELIABILITY AND STABILITY

Wechsler's manual (101, p. 13) reported corrected split-half reliability coefficients of 0.77, 0.91, and 0.90, respectively, for Vocabulary, and 0.84, 0.87, and 0.88, respectively, for Block Design for samples of 200 children at each of the following age levels: 7 1/2, 10 1/2, and 13 1/2 years. The corresponding FS reliabilities were 0.92, 0.95, and 0.94, respectively. As noted above, these two subtests were the most reliable of all the WISC subtests. These results for Voc. and BD have been confirmed by Hite (112) for children in the age range of 5 to 7 years.

Stability of the WISC on retest has also been found satisfactory by Gehman and Matyas (113) over a 4-year period (age 11 years at initial test), by Reger (115), who tested a sample at ages 10, 11, and 12 years, and by Whatley and Plant (116), who used a 17-month interval. In these studies, retest correlations were generally of the order of the corrected split-half reliabilities. These and related data are summarized in table 1.

VALIDITY

Despite the fact that Wechsler developed the WISC in protest against the measurement concept of *mental age* (and the IQ based on it) implicit in the Stanford-Binet test, and despite the additional

Table 1. Studies reporting reliability coefficients of the WISC

Investigator	Year	Subjects[a]	Age range	Number			Coefficient					Type of coefficient
				Σ	M	F	Voc.	BD	VS	PS	FS	
Throne, Schulman, and Kaspar (227).	1962	Retarded	11-0 - 14-11	39	39	-	0.79	0.82	0.92	0.89	0.95	Test-retest
Armstrong (175)	1955	Guidance clinic	5-0 - 14-11	200	100	100	0.94	N.R.	N.R.	N.R.	N.R.	Split-half, Spearman-Brown
			5-7 years	20	20	-	0.92	N.R.	N.R.	N.R.	N.R.	
			5-7 years	20	-	20	0.90	N.R.	N.R.	N.R.	N.R.	
			7-9 years	20	20	-	0.93	N.R.	N.R.	N.R.	N.R.	
			7-9 years	20	-	20	0.91	N.R.	N.R.	N.R.	N.R.	
			9-11 years	20	20	-	0.87	N.R.	N.R.	N.R.	N.R.	
			9-11 years	20	-	20	0.89	N.R.	N.R.	N.R.	N.R.	
			11-13 years	20	20	-	0.88	N.R.	N.R.	N.R.	N.R.	
			11-13 years	20	-	20	0.88	N.R.	N.R.	N.R.	N.R.	
			13-15 years	20	20	-	0.90	N.R.	N.R.	N.R.	N.R.	
			13-15 years	20	-	20	0.96	N.R.	N.R.	N.R.	N.R.	
Gehman and Matyas (113).	1956	Normals	11-1	60	29	31	N.R.	N.R.	0.77	0.74	0.77	Test-retest[b]
Caldwell (252)	1954	Normals (Negro)	9-7 - 10-6	60	---	---	0.70	0.89	0.82	0.90	0.84	Split-half
Jones (154)	1962	Normals (England)		240	120	120						Split-half, Kuder-Richardson
			7-6 - 8-5	80	40	40	0.70	0.74	0.86	0.80	0.89	
			8-6 - 9-5	80	40	40	0.70	0.68	0.87	0.81	0.90	
			9-6 - 10-5	80	40	40	0.70	0.75	0.90	0.85	0.94	
Wechsler (101)	1949	Normals (WISC standardization data).		600	300	300						Split-half, Spearman-Brown
			7-6	200	100	100	0.77	0.84	0.88	0.86	0.92	
			10-6	200	100	100	0.91	0.87	0.96	0.89	0.95	
			13-6	200	100	100	0.90	0.88	0.96	0.90	0.94	
Hite (112)	1953	Normals		200	117	83						Split-half
			5-6	50	34	16	0.71	0.77	0.77	0.81	0.90	
			6-6	100	56	44	0.72	0.84	0.89	0.89	0.91	
			7-6	50	27	23	0.76	0.89	0.89	0.86	0.94	
Hagen (109)[c]	1952	Normals (WISC standardization data).		400	200	200						Split-half, Spearman-Brown
			5 years	200	100	100	0.68	0.77	N.R.	N.R.	N.R.	
			15 years	200	100	100	0.91	0.89	N.R.	N.R.	N.R.	

[a]Designations of subjects are always white Americans unless otherwise specified.
[b]Time between testings was 49 months.
[c]Data are from the WISC standardization sample, but were not reported in the WISC manual.

NOTES: All correlation coefficients are Pearson Product-Moment unless otherwise specified.

Σ —Total population; M—male; F—female; Voc.—Vocabulary; BD—Block Design; VS—Verbal Scale; PS—Performance Scale; FS—Full Scale; N.R.—not reported.

Table 2. Studies reporting correlation between the WISC and Stanford-Binet[a]

Investigator	Year	Subjects[b]	Age range	Number			Correlation				
				Σ	M	F	Voc.	BD	VS	PS	FS
Nale (216)	1951	Mental defectives	8-10 - 15-11	104	54	50	N.R.	N.R.	N.R.	N.R.	0.91
Stacey and Levin (228)	1951	Mental defectives	7-2 - 15-11	70	---	---	N.R.	N.R.	N.R.	N.R.	0.68
Sloan and Schneider (217)	1951	Mental defectives	N.R.	40	20	20	N.R.	N.R.	0.75	0.64	0.76
Orr (188)	1950	Retarded	N.R.	10	---	---	N.R.	N.R.	[c]0.81	[c]0.49	[c]0.71
Sharp (229)	1957	Slow learners	8-0 - 16-5	50	---	---	N.R.	N.R.	0.62	0.67	0.69
Post (198)	1952	Stutterers	5-5 - 15-10	30	27	3	N.R.	N.R.	0.80	0.37	0.78
Kent and Davis (207)	1957	Normals and clinic referrals (England)	8-12 years	213	133	80	N.R.	N.R.	N.R.	0.58	N.R.
		Normals		118	59	59					
		Delinquents		55	48	7					
		Psychiatric outpatients		40	26	14					
Muhr (119)	1952	Institutional (orphans and various problems)	5-0 - 6-11	42	---	---	N.R.	N.R.	0.46	0.52	0.62
			5 years	21	---	---	N.R.	N.R.	0.65	0.66	0.74
			6 years	21	---	---	N.R.	N.R.	0.44	0.39	0.49
Davidson (162)	1954	Normals	14-0 - 14-3	30	---	---	N.R.	N.R.	0.79	0.71	0.83
Kardos (161)	1954	Normals	11-11 - 13-0	100	50	50	N.R.	N.R.	0.87	0.82	0.89
Matyas[d] (114)	1954	Normals		60	29	31					
		Grade 5	11-1 (mean)	60	29	31	N.R.	N.R.	0.78	0.46	0.73
		Grade 9 (retest)	15-2 (mean)	60	29	31	N.R.	N.R.	0.76	0.64	0.77
Raleigh (191)	1952	Normals	10-8 - 14-9	100	52	48	N.R.	N.R.	0.77	0.59	0.80
Schwitzgoebel (189)	1952	Normals	9-11 - 13-8	100	52	48	N.R.	N.R.	0.78	0.61	0.84
Clarke (160)	1950	Normals	9-7 - 12-9	84	39	45	N.R.	N.R.	0.83	0.57	0.79
Frandsen and Higginson (159)	1951	Normals	9-1 - 10-3	54	---	---	N.R.	N.R.	0.71	0.63	0.80
Reidy (171)	1952	Normals	9-0 - 11-11	60	30	30	N.R.	N.R.	0.87	0.69	0.86
Jones (154)	1962	Normals (England)	8-10 years	240	120	120	N.R.	N.R.	0.84	0.59	0.81
			8 years	40	40	-	N.R.	N.R.	0.77	0.48	0.72
			8 years	40	-	40	N.R.	N.R.	0.79	0.46	0.76
			Σ 8 years	80	40	40	N.R.	N.R.	0.78	0.47	0.74
			9 years	40	40	-	N.R.	N.R.	0.89	0.65	0.90
			9 years	40	-	40	N.R.	N.R.	0.78	0.58	0.75
			Σ 9 years	80	40	40	N.R.	N.R.	0.84	0.61	0.84
			10 years	40	40	-	N.R.	N.R.	0.86	0.64	0.83
			10 years	40	-	40	N.R.	N.R.	0.90	0.67	0.86
			Σ 10 years	80	40	40	N.R.	N.R.	0.88	0.66	0.85
Arnold and Wagner (158)	1955	Normals	8-9 years	50	---	---	N.R.	N.R.	0.85	0.75	0.88
Wagner (156)	1951	Normals	8-9 years	50	---	---	N.R.	N.R.	0.77	0.87	0.81
Scott (155)	1950	Normals	7-7 - 11-1	30	---	---	0.63	0.60	0.86	0.86	0.92
Beeman (153)	1960	Normals	7-2 - 11-9	36	---	---	N.R.	N.R.	0.64	0.42	0.67
Harlow, Price, Tatham, and Davidson (145).	1957	Normals		60	---	---					
			6-6 - 6-7	30	---	---	N.R.	N.R.	0.64	0.61	0.64
			10-0 - 10-1	30	---	---	N.R.	N.R.	0.88	0.52	0.83
Cohen and Collier (124)	1952	Normals	6-5 - 8-9	51	---	---	N.R.	N.R.	0.82	0.80	0.85
Tatham (152)	1952	Normals	6-5 - 6-7	30	---	---	N.R.	N.R.	0.64	0.51	0.64
Mussen, Dean, and Rosenberg (117).	1952	Normals	6-0 - 13-1	39	---	---	N.R.	N.R.	0.83	0.72	0.85

See footnotes at end of table.

Table 2. Studies reporting correlation between the WISC and Stanford-Binet[a]—Con.

Investigator	Year	Subjects[b]	Age range	Number			Correlation				
				Σ	M	F	Voc.	BD	VS	PS	FS
Krugman, Justman, Wrightstone, and Krugman (144)	1951	Normals		222	---	---					
			6 years	38	---	---	N.R.	N.R.	0.73	0.74	0.82
			7 years	43	---	---	N.R.	N.R.	0.64	0.49	0.73
			8 years	44	---	---	N.R.	N.R.	0.78	0.57	0.82
			9 years	31	---	---	N.R.	N.R.	0.83	0.79	0.87
			10 years	29	---	---	N.R.	N.R.	0.88	0.54	0.86
			11 years	37	---	---	N.R.	N.R.	0.69	0.53	0.76
Pastovic[d] (121)	1951	Normals		100	---	---					
			5-6	50	---	---	N.R.	N.R.	0.63	0.57	0.71
			7-6	50	---	---	N.R.	N.R.	0.82	0.71	0.88
Winpenny (105)	1951	Normals		185	---	---					
		Kindergarten	5-4 - 5-8	50	---	---	N.R.	N.R.	N.R.	N.R.	0.71
		Grade 2	7-4 - 7-8	50	---	---	N.R.	N.R.	N.R.	N.R.	0.88
		Grade 5	9-7 - 12-9	85	---	---	N.R.	N.R.	N.R.	N.R.	0.79
Dunsdon and Roberts (170)	1955	Normals (England)	5-0 - 14-11	1,947	980	967					
				980	980	-	N.R.	N.R.	N.R.	N.R.	0.82
				967	-	967	N.R.	N.R.	N.R.	N.R.	0.77
oruszak (146)	1954	Normals	5-14 years	80	40	40	N.R.	N.R.	0.87	0.78	0.90
			5-14 years	40	40	-	N.R.	N.R.	0.89	0.72	0.93
			5-14 years	40	-	40	N.R.	N.R.	0.86	0.71	0.93
olland (149)	1953	Normals	5-13 years	52	---	---	N.R.	N.R.	0.88	0.73	0.87
eider, Noller, and Schraumm (150)	1951	Normals	5-0 - 11-11	106	---	---	N.R.	N.R.	0.89	0.77	0.89
			5-0 - 7-11	44	---	---	N.R.	N.R.	0.82	0.79	0.90
			8-0 - 11-11	62	---	---	N.R.	N.R.	0.92	0.78	0.90
ureth, Muhr, and Weisgerber (118)	1952	Normals	5-6 years	100	---	---	0.51	0.61	0.75	0.71	0.81
			5 years	50	---	---	0.42	0.65	0.79	0.73	0.84
			6 years	50	---	---	0.65	0.55	0.71	0.71	0.79
ottersman (151)	1950	Normals	6 years	50	21	29	N.R.	N.R.	0.71	0.49	0.71
riggs and Cartee (148)	1953	Normals (S-B, Form M)	5 years	46	---	---	N.R.	N.R.	0.58	0.48	0.61
rr (188)	1950	Normals		40	---	---					
		Grade 1	N.R.	15	---	---	N.R.	N.R.	[c]0.63	[c]0.62	[c]0.77
		Grade 4	N.R.	14	---	---	N.R.	N.R.	[c]0.64	[c]0.65	[c]0.67
		Grade 7	N.R.	11	---	---	N.R.	N.R.	[c]0.88	[c]0.66	[c]0.79
Stanley (157)	1955	Normals (from Frandsen and Higginson, 159, above)	N.R.	50	---	---	N.R.	N.R.	N.R.	N.R.	[f]0.71
Schachter and Apgar (147)	1958	Normals, mixed sample	N.R.	113	61	52	N.R.	N.R.	[g]0.64	[g]0.48	[g]0.67
		White		39	---	---					
		Negro		66	---	---					
		Puerto Rican		6	---	---					
		Oriental		2	---	---					
Estes, Curtin, DeBurger, and Denny (125)	1961	Normals, Grades 1-8		82	47	35					
		Form L	N.R.	82	47	35	N.R.	N.R.	N.R.	N.R.	0.80
		Form L-M	N.R.	82	47	35	N.R.	N.R.	N.R.	N.R.	0.74

[a]Unless otherwise noted, Stanford-Binet, Form L.
[b]Designation of subjects are always white Americans unless otherwise specified.
[c]Rank difference correlation.
[d]Also reported by Gehman and Matyas in 1956.
[e]Also reported by Pastovic and Guthrie in 1951.
[f]Intraclass correlation.
[g]Average time between S-B and WISC administration was 50.8 months.

NOTES: All correlation coefficients are Pearson Product-Moment unless otherwise specified.

Σ —Total population; M—male; F—female; Voc.—Vocabulary; BD—Block Design; VS—Verbal Scale; PS—Performance Scale; FS—Full Scale; N.R.—not reported.

Table 3. Studies reporting correlation between the WISC and other measures

Investigator	Year	Test or criterion variable	Subjects[a]	Age range	Number			Correlation				
					Σ	M	F	Voc.	BD	VS	PS	FS
Smith (126)	1961	Full Range Picture Vocabulary Test.	Normals	6-11 - 8-10	100	51	49	N.R.	N.R.	0.63	0.42	0.60
McBrearty (123)	1951	Arthur Point Scale of Performance Tests.	Normals	10-3 - 12-11	52	22	30	N.R.	N.R.	N.R.	0.65	0.71
Cohen and Collier (124).	1952	Arthur Point Scale of Performance Tests.	Normals	6-5 - 8-9	49	---	---	N.R.	N.R.	[b]0.77	[b]0.81	[b]0.80
Winpenny (105)	1951	Arthur Point Scale of Performance Tests.	Normals	9-7 - 12-9	85	---	---	N.R.	N.R.	N.R.	N.R.	0.70
Armstrong and Hauck (130).	1960	Visual Motor Gestalt Test.	Nonorganic child guidance population.	6-12 years	98	49	49	N.R.	N.R.	-0.22	-0.07	-0.23
Winpenny (105)	1951	Bernreuter-Winpenny-	Normals									
			Kindergarten	5-4 - 5-8	50	---	---	N.R.	N.R.	N.R.	N.R.	0.92
			Grade 2	7-4 - 7-8	50	---	---	N.R.	N.R.	N.R.	N.R.	0.92
			Grade 5	9-7 - 12-9	85	---	---	N.R.	N.R.	N.R.	N.R.	0.97
Cooper (242)	1958	California Achievement Tests.	Bilinguals (Guam), Grade 5.	N.R.	51	---	---	N.R.	N.R.	0.80	0.54	0.77
Altus (122)	1952	California Test of Mental Maturity.	Normals, junior high.	N.R.	55	---	---	N.R.	N.R.	N.R.	N.R.	0.81
Altus (134)	1955	California Test of Mental Maturity	Retarded, elementary school.	N.R.	100	71	29					
		Language			---	---	---	N.R.	N.R.	0.71	0.57	0.70
		Non-language			---	---	---	N.R.	N.R.	0.65	0.67	0.68
		Total			---	---	---	N.R.	N.R.	0.76	0.68	0.77
Cooper (242)	1958	California Test of Mental Maturity.	Bilinguals (Guam), Grade 5.	N.R.	51	---	---	N.R.	N.R.	0.66	0.68	0.74
Schwitzgoebel (189).	1952	California Test of Mental Maturity.	Normals	9-11 - 13-8	100	52	48	N.R.	N.R.	0.55	0.59	0.75
Barratt (138)	1956	Columbia Mental Maturity Scale.	Normals	9-2 - 10-1	60	26	34	[c]0.45	[c]0.47	[c]0.56	[c]0.48	[c]0.61
Warren and Collier (224).	1960	Columbia Mental Maturity Scale.	Retarded	9-30 years	49	---	---	N.R.	N.R.	N.R.	N.R.	0.68
Thompson (193)	1961	Gates Advanced Primary Reading Tests.	Normals	6-4 - 8-0	105	62	43					
		Word Recognition			---	---	---	N.R.	N.R.	0.58	0.42	0.55
		Paragraph Reading			---	---	---	N.R.	N.R.	0.55	0.46	0.56
		Composite Reading			---	---	---	N.R.	N.R.	0.57	0.47	0.58
Warren and Collier (224).	1960	Goodenough Intelligence Test.	Retarded	9-30 years	49	---	---	N.R.	N.R.	N.R.	N.R.	0.43
Armstrong and Hauck (130).	1960	Goodenough Intelligence Test.	Child guidance clinic.	6-12 years	98	49	49	N.R.	N.R.	0.37	0.51	0.49
Rottersman (151)	1950	Goodenough Intelligence Test.	Normals	6 years	50	21	29	N.R.	N.R.	0.38	0.43	0.47
Kimbrell (136)	1960	Grade placement	Mental defectives.	10.5 - 15.8	62	---	---	N.R.	N.R.	N.R.	N.R.	0.40
Smith (126)	1961	Wide Range Achievement Test.	Normals	6-11 - 8-10	100	51	49	N.R.	N.R.	0.55	0.47	0.61
Delp (135)	1953	Kent EGY Test	Normals	6-15 years	74	---	---	N.R.	N.R.	0.60	0.55	0.62
Cooper (242)	1958	Leiter International Performance Scale.	Bilinguals (Guam), Grade 5.	N.R.	51	---	---	N.R.	N.R.	0.73	0.78	0.83
Sharp (229)	1957	Leiter International Performance Scale.	Slow learners	8-0 - 16-5	50	---	---	N.R.	N.R.	0.78	0.80	0.83

See footnotes at end of table.

Table 3. Studies reporting correlation between the WISC and other measures—Con.

Investigator	Year	Test or criterion variable	Subjects[a]	Age range	Number			Correlation				
					Σ	M	F	Voc.	BD	VS	PS	FS
Alper (221)-------	1958	Leiter International Performance Scale.	Mental defectives.	7-2 - 17-3	30	15	15	N.R.	N.R.	0.40	0.79	0.77
Dunn and Brooks (234).	1960	Peabody Picture Vocabulary Test.	Retarded---------	N.R.	56	---	---	N.R.	N.R.	N.R.	N.R.	0.61
Kimbrell (136)----	1960	Peabody Picture Vocabulary Test.	Mental defectives.	10.5 - 15.8	62	---	---	N.R.	N.R.	N.R.	N.R.	0.30
Himelstein and Herndon (137).	1962	Peabody Picture Vocabulary Test.	Emotionally disturbed.	6-2 - 14-8	48	---	---	N.R.	N.R.	0.64	0.52	0.63
McBrearty (123)---	1951	Progressive Achievement Tests.	Normals----------	10-3 - 12-11	52	22	30	N.R.	N.R.	0.78	0.50	0.81
Dunsdon and Roberts (170).	1955	Mill Hill Vocabulary Scale.	Normals (England).	5-0 - 14-11	1947	980	967					
		Form A----------	[b]----------	----------	980	980	-	0.83	N.R.	N.R.	N.R.	N.R.
		Form A----------	----------	----------	967	-	967	0.81	N.R.	N.R.	N.R.	N.R.
		Form B----------	----------	----------	980	980	-	0.85	N.R.	N.R.	N.R.	N.R.
		Form B----------	----------	----------	967	-	967	0.82	N.R.	N.R.	N.R.	N.R.
Brown, Hakes, and Malpass (233).	1959	Raven Progressive Matrices.	Retarded---------	N.R.	N.R.	---	---	N.R.	N.R.	N.R.	N.R.	0.39-0.49
Malpass, Brown, and Hakes (140).	1960	Raven Progressive Matrices.	Retarded---------	11-8 (mean)	104	---	---	N.R.	N.R.	N.R.	N.R.	[d]0.51
Barratt (138)-----	1956	Raven Progressive Matrices.	Normals----------	9-2 - 10-1	60	26	34	[c]0.56	[c]0.60	[c]0.69	[c]0.70	[c]0.75
Wilson (139)------	1952	Raven Progressive Matrices.	British Columbia	5-6 - 13-0	90	---	---	N.R.	N.R.	N.R.	N.R.	------
			Hospitalized Americans Indians.	----------	30	---	---	------	------	------	------	[c]0.75 [f]0.27
			Hospitalized whites.	----------	30	---	---	------	------	------	------	[c]0.83 [f]0.42
			High socioeconomic whites.	----------	30	---	---	------	------	------	------	[c]0.81 [f]0.49
Martin and Wiechers (142).	1954	Coloured Progressive Matrices.	Normals----------	9-0 - 10-0	100	60	40	0.73	0.74	0.84	0.83	0.91
Stacey and Carleton (141).	1955	Coloured Progressive Matrices.	Mental defectives.	7-5 - 15-9	150	---	---	N.R. [c]0.36	N.R. [c]0.41	[c]0.54 [c]0.51	[c]0.52 [c]0.55	0.55 [c]0.62
Hite (112)--------	1953	SRA Primary Mental Abilities Test.	Normals----------	5-6 years	50	34	16	------	------	------	------	------
		Verbal----------	----------	----------	------	------	------	0.45	0.38	N.R.	N.R.	N.R.
		Perception------	----------	----------	------	------	------	0.30	0.83	N.R.	N.R.	N.R.
		Quantitative----	----------	----------	------	------	------	0.35	0.53	N.R.	N.R.	N.R.
		Space-----------	----------	----------	------	------	------	0.39	0.68	N.R.	N.R.	N.R.
Stempel (143)-----	1953	SRA Primary Mental Abilities.	Superior intelligence.	8-5 - 10-4	50	---	---	------	------	------	------	------
		Space-----------	----------	----------	------	------	------	N.R.	N.R.	0.45	0.34	N.R.
		Number----------	----------	----------	------	------	------	N.R.	N.R.	0.15	0.38	N.R.
		Reasoning-------	----------	----------	------	------	------	N.R.	N.R.	0.63	0.55	N.R.
		Perception------	----------	----------	------	------	------	N.R.	N.R.	0.18	0.42	N.R.
		Verbal----------	----------	----------	------	------	------	N.R.	N.R.	0.68	0.40	N.R.
		IQ----------	----------	----------	------	------	------	N.R.	N.R.	N.R.	N.R.	0.68
Jones (154)-------	1962	Teacher ratings---	Normals (England).	7-6 - 10-5	240	120	120	N.R.	N.R.	0.73	0.57	0.74
				8 years	80	40	40	N.R.	N.R.	0.70	0.48	0.70
				9 years	80	40	40	N.R.	N.R.	0.71	0.59	0.73
				10 years	80	40	40	N.R.	N.R.	0.76	0.62	0.76
Stark (163)-------	1954	The Drawing-Completion Test.	Normals----------	8-4 - 9-10	50	30	20	0.72	0.49	N.R.	N.R.	0.79
Bacon (127)-------	1954	Wechsler-Bellevue Intelligence Scale, Form I.	Normals----------	11-9 - 12-3	32	16	16	0.84	0.65	0.86	0.65	0.77
Delattre and Cole (128).	1952	Wechsler-Bellevue Intelligence Scale, Form I.	Normals----------	10-5 - 15-7	50	---	---	0.55	0.49	0.86	0.82	0.87

[a]Designation of subjects are always white Americans unless otherwise specified. [b]ETA coefficient.
[c]WISC scaled scores. [d]Partial correlations with chronological age removed.
[e]Raw scores. [f]Scaled scores.

NOTES: All correlation coefficients are Pearson Product-Moment unless otherwise specified.

Σ—Total population; M—male; F—female; Voc.—Vocabulary; BD—Block Design; VS—Verbal Scale; PS—Performance Scale; FS—Full Scale; N.R.—not reported.

fact that the validity of the WISC must be judged principally in relation to the logic of Wechsler's approach and the adequacy of his development and standardization of the test, a surprisingly large number of papers dealing with the validity of the WISC have used the Stanford-Binet as a criterion. As may be expected, unless one assumes naïvely that the theoretical objections to mental age scores involve gross discrepancies, which they usually do not, the correlations between WISC Full Scale IQ's and Stanford-Binet IQ's are generally high, in about the same range as the respective reliabilities of these tests. (See table 2.) There seems to be little doubt that both the WISC and the Stanford-Binet merit their reputations as outstanding individual intelligence tests.

There are, however, differences between the WISC and Stanford-Binet in score levels. As noted above, the WISC IQ's tend to be substantially lower than the corresponding Stanford-Binet IQ's for the very young and for the gifted (153 and 215), as well as for many samples reported across the normal range (119, 120, 124, 147, 148, 151, 154, 156, 159, and 161). This problem is discussed below.

The WISC has been correlated with a wide range of verbal and performance tests that purport to measure various aspects of *intelligence.* Correlations with the Wechsler-Bellevue, Form I, have been reported by Bacon (127) for a sample of 36 children in the age range 11 years 9 months to 12 years 3 months and by Delattre (128) for 50 students aged 10-5 to 15-7. Their results for FS were 0.77 and 0.87, respectively, while both correlated 0.86 for VS. For PS their respective correlations were 0.65 and 0.82; for Voc., 0.84 and 0.55. Finally, for BD their results were 0.65 and 0.49. Variations of the magnitude indicated must be expected for small samples from different settings. Dunsdon and Roberts (170) administered four vocabulary tests including the WISC to 2,000 British children and obtained intercorrelations exceeding 0.8 for both sexes.

Table 3 summarizes reported correlation coefficients between WISC scores and other tests of intelligence, mental maturity, and achievement in school subjects, teacher ratings, and related criteria. For the FS IQ these are generally quite high and positive, considering sample size and variation in sample composition and setting. In view of these variations, the specific coefficients are of less interest than the general trend, which supports the validity of the WISC as a general measure of what Wechsler labels "the total effective intelligence of the individual" (101, pp. 4 and 5).

For the purposes of a national survey, the robustness of the validity data over wide sample fluctuations is very encouraging, as is revealed by its use on samples of varying geographic and ethnic characteristics, of varying abilities ranging from defective to gifted samples, and by its use with special groups such as retarded readers (133), bilinguals (242), stutterers (198), and low school achievers (190).

FACTORS AFFECTING WISC SCORES

Both qualitative and quantitative variations in WISC scores have been reported by various investigators in relation to a wide range of factors. Those discussed in this section are considered relevant to the objectives and problems of the Survey. Where feasible and appropriate, implications and recommendations are noted.

Anxiety

Hafner, Pollie, and Wapner (132) and Carrier, Orton, and Malpass (205) have both reported negative correlations between the WISC FS and the Children's Manifest Anxiety Scale (CMAS), indicating that anxiety, as measured by this scale, tends to interfere with effective WISC performance. Hafner and others found a significant correlation of -0.31 between CMAS and BD. The Carrier study observed the relationship (-0.54) over a range of ability but not among the exceptionally bright. It appears to be most marked in the subnormal; Feldhusen and Klausmeier (167) found the following mean differences in CMAS scores for three groups at different IQ levels: low IQ, 20.2; average, 14.8; and high, 12. These results are not entirely consistent with those of Burns (206), however, who found similar correlations between WISC Vocabulary and California Personality Test measures of Social Adjustment (0.55) and Personal Adjustment (0.45) but obtained nonsignificant coefficients of 0.12 and 0.10, respectively, for Block Design.

Although anxiety and adjustment may be regarded generally as factors that tend to depress WISC (Voc. and BD) scores for some segments of the child population on some occasions, it would seem unwise to attempt any correction for these factors. Presumably, some valid evidence on adjustment will become available from the Thematic Apperception Test (TAT), the School Information Form, and the extensive background and medical information being collected in the Health Examination Survey. However, the relationships are not clearly enough defined for fine quantitative manipulation. One alternative is to regard fluctuations on these variables as a source of error which may possibly be crudely estimated later but is probably well randomized in the total sample. Another is to accept the error pragmatically with the attitude that depressed scores resulting from affective factors probably reflect depressed ability of the individual to function effectively.

Sex Differences

The statement by McCandless (103), cited earlier, that boys do better on the WISC than girls, is not supported by the present review. Data on sex differences are presented in nine studies (130, 146, 154, 169, 175, 192, 194, 196, and 232), and only one (130) reports a significant mean difference favoring boys on FS IQ. However, none of them employed a sampling design encouraging confidence in the group comparisons.

Some correlational differences mentioned by several authors do appear interesting: The correlation of WISC Full Scale IQ with Bender-Gestalt was negative and higher for boys (-0.34 $p<0.01$) than for girls (-0.09 ns) (130). The correlation of WISC Full Scale IQ with the Ammons Picture Vocabulary Test was 0.71 for boys and 0.45 for girls (169). The correlations of WISC FS and VS IQ's with the spelling subtest of the Iowa Test of Basic Skills were higher for boys than for girls. No data were reported in which sex differences favored girls. The absence of sex differences in studies of normal American (146) and English (154) children, deaf American (194) and English (196) children, and retarded American children (232) suggests considerable generality for the negative conclusion.

Qualitative Differences by Level

Gallagher and Lucito (164) found a negative rank order between the mean scores of gifted and retarded children on the WISC. The three highest and three lowest subtests for five comparison groups in their study are shown below. These results agree with others, to be discussed below, which indicate that Block Design scores are least affected by population variations, in contrast with Vocabulary, which is the highest test of the gifted groups and the lowest of the retarded.

Baroff (223) described a WISC profile for a sample of 53 low-IQ patients with a mean FS IQ of 63; Block Design was highest, and Vocabulary ranked 11 out of 12. Although Fisher (225) failed to verify the Baroff patterning, Baroff's results are in agreement with those of Gallagher and Lucito with respect to Vocabulary. Matthews (230) found that nonachievers in school tend to be higher on Block Design than on Vocabulary. Levinson (243 and 244), working with Jewish children in New York, and Altus (240), with Mexican and Anglo-American children in California, both found that monolinguals exceeded bilinguals on Vocabulary, but that the differences on Block Design

	Group classification	Number of subjects (N)	Three highest subtests	Three lowest subtests
1	Gifted------	50	Similarities, Information, Vocabulary	Picture Completion, Picture Arrangement, Digit Span
2	Gifted------	43	Vocabulary, Information, Similarities	Picture Completion, Picture Arrangement, Digit Span
3	Average-----	565	Arithmetic, Digit Symbol, Picture Arrangement	Block Design, Information, Similarities
4	Retarded----	150	Object Assembly, Picture Completion, Digit Span	Information, Vocabulary, Arithmetic
5	Retarded----	52	Object Assembly, Digit Span, Picture Completion	Vocabulary, Information, Picture Arrangement

were not significant. Burks and Bruce (186) found that poor readers score significantly high on Block Design, and Kallos, Grabow, and Guarino (180) obtained a significant difference between Block Design and Vocabulary, favoring Block Design, for a sample of poor readers.

Results such as these suggest the possibility of investigating a Voc.-BD ratio which may prove to have some diagnostic use, in conjunction with the Goodenough Draw-A-Man Test, the Wide Range Achievement Test (WRAT), the Thematic Apperception Test, and school information, in evaluating various categories of subnormal and deviant performance such as those enumerated above.

On the Vocabulary subtest, Stacey and Portnoy (168) also observed qualitative differences between a borderline group (IQ range 66-79) and a defective group (IQ range 50-65) in conceptual approaches to word definition. Defectives exceeded borderlines significantly in the use of functional definitions, while the borderlines were significantly higher in use of descriptive definitions. Neither group used abstract concepts to more than a slight degree.

Carleton and Stacey (219) made an item analysis of the Vocabulary and Block Design subtests with a sample of 366 low-IQ children (mean FS IQ 67) and found four Voc. items and two BD items displaced. In view of the greater dependence on these two subtests in a short form than is usually required with the full test, consideration might well be given by the Survey staff to a repetition of this study for a substantial sample.

Maxwell (211) observed that the WISC variances for a sample of neurotic children were greater than for a normal sample, which led him to criticize the transformations of raw scores to scaled scores. This point was also made by Wilson (139), whose work was with Indian children. Walker (209), in a highly creative study, enumerated a lengthy list of qualitative variations of WISC responses that appear to have promise for personality diagnosis. Walker's study merits further followup.

Developmental Factors

Klausmeier and Check (166) investigated a number of developmental correlates of the WISC. They reported that children with high intelligence quotients grow *taller* than those in the average or low range, but that *weight* is not significantly related to sex or IQ. On *strength of grip*, they found low-IQ children weaker than those with average or high IQ's, the average group weaker than the high-IQ group, and girls weaker than boys. Girls were found to have more *permanent teeth* and a higher *carpal age* than boys of the same age. No sex differences or IQ differences were found in relation to *emotional adjustment.* Girls also exceeded boys on *achievement in relation to capacity, integration of self concept,* and *estimation of own ability.* These observations are of interest in suggesting cross-disciplinary analysis of psychological and biomedical data.

SPECIAL GROUPS

The following discussion includes research on the WISC with reference to a number of special groups—those involving various disabilities, afflictions, deviations, social and ethnic characteristics, and other definitive attributes commonly recognized in the literature—for which at least some information has been found. Each of these groups involves some variables which affect WISC scores, and this review might properly have been included in the preceding section. However, most of the research referred to here was organized in terms of samples of persons in various categories rather than by underlying variables. As a result, the organization of the discussion follows the organization of the material reviewed.

Reading Disability

As noted earlier, Kallos and others (180) found that Block Design scores were significantly higher than Vocabulary scores for a reading disability sample of 37 boys aged 9 to 14 years whose IQ's ranged from 90 to 109. The elevation of BD was supported by Burks and Bruce (186). Altus (181), Sheldon and Garton (182), and Karlsen (185) published WISC profiles for retarded readers, based on small but similar groups. No consistent pattern is unequivocally shown. Robeck (183) used a more sophisticated method to study subtest patterning of problem readers on the WISC, representing subtest scores as deviations of scaled

scores from the respective age-group means. By this method problem readers were significantly higher than the norms on both Block Design and Vocabulary (as well as on Comprehension, Similarities, and Picture Arrangement) and lower on Digit Span, Arithmetic, Information, and Coding. Rogge (187) reported no significant differences on WISC VS, PS, or FS IQ's between a sample of 132 delinquents 14 to 16 years of age and a control sample of good readers.

Correlations of WISC scales with reading tests are generally moderate, in the range of 0.3 to 0.5 (171, 172, and 173). On the other hand, approaches involving score patterns or profiles, such as discussed above, and qualitative analyses of responses, exemplified by the analyses of the understanding of the concept of *opposite*, by Robinowitz (108) and by Flamand (172), appear to offer greater promise than linear regression methods for the evaluation of reading disability cases. The latter approach does not appear feasible with only Voc. and BD in the battery, but the pattern approach, as discussed above, merits consideration. In the Survey battery the WRAT is, of course, most directly related to estimation of reading disability, but a Voc.-BD ratio may be a useful supplement.

Auditory Disability

Murphy (196) administered the WISC to an equally divided sample of 300 deaf boys and girls in English schools for the deaf. Deaf children did not differ significantly from normal children on the Performance Scale in this study, and there was no meaningful relation between hearing loss and PS. It is of interest, though, that Block Design correlated 0.71 with PS in this sample. In addition, teacher ratings of emotional adjustment correlated 0.76 with PS, suggesting that here also, as in the samples evaluated in relation to the Children's Manifest Anxiety Scale, anxiety may be a deterrent to effective performance.

Graham and Shapiro (195) compared the performance of the deaf and normal children on the WISC with standard and pantomime instructions. Both groups did equally well on PS with pantomime instructions, but the normals were superior with standard instructions. Mean scores on BD were approximately equal under all three conditions. For deaf children, then, the pantomime instructions are appropriate on BD.

Glowatsky (194) found that WISC Performance Scale IQ's were comparable with Draw-A-Man Test IQ's for a sample of 24 deaf and hard-of-hearing children in Santa Fe. PS scores were substantially higher than VS scores in this group, but bilingualism (noted in 13 cases) was not a factor.

Thompson gave Wepman's Auditory Discrimination Test, the WISC, and other tests of reading and auditory acuity to 105 children, including good and poor readers. She found that a significant and substantial proportion of first graders (71 percent) had inadequate auditory discrimination, but that this number was reduced to 24 percent by the second grade. Auditory Discrimination scores correlated more highly with reading (0.59 to 0.66) than with WISC IQ's (0.55 to 0.58). The correlation of Auditory Discrimination with WISC Verbal Scale IQ, the highest correlation reported, was 0.61.

Where hearing disability is noted by audiometer test it would be advantageous to estimate intelligence level by a combination of Draw-A-Man and Block Design scores.

Visually Handicapped

According to a study by Scholl (197), the Block Design test may be administered with normal procedures to the partially blind. For the totally blind only the Vocabulary test would be appropriate in the Survey, and no data are available to evaluate their scores adequately.

Stutterers

Post (198) found no significant differences between the mean scores of 30 stutterers and 30 controls, predominantly boys in the age range of 5-5 to 15-10, on the Stanford-Binet (S-B) and the WISC. The correlation of WISC Full Scale IQ with the S-B was 0.78 for the stutterers. The only difference found between the two groups was in the correlation of WISC Verbal Scale and Performance Scale IQ's, which was 0.26 for the stutterers and 0.60 (the same as in Wechsler's standardization sample) for the controls. Both group means were higher on PS than VS.

Cerebral Palsy

Bortner and Birch (199) studied the administration of the Block Design subtests with twenty-eight 13-year-old cerebral palsied children. They found, as may be expected, that the ability to discriminate block designs in a choice situation may be intact even though motor factors impair reproductive ability.

Organic Impairment of Central Nervous System

Beck and Lam (200) found that WISC Full Scale IQ's of diagnosed organics were lower than those of nonorganics, but failed, as others have, to verify Wechsler's subtest diagnostic pattern for organics. Young and Pitts (202) compared the WISC scores of 40 rural juvenile congenital syphilitics (aged 6 to 16 years) with 40 normal controls matched on age, sex, race, region, and father's occupation. The controls were significantly superior on IQ's and on Vocabulary, but not on Block Design, where the critical ratio was marginal.

Gifted

In Edmonton, Chalmers (213) administered the WISC to 57 superior children with IQ's above 120 (mean FS IQ 128) and found that 11 obtained perfect scores on one or more tests. However, there were no perfect scores on Vocabulary and only one on Block Design. Nevertheless, Chalmers questioned the adequacy of the WISC ceilings for precise measurement in the very high range. Trauba (214), with a similar sample of 71 gifted Kansas children, found that WISC Vocabulary has a correlation of 0.71 with the McCall-Crabbs Standard Test Lesson in Reading. Lucito and Gallagher (215) obtained a mean WISC Full Scale IQ of 141 for a sample of 50 children whose mean S-B IQ was 161. In this group the boys' scores were slightly higher than those of the girls. In agreement with Gallagher and Lucito (164), mentioned earlier, Similarities, Information, and Vocabulary were the three highest tests for boys and girls. Object Assembly, Coding, and Picture Arrangement were lowest for boys, while Digit Span, Picture Arrangement, and Picture Completion were lowest for girls (only partially in agreement with Gallagher and Lucito).

The adequacy of the WISC for precise measurement of the gifted may be questioned, but it is possible that more accurate measurement may be obtained by use of the present short form of Vocabulary and Block Design than with the Full Scale. This is a problem, however, that will require further attention.

Mentally Retarded and Defective

The research on the use of the WISC with retarded and defective groups is very favorable, in contrast with research on its use for the gifted. This is indicated by virtually all the studies reviewed: (a) reliabilities reported—Throne and others (227) obtained retest reliabilities over 3 to 4 months of 0.79 for Vocabulary and 0.82 for Block Design on a sample of 39 retarded boys aged 11 to 14 years; (b) correlations of the WISC with other tests—Stanford-Binet (216, 217, 228, and 229), Leiter International Performance Scale (221 and 229), Wechsler Adult Intelligence Scale (222), Columbia Mental Maturity Scale (224), Goodenough Draw-A-Man Test (224), Progressive Matrices (233), Peabody Picture Vocabulary Test (234), and grade placement (238); (c) patterning studies, mentioned earlier; (d) absence of sex differences (232); and (e) amenability to short forms based on Vocabulary and Block Design, as discussed above. (See Research on Short Forms of the WISC.) Differences between WISC and Stanford-Binet IQ's are smaller in this range than in any other. It appears that estimates of retardation in the population should be justified on the basis of a composite score of Voc. and BD, but the desirability of further research to develop a conversion table to the Full Scale should not be minimized.

Bilingual

The effect of bilingualism appears to be in the direction of lowering the Vocabulary scores; no effects have been reported on Block Design. Altus (240) reported such results for Mexicans in California; Kralovich (241), for children of Slavic origin in New Jersey; and Levinson (243 and 244), for Jewish children in New York. Kralovich reported a correlation of 0.61 between the Verbal and Performance scales of the WISC for 28 monolinguals and -0.04 for 28 bilinguals. Where bi-

lingualism is known to exist, verbal tests may be expected to be invalid measures and greater reliance on performance-type tests such as Block Design and Draw-A-Man is indicated.

Negro

The WISC norms do not apply to Negro children, and research by Young and Bright (251), Caldwell (252), Blakemore (253), and Racheile (254), as well as others, does nothing to alter this fact. Negroes score lower than whites, and it is generally accepted that cultural experience and caste factors not only account for the Negro-white differences, but also render comparable measurement by culture-fair or culture-free methods as difficult as other ethnic comparisons. The sampling designs of the studies cited, which used the WISC, were not adequate to qualify them for any detailed comment on differences found.

Socioeconomic Status

Laird (250) compared children of different socioeconomic status (SES) on the WISC and noted, in common with the general trend in the literature, superior performance at upper levels. Estes (247 and 248) found similar differences at grade 2 but not at grade 5. At both grades the WISC Full Scale IQ was more highly correlated with the Metropolitan Achievement Test for the higher SES sample.

COMPARISON OF WISC AND STANFORD-BINET IQ'S

Despite the theoretical objections to the mental age concept, discussed earlier, which led to the adoption of the deviation IQ as a distinctive feature of the Wechsler scales and which set them apart from the venerable Stanford-Binet test, the relation of the WISC to the S-B has been a matter of great interest, as evidenced by the number of papers on this topic in the present review.

The Stanford-Binet is indeed one of the giants among psychological tests, a veritable landmark in the history of psychological measurement, and still enjoys extensive school and clinical use, notwithstanding the fact that its popularity has been somewhat reduced by the success of the relatively recent WISC. Although the standardization of the WISC has been impressive and supported by sophisticated conceptualization, many users have been relieved to find that it is highly correlated with the Stanford-Binet. The correlation is in fact so high (accounting for over 80 percent of common variance) that one wonders about the significance of the theorizing which describes them so differently.

The impression of similarity of measurement results given by the correlations does not, however, stand up when mean scores of different groups are compared. As noted earlier, WISC IQ's tend to be lower than Stanford-Binet IQ's at the lower age levels and among the gifted. These observations are illustrated by data extracted from the following 12 studies in which comparison means were cited: 119, 120, 124, 147, 148, 151, 153, 156, 159, 161, 215, and 216. Their results are epitomized briefly on the following page. Data from Jones' (154) British study of 240 children in the age range 8 to 10 years are also of interest. For this group the WISC means were, on the average, 7.2 IQ points below the S-B, the WISC always being administered first.

Allowing for sampling fluctuations and errors of measurement in routine testing, there never-

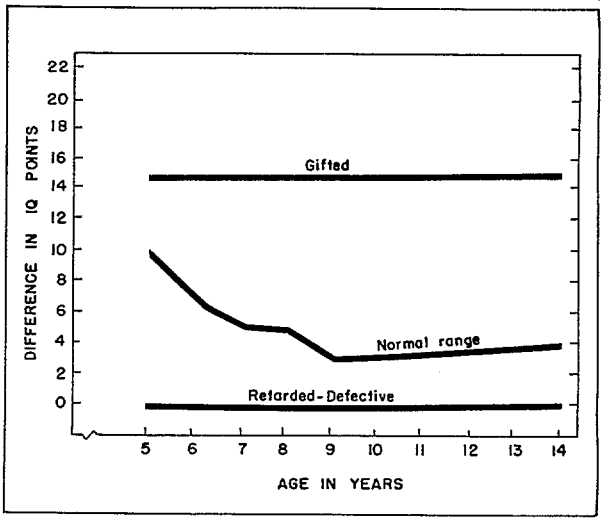

Figure 1. Summary of the amount Stanford-Binet Intelligence Test scores differ from Wechsler Intelligence Test scores.

Normal (White) Samples			
Schachter and Apgar (147)[1]	Mean age 4-1 Mean age 8-3 N 113 (61m, 62f)	Mean S-B Mean WISC	104.3 98.9 -5.4
Triggs and Cartee (148)	Kindergarten-Age 5 N 48	Mean S-B Mean WISC	124.1 107.6 -16.5
Muhr (119)	5-year group N 21	Mean S-B Mean WISC	97.4 88.1 -9.3
	6-year group N 21	Mean S-B Mean WISC	102.2 96.6 -5.6
Pastovic and Guthrie (120)	5-year group N 50	Mean S-B Mean WISC	113.0 103.2 -9.8
	7-year group N 50	Mean S-B Mean WISC	115.1 111.5 -3.6
Rottersman (151)	6-year group N 50	Mean S-B Mean WISC	110.2 101.5 -8.7
Cohen and Collier (124)	6- to 9-year group Ages 6-5 to 8-9 N 53	Mean S-B Mean WISC	104.8 99.8 -5.0
Wagner (156)	8- to 9-year group N 50	Mean S-B Mean WISC	104.5 103.3 -1.2
Frandsen and Higginson (159)	9-year group N 50	Mean S-B Mean WISC	105.8 102.4 -3.4
Kardos (161)	13- to 14-year group N 100	Mean S-B Mean WISC	113.7 109.4 -4.3
Gifted (White) Samples			
Beeman (153)	N 36 Full sample: Mean WISC compared with Mean S-B: IQ over 130: Mean WISC compared with Mean S-B: IQ 120-129: Mean WISC compared with Mean S-B:		-15 -20 -11.4
Lucito and Gallagher (215)	N 50	Mean S-B Mean WISC	160.8 141.2 -19.6
Retarded Samples			
Nale (216)	9- to 11-year group N 104	Mean S-B Mean WISC	55.4 58.0 +2.6

[1]Interval between S-B and WISC administration, 50 months.

NOTE: N—number; m—male; f—female.

theless appears to be a common trend in these reports which can be summarized as follows. The differences between WISC and S-B IQ's are greatest among the gifted. In the normal range they are high among the very young, dropping off as age increases, but persisting to some degree throughout the age range 5 to 14 years. The data suggest an upturn after age 9, but this is not certain. No significant differences appear for the subnormal. The schematic chart in figure 1 suggests the nature of the age- and level-related difference functions on the basis of the results cited.

Unfortunately it is possible only to speculate on the nature of the true curves which those in figure 1 are intended to suggest, and speculation on what they would be for a short form composed only of Vocabulary and Block Design is difficult. Some of the data presented earlier for these subtests suggest that the differences might be smaller, but in the absence of empirical evidence this is only an educated guess.

For the purposes of the Survey there are only two alternatives. One is to carry out some *ad hoc* research on the short form, as suggested earlier, for the purpose of estimating the Full Scale IQ from Voc. and BD, using the results to conform to Wechsler's norms. The other is to regard the full Survey sample as the unprecedented opportunity to carry out a complete new standardization of the short form on a basis that, in sampling sophistication, far exceeds any work of its kind in the history of testing. There are a number of problems related to the second alternative, including the availability of funds for this purpose. However, if this standardization were accomplished, the new norms for Voc. and BD would be superior to those now available, and the computations of FS IQ based on them would permit more accurate population estimates than any others conceivable for the age range included.

SUMMARY AND CONCLUSIONS

This review is based on 154 published studies, reviews, and unpublished theses and dissertations related to the WISC, interpreted in a frame of reference of measurement theory and psychometric principles. The evidence considered strongly supports the judgment of the Survey staff in the selection of the WISC Vocabulary and Block Design subtests as a short form of the WISC for the national survey, but at the same time it raises questions concerning the acceptance of either the scaled scores of these subtests or of prorated Full Scale Intelligence Quotients based on them without further empirical research. It is the reviewer's considered opinion that, given the alternatives presented, the selection was an eminently wise one. The research recommended reflects principally the nature of the unprecedented testing problems and the generally imprecise nature of psychological measurement.

The most important recommended investigations discussed in this section involve the following steps:

1. Restandardization of the Vocabulary and Block Design tests on the full Survey sample. As part of this study, item difficulties should be checked and a formula or set of formulas should be developed for estimating Full Scale IQ's from revised Voc. and BD scaled scores (based on samples of normal, gifted, and retarded groups—and if possible several ethnic groups, such as Negroes or Mexicans—to whom the Full Scale has been administered). Consideration should be given to estimation of IQ's directly from raw scores by age group.
2. Research on correlates of a Voc.-BD ratio, for use with the WRAT and with the Draw-A-Man Test in the identification of poor readers, bilinguals, and verbally impaired children and in estimating IQ's of culturally deviant ethnic groups.
3. Cross-disciplinary developmental analyses of Vocabulary, Block Design, and derived scores and of item responses with biomedical data obtained in other sections of the Survey. This area is discussed in detail elsewhere. See Klausmeier and Check (166).

BIBLIOGRAPHY

General References to WISC

101. Wechsler, D.: *Wechsler Intelligence Scale for Children.* New York. Psychological Corp., 1949.

102. Littell, W. M.: The Wechsler Intelligence Scale for Children, review of a decade of research. *Psychological Bull.* 57:132-156, 1960.

103. McCandless, B. R.: Review of the WISC, in O. K. Buros, ed., *Fourth Mental Measurements Yearbook.* Highland Park, N.J. The Gryphon Press, 1953. pp. 480-481.

104. Frost, B. P.: An application of the method of extreme deviations to the Wechsler Intelligence Scale for Children. *J.Clin.Psychol.* 16:420, 1960.

105. Winpenny, N.: *An Investigation of the Use and the Validity of Mental Age Scores on the Wechsler Intelligence Scale for Children.* Unpublished master's thesis, Pennsylvania State College, 1951.

106. Maxwell, A. E.: Inadequate reporting of normative test data. *J.Clin.Psychol.* 17:99-101, 1961.

107. Seashore, H. G.: Differences between verbal and performance IQ's on the Wechsler Intelligence Scale for Children. *J.Consult.Psychol.* 15:62-67, 1951.

108. Robinowitz, R.: Learning the relation of opposition as related to scores on the Wechsler Intelligence Scale for Children. *J.Genet.Psychol.* 88:25-30, 1956.

Factor Analytic Studies

109. Hagen, E. P.: *A Factor Analysis of the Wechsler Intelligence Scale for Children.* Unpublished doctoral dissertation, Columbia University, 1952.

110. Gault, U.: Factorial patterns on the Wechsler Intelligence Scales. *Aust.J.Psychol.* 6:85-90, 1954.

111. Cohen, J.: The factorial structure of the WISC at ages 7-6, 10-6, and 13-6. *J.Consult.Psychol.* 23:285-299, 1959.

Reliability and Stability

112. Hite, L.: *Analysis of Reliability and Validity of the Wechsler Intelligence Scale for Children.* Unpublished doctoral dissertation, Western Reserve University, 1953.

113. Gehman, I. H., and Matyas, R. P.: Stability of the WISC and Binet tests. *J.Consult.Psychol.* 20:150-152, 1956.

114. Matyas, R. P.: *A Longitudinal Study of the Revised Stanford-Binet and the WISC.* Unpublished master's thesis, Pennsylvania State University, 1954.

115. Reger, R.: Repeated measurements with the WISC. *Psychol.Rep.* 11:418, 1962.

116. Whatley, R. G., and Plant, W. T.: The stability of WISC IQ's for selected children. *J.Psychol.* 44:165-167, 1957.

Validity

117. Mussen, P., Dean, S., and Rosenberg, M.: Some further evidence on the validity of the WISC. *J.Consult.Psychol.* 16:410-411, 1952.

118. Kureth, G., Muhr, J. P., and Weisgerber, C. A.: Some data on the validity of the Wechsler Intelligence Scale for Children. *Child Development* 23:281-287, 1952.

119. Muhr, J. P.: *Validity of the Wechsler Intelligence Scale for Children at the Five and Six Year Level.* Unpublished master's thesis, University of Detroit, 1952.

120. Pastovic, J. J., and Guthrie, G. M.: Some evidence on the validity of the WISC. *J.Consult.Psychol.* 15:385-386, 1951.

121. Pastovic, J. J.: *A Validation Study of the Wechsler Intelligence Scale for Children at the Lower Age Level.* Unpublished master's thesis, Pennsylvania State College, 1951.

122. Altus, G. T.: A note on the validity of the Wechsler Intelligence Scale for Children. *J.Consult.Psychol.* 16: 231, 1952.

Relations with Other Tests: Batteries

123. McBrearty, J. F.: *Comparison of the WISC With the Arthur Performance Scale, Form I, and Their Relationship to the Progressive Achievement Test.* Unpublished master's thesis, Pennsylvania State College, 1951.

124. Cohen, B. D., and Collier, M. J.: A note on WISC and other tests of children six to eight years old. *J.Consult. Psychol.* 16:226-227, 1952.

125. Estes, B. W., Curtin, M. E., DeBurger, R. A., and Denny, C.: Relationships between 1960 Stanford-Binet, 1937 Stanford-Binet, WISC, Raven, and Draw-A-Man. *J.Consult.Psychol.* 25:388-391, 1961.

126. Smith, B. S.: The relative merits of certain verbal and non-verbal tests at the second-grade level. *J.Clin.Psychol.* 17:53-54, 1961.

Relations with Other Tests: Wechsler-Bellevue

127. Bacon, C. S.: *A Comparative Study of the Wechsler-Bellevue Intelligence Scale for Adolescents and Adults, Form I, and the Wechsler Intelligence Scale for Children at the Twelve-Year Level.* Unpublished master's thesis, University of North Dakota, 1954.

128. Delattre, L., and Cole, D.: A comparison of the WISC and the Wechsler-Bellevue. *J.Consult.Psychol.* 16:228-230, 1952.

Relations with Other Tests: Bender-Gestalt Perceptual Tests

129. Koppitz, E. M.: Relationships between the Bender-Gestalt Test and the Wechsler Intelligence Scale for Children. *J.Clin.Psychol.* 14:413-416, 1958.

130. Armstrong, R. G., and Hauck, P. A.: Correlates of the Bender-Gestalt scores in children. *J.Psychol.Stud.* 11: 153-158, 1960.

131. Goodenough, D. R., and Karp, S. A.: Field dependence and intellectual functioning. *J.Abnorm.&Social Psychol.* 63:241-246, 1961.

Relations with Other Tests: CMAS

132. Hafner, A. J., Pollie, D. M., and Wapner, I.: The relationship between the CMAS and WISC functioning. *J.Clin. Psychol.* 16:322-323, 1960.

Relations with Other Tests:
Ammons Full Range Picture Vocabulary

133. Smith, L. M., and Fillmore, A. R.: The Ammons FRPV Test and the WISC for remedial reading cases; abstracted, *J.Consult.Psychol.* 18:332, 1954.

Relations with Other Tests: CTMM

134. Altus, G. T.: Relationships between verbal and non-verbal parts of the CTMM and WISC. *J.Consult.Psychol.* 19:143-144, 1955.

Relations with Other Tests: Kent EGY

135. Delp, H. A.: Correlations between the Kent EGY and the Wechsler batteries. *J.Clin.Psychol.* 9:73-75, 1953.

Relations with Other Tests: Peabody Picture Vocabulary Test

136. Kimbrell, D. L.: Comparison of Peabody, WISC, and academic achievement scores among educable mental defectives. *Psychol.Rep.* 7:502, 1960.

137. Himelstein, P., and Herndon, J. D.: Comparison of the WISC and Peabody Picture Vocabulary Test with emotionally disturbed children. *J.Clin.Psychol.* 18:82, 1962.

Relations with Other Tests: Raven Progressive Matrices

138. Barratt, E. S.: The relationship of the Progressive Matrices (1938) and the Columbia Mental Maturity Scale to the WISC. *J.Consult.Psychol.* 20:294-296, 1956.

139. Wilson, L.: *A Comparison of the Raven Progressive Matrices (1947) and the Performance Scale of the Wechsler Intelligence Scale for Children for Assessing the Intelligence of Indian Children.* Unpublished master's thesis, University of British Columbia, 1952.

140. Malpass, L. F., Brown, R., and Hade, D.: The utility of the Progressive Matrices (1956 edition) with normal and retarded children. *J.Clin.Psychol.* 16:350, 1960.

141. Stacey, C. L., and Carleton, F. O.: The relationship between Raven's Colored Progressive Matrices and two tests of general intelligence. *J.Clin.Psychol.* 11:84-85, 1955.

142. Martin, A. W., and Wiechers, J. E.: Raven's Colored Progressive Matrices and the Wechsler Intelligence Scale for Children. *J.Consult.Psychol.* 18:143-144, 1954.

Relations with Other Tests: SRA-PMA

143. Stempel, E. F.: The WISC and the SRA Primary Mental Abilities Test. *Child Development* 24:257-261, 1953.

Relations with Other Tests: Stanford-Binet

144. Krugman, J. I., Justman, J., Wrightstone, J. W., and Krugman, M.: Pupil functioning on the Stanford-Binet and the Wechsler Intelligence Scale for Children. *J.Consult. Psychol.* 15:475-483, 1951.

145. Harlow, J. E., Jr., Price, A. C., Tatham, L. J., and Davidson, J. F.: Preliminary study of comparison between Wechsler Intelligence Scale for Children and Form L of the Revised Stanford Binet Scale at three age levels. *J.Clin.Psychol.* 13:72-73, 1957.

146. Boruszak, R. J.: *A Comparative Study to Determine the Correlation Between the IQ's of the Revised Stanford Binet Scale, Form L, and the IQ's of the Wechsler Intelligence Scale for Children.* Unpublished master's thesis, Wisconsin State College, 1954.

147. Schachter, F. F., and Apgar, V.: Comparison of preschool Stanford-Binet and school-age WISC IQ's. *J.Educ. Psychol.* 49:320-323, 1958.

148. Triggs, F. O., and Cartee, J. K.: Pre-school pupil performance on the Stanford-Binet and the Wechsler Intelligence Scale for Children. *J.Clin.Psychol.* 9:27-29, 1953.

149. Holland, G. A.: A comparison of the WISC and Stanford-Binet IQ's of normal children. *J.Consult.Psychol.* 17: 147-152, 1953.

150. Weider, A., Noller, P. A., and Schraumm, T. A.: The Wechsler Intelligence Scale for Children and the Revised Stanford-Binet. *J.Consult.Psychol.* 15:330-333, 1951.

151. Rottersman, L.: *A Comparison of the IQ Scores on the New Revised Stanford Binet, Form L, the Wechsler Intelligence Scale for Children, and the Goodenough "Draw A Man" Test at the Six Year Age Level.* Unpublished master's thesis, University of Nebraska, 1950.

152. Tatham, L. J.: *Statistical Comparison of the Revised Stanford-Binet Intelligence Test Form L With the Wechsler Intelligence Scale for Children Using the Six and One-Half Year Level.* Unpublished master's thesis, University of Florida, 1952.

153. Beeman, G.: A comparative study of the WISC and Stanford-Binet with a group of more able and gifted 7-11 year old students. *Calif.J.Educ.Res.* 11:77, 1960.

154. Jones, S.: The Wechsler Intelligence Scale for Children applied to a sample of London primary school children. *Br.J.Educ.Psychol.* 32(2):119-133, 1962.

155. Scott, G. R.: *A Comparison Between the Wechsler Intelligence Scale for Children and the Revised Stanford-Binet Scales.* Unpublished master's thesis, Southern Methodist University, 1950.

156. Wagner, W. K.: *A Comparison of Stanford-Binet Mental Ages and Scaled Scores on the Wechsler Intelligence Scale for Children for Fifty Bowling Green Pupils.* Unpublished master's thesis, Bowling Green State University, 1951.

157. Stanley, J. C.: Statistical analysis of scores from counterbalanced tests. *J.Exp.Educ.* 23:187-207, 1955.

158. Arnold, F. C., and Wagner, W. K.: A comparison of Wechsler Children's Scale and Stanford-Binet scores for eight- and nine-year-olds. *J.Exp.Educ.* 24:91-94, 1955.

159. Frandsen, A. N., and Higginson, J. B.: The Stanford-Binet and the Wechsler Intelligence Scale for Children. *J.Consult.Psychol.* 15:236-238, 1951.

160. Clarke, F. R.: *A Comparative Study of the Wechsler Intelligence Scale for Children and the Revised Stanford Binet Intelligence Scale, Form L, in Relation to the Scholastic Achievement of a 5th Grade Population.* Unpublished master's thesis, Pennsylvania State College, 1950.

161. Kardos, M. S.: *A Comparative Study of the Performance of Twelve-Year-Old Children on the WISC and the Revised Stanford-Binet, Form L, and the Relationship of Both to the California Achievement Tests.* Unpublished master's thesis, Marywood College, 1954.

162. Davidson, J. F.: *A Preliminary Study in Statistical Comparison of the Revised Stanford-Binet Intelligence Test Form L With the Wechsler Intelligence Scale for Children Using the Fourteen Year Level.* Unpublished master's thesis, University of Florida, 1954.

Relations with Other Tests: Wartegg Drawing Completion Test

163. Stark, R.: *A Comparison of Intelligence Test Scores on the Wechsler Intelligence Scale for Children and the Wartegg Drawing Completion Test with School Achievement of Elementary School Children.* Unpublished master's thesis, University of Detroit, 1954.

WISC: Response Patterns of Gifted, Average, and Retarded

164. Gallagher, J. J., and Lucito, L. L.: Intellectual patterns of gifted compared with average, and retarded. *Except. Children* 27:479-482, 1961.

165. Klausmeier, H. J., and Feldhusen, J. F.: Retention in arithmetic among children of low, average, and high intelligence at 117 months of age. *J.Educ.Psychol.* 50: 88-92, 1959.

166. Klausmeier, H. J., and Check, J.: Relationships among physical, mental, achievement, and personality measures in children of low, average, and high intelligence at 113 months of age. *Am.J.Ment.Deficiency* 63:1059-1068, 1959.

167. Feldhusen, J. F., and Klausmeier, H. J.: Anxiety, intelligence, and achievement in children of low, average, and high intelligence. *Child Development* 33:403-409, 1962.

WISC: Vocabulary, Language Skills, Reading

168. Stacey, C. L., and Portnoy, B.: A study of the differential responses on the vocabulary subtest of the Wechsler Intelligence Scale for Children. *J.Clin.Psychol.* 6:401-403, 1950.

169. Winitz, H.: *A Comparative Study of Certain Language Skills in Male and Female Kinaeryarten Children.* Unpublished doctoral dissertation, State University of Iowa, 1959.

170. Dunsdon, M. I., and Roberts, J. A. F.: A study of the performance of 2,000 children on four vocabulary tests. *Br.J.Statist.Psychol.* 8:3-15, 1955.

171. Reidy, M. E.: *A Validity Study of the Wechsler-Bellevue Intelligence Scale for Children and Its Relationship to Reading and Arithmetic.* Unpublished master's thesis, Catholic University of America, 1952.

172. Flamand, R. K.: *The Relationship Between Various Measures of Vocabulary and Performance in Beginning Reading.* Unpublished doctoral dissertation, Temple University, 1961.

173. Triggs, F. O., Cartee, J. K., Binks, V., Foster, D., and Adams, N. A.: The relationship between specific reading skills and general ability at the elementary and junior-senior high school levels. *Educ.Psychol.Measur.* 14: 176-185, 1954.

174. Fitzgerald, L. A.: *Some Effects of Reading Ability on Group Intelligence Test Scores in the Intermediate Grades.* Unpublished doctoral dissertation, State University of Iowa, 1960; abstracted, *Diss.Abstr.* 21:1844, 1961.

WISC: Short Forms

175. Armstrong, R. G.: A reliability study of a short form of the WISC vocabulary subtest. *J.Clin.Psychol.* 11:413-414, 1955.

176. Throne, J. M.: *A Short Form of the Wechsler-Bellevue Intelligence Test for Children.* Unpublished master's thesis, University of Florida, 1951.

177. Simpson, W. H., and Bridges, C. C., Jr.: A short form of the Wechsler Intelligence Scale for Children. *J.Clin.Psychol.* 15:424, 1959.

178. Carleton, F. O., and Stacey, C. L. : Evaluation of selected short forms of the Wechsler Intelligence Scale for Children. *J.Clin.Psychol.* 10:258-261, 1954.

179. Yalowitz, J. M., and Armstrong, R. G.: Validity of short forms of the Wechsler Intelligence Scale for Children (WISC). *J.Clin.Psychol.* 11:275-277, 1955.

WISC: Reading Disability

180. Kallos, G. L., Grabow, J. M., and Guarino, E. A.: The WISC profile of disabled readers. *Personnel Guid.J.* 39:476-478, 1961.

181. Altus, G. T.: A WISC profile for retarded readers. *J.Consult.Psychol.* 20:155-156, 1956.

182. Sheldon, M. S., and Garton, J.: A note on "a WISC profile for retarded readers." *Alberta J.Educ.Res.* 5:264-267, 1959.

183. Robeck, M. C.: Subtest patterning of problem readers on WISC. *Calif.J.Educ.Res.* 11:110-115, 1960.

184. Abrams, J. C.: *A Study of Certain Personality Characteristics of Non-Readers and Achieving Readers.* Unpublished doctoral dissertation, Temple University, 1955.

185. Karlsen, B.: *A Comparison of Some Educational and Psychological Characteristics of Successful and Unsuccessful Readers at the Elementary School Level.* Unpublished doctoral dissertation, University of Minnesota, 1954.

186. Burks, H. F., and Bruce, P.: The characteristics of poor and good readers as disclosed by the Wechsler Intelligence Scale for Children. *J.Educ.Psychol.* 46:488-493, 1955.

187. Rogge, H. J.: *A Study of the Relationships of Reading Achievement to Certain Other Factors in a Population of Delinquent Boys.* Unpublished doctoral dissertation, University of Minnesota, 1959.

WISC: School Achievement

188. Orr, K. N.: *The Wechsler Intelligence Scale for Children as a Predictor of School Success.* Unpublished master's thesis, Indiana State Teachers College, 1950.

189. Schwitzgoebel, R. R.: *The Predictive Value of Some Relationships Between the Wechsler Intelligence Scale for Children and Academic Achievement in Fifth Grade.* Unpublished doctoral dissertation, University of Wisconsin, 1952.

190. Barratt, E. S., and Baumgarten, D. L.: The relationship of the WISC and Stanford-Binet to school achievement. *J.Consult.Psychol.* 21:144, 1957.

191. Raleigh, W. H.: *A Study of the Relationships of Academic Achievement in Sixth Grade With the Wechsler Intelligence Scale for Children and Other Variables.* Unpublished doctoral dissertation, Indiana University, 1952.

192. Stroud, J. B., Blommers, P., and Lauber, M.: Correlation of WISC and achievement tests. *J.Educ.Psychol.* 48: 18-26, 1957.

WISC: Auditory Disability, Visual Handicap, Stuttering, Cerebral Palsy, Brain Damage

193. Thompson, B. B.: *The Relation of Auditory Discrimination and Intelligence Test Scores to Success in Primary Reading.* Unpublished doctoral dissertation, Indiana University, 1961.

194. Glowatsky, E.: The verbal element in the intelligence scores of congenitally deaf and hard of hearing children. *Amer.Ann.Deaf* 98:328-335, 1953.

195. Graham, E. E., and Shapiro, E.: Use of the Performance Scale of the Wechsler Intelligence Scale for Children with the deaf child. *J.Consult.Psychol.* 17:396-398, 1953.

196. Murphy, L. J.: Tests of abilities and attainments, pupils in schools for the deaf aged six to ten, in A. W. G. Ewing, ed., *Educational Guidance and the Deaf Child.* Manchester, England. Manchester University Press, 1957. pp. 213-251.

197. Scholl, G.: Intelligence tests for visually handicapped children. *Excep.Children* 20:116-120, 1953..

198. Post, D. P.: *A Comparative Study of the Revised Stanford Binet and the Wechsler Intelligence Scale for Children Administered to a Group of Thirty Stutterers.* Unpublished master's thesis, University of Southern California, 1952.

199. Bortner, M., and Birch, H. G.: Perceptual and perceptual-motor dissociation in cerebral palsied children. *J.Nerv. &Ment.Dis.* 134:103-108, 1962.

200. Beck, H. S., and Lam, R. L.: Use of the WISC in predicting organicity. *J.Clin.Psychol.* 11:154-157, 1955.

201. Kilman, B. A., and Fisher, G. M.: An evaluation of the Finley-Thompson abbreviated form of the WISC for undifferentiated, brain damaged and functional retardates. *Am.J.Ment.Deficiency* 64:742-746, 1960.

202. Young, F. M., and Pitts, V. A.: The performance of congenital syphilitics on the Wechsler Intelligence Scale for Children. *J.Consult.Psychol.* 15:239-242, 1951.

203. Rowley, V. N.: Analysis of the WISC performance of brain damaged and emotionally disturbed children. *J.Consult.Psychol.* 25:553, 1961.

WISC: Personality Measures (Normal), Discipline, Delinquency

204. Gourevitch, V., and Feffer, M. H.: A study of motivational development. *J.Genet.Psychol.* 100:361-375, 1962.

205. Carrier, N. A., Orton, K. D., and Malpass, L. F.: Responses of bright, normal, and EMH children to an orally-administered manifest anxiety scale. *J.Educ.Psychol.* 53:271-274, 1962.

206. Burns, L.: *A Correlation of Scores on the Wechsler Intelligence Scale for Children and the California Test of Personality Obtained by a Group of 5th Graders.* Unpublished master's thesis, Pennsylvania State College, 1954.

207. Kent, N., and Davis, D. R.: Discipline in the home and intellectual development. *Brit.J.M.Psychol.* 30:27-33, 1957.

208. Wall, H. R.: *A Differential Analysis of Some Intellective and Affective Characteristics of Peer Accepted and Rejected Pre-Adolescent Children.* Unpublished doctoral dissertation, University of Kansas, 1960.

209. Walker, H. A.: *The Wechsler Intelligence Scale for Children as a Diagnostic Device.* Unpublished master's thesis, Utah State Agricultural College, 1956.

210. Schonborn, R.: A comparative study of the differences between adolescent and child male enuretics and non-enuretics as shown by an intelligence test. *Psychol. Newsletter* 6:1-9, 1954.

211. Maxwell, A. E.: Discrepancies in the variances of test results for normal and neurotic children. *Br.J.Statist. Psychol.* 13:165-172, 1960.

212. Richardson, H. M., and Surko, E. F.: WISC scores and status in reading and arithmetic of delinquent children. *J.Genet.Psychol.* 89:251-262, 1956.

WISC: Gifted

213. Chalmers, J. M.: *An Analysis of Results Obtained on the Wechsler Intelligence Scale for Children by Mentally Superior Subjects.* Unpublished master's thesis, University of Alberta, 1953.

214. Traube R. G.: *A Study of the Aspects of Differentiation of Abilities in Interpretation of Reading With a Group of Gifted Children.* Unpublished doctoral dissertation, University of Kansas, 1959.

215. Lucito, L., and Gallagher, J.: Intellectual patterns of highly gifted children on the WISC. *Peabody J.Educ.* 38:131-136, 1960.

WISC: Mental Defectives

216. Nale, S.: The Childrens-Wechsler and the Binet on 104 mental defectives at the Polk State School. *Am.J.Ment. Deficiency* 56:419-423, 1951.

217. Sloan, W., and Schneider, B.: A study of the Wechsler Intelligence Scale for Children with mental defectives. *Am.J.Ment.Deficiency* 55:573-575, 1951.

218. Atchison, C. O.: Use of the Wechsler-Intelligence Scale for Children with eighty mentally defective Negro children. *Am.J.Ment.Deficiency* 60:378-379, 1955.

219. Carleton, F. O., and Stacey, C. L.: An item analysis of the Wechsler Intelligence Scale for Children. *J.Clin. Psychol.* 11:149-154, 1955.

220. Newman, J. R., and Loos, F. M.: Differences between verbal and performance IQ's with mentally defective children on the Wechsler Intelligence Scale for Children. *J.Consult.Psychol.* 19:16, 1955.

221. Alper, A. E.: A comparison of the WISC and the Arthur adaptation of the Leiter International Performance Scale with mental defectives. *Am.J.Ment.Deficiency* 63:312-316, 1958.

222. Fleming, J. W.: *The Relationships Among Psychometric, Experimental, and Observational Measures of Learning Ability in Institutionalized Endogenous Mentally Retarded Persons.* Unpublished doctoral dissertation, University of Colorado, 1959.

223. Baroff, G. S.: WISC patterning in endogenous mental deficiency. *Am.J.Ment.Deficiency* 64:482-485, 1959.

224. Warren, S. A., and Collier, H. L.: Suitability of the Columbia Mental Maturity Scale for mentally retarded institutionalized females. *Am.J.Ment.Deficiency* 64:916-920, 1960.

225. Fisher, G. M.: A cross-validation of Baroff's WISC patterning in endogenous mental deficiency. *Am.J.Ment.Deficiency* 65:349-350, 1960.

226. Baumeister, A., and Bartlett, C. J.: Further factorial investigations of WISC performance of mental defectives. *Am.J.Ment.Deficiency* 67:257-261, 1962.

227. Throne, F. M., Schulman, J. L., and Kasper, J. C.: Reliability and stability of the Wechsler Intelligence Scale for Children for a group of mentally retarded boys. *Am. J.Ment.Deficiency* 67:455-457, 1962.

WISC: Mentally Retarded

228. Stacey, C. L., and Levin, J.: Correlation analysis of scores of subnormal subjects on the Stanford-Binet and Wechsler Intelligence Scale for Children. *Am.J.Ment. Deficiency* 55:590-597, 1951.

229. Sharp, H. C.: A comparison of slow learner's scores on three individual intelligence scales. *J.Clin.Psychol.* 13:372-374, 1957.

230. Matthews, C. G.: *Differential Performances of Non-Achieving Children on the Wechsler Intelligence Scale.* Unpublished doctoral dissertation, Purdue University, 1958.

231. Finley, C. J., and Thompson, J.: An abbreviated Wechsler Intelligence Scale for Children for use with educable mentally retarded. *Am.J.Ment.Deficiency* 63:473-480, 1958.

232. Finley, C., and Thompson, J.: Sex differences in intelligence of educable mentally retarded children. *Calif. J.Educ.Res.* 10:167-170, 1959.

233. Brown, R., Hakes, D., and Malpass, L.: The utility of the Progressive Matrices Test(1956 revision); abstracted, *Am.Psychologist* 14:341, 1959.

234. Dunn, L. M., and Brooks, S. T.: Peabody Picture Vocabulary Test performance of educable mentally retarded children. *Train.Sch.Bull.* 57:35-40, 1960.

235. Schwartz, L., and Levitt, E.: Short forms of the Wechsler Intelligence Scale for Children in the educable, non-institutionalized mentally retarded. *J.Educ.Psychol.* 51: 187-190, 1960.

236. Salvati, S. R.: *A Comparison of WISC IQ's and Altitude Scores as Predictors of Learning Ability of Mentally Retarded Subjects.* Unpublished doctoral dissertation, New York University, 1960; abstracted, *Diss.Abstr.* 21:2370, 1961.

237. Baumeister, A. A.: *The Dimensions of Abilities in Retardates as Measured by the Wechsler Intelligence Scale for Children.* Unpublished doctoral dissertation, George Peabody College for Teachers, 1961.

238. Thompson, J. M., and Finley, C. J.: The validation of an abbreviated Wechsler Intelligence Scale for Children for use with the educable mentally retarded. *Educ.Psychol.Measur.* 22:539-542, 1962.

239. Osborne, R. T., and Allen, J.: Validity of short forms of the WISC for mental retardates. *Psychol.Rep.* 11:167-170, 1962.

WISC: Bilingualism

240. Altus, G. T.: WISC patterns of a selective sample of bilingual school children. *J.Genet.Psychol.* 83:241-248. 1953.

241. Kralovich, A. M.: *The Effect of Bilingualism on Intelligence Test Scores as Measured by the Wechsler Intelligence Scale for Children.* Unpublished master's thesis, Fordham University, 1954.

242. Cooper, J. G.: Predicting school achievement for bilingual pupils. *J.Educ.Psychol.* 49:31-36, 1958.

243. Levinson, B. M.: A comparison of the performance of bilingual and monolingual native born Jewish preschool children of traditional parentage on four intelligence tests. *J.Clin.Psychol.* 15:74-76, 1959.

244. Levinson, B. M.: A comparative study of the verbal and performance ability of monolingual and bilingual native born Jewish preschool children of traditional parentage. *J.Genet.Psychol.* 97:93-112, 1960.

WISC: Cultural Variations

245. Levinson, B. M.: Traditional Jewish cultural values and performance on the Wechsler tests. *J.Educ.Psychol.* 50:177-181, 1959.

246. Levinson, B. M.: Subcultural variations in verbal and performance ability at the elementary school level. *J. Genet.Psychol.* 97:149-160, 1960.

WISC: Socioeconomic Status

247. Estes, B. W.: Influence of socioeconomic status on Wechsler Intelligence Scale for Children, an exploratory study. *J.Consult.Psychol.* 17:58-62, 1953.

248. Estes, B. W.: Influence of socioeconomic status on Wechsler Intelligence Scale for Children, addendum. *J.Consult.Psychol.* 19:225-226, 1955.

249. Roy, I., and Cohen, N.: Some psychometric variables relative to change in sociometric status; abstracted, *Am.Psychologist* 10:328, 1955.

250. Laird, D. S.: The performance of two groups of eleven-year-old boys on the Wechsler Intelligence Scale for Children. *J.Educ.Res.* 51:101-107, 1957.

WISC: Negro Samples, Negro-White Comparisons

251. Young, F. M., and Bright, H. H.: Results of testing 81 Negro rural juveniles with the Wechsler Intelligence Scale for Children. *J.Soc.Psychol.* 39:219-226, 1954.

252. Caldwell, M. B.: *An Analysis of Responses of a Southern Urban Negro Population to Items on the Wechsler Intelligence Scale for Children.* Unpublished doctoral dissertation, Pennsylvania State University, 1954.

253. Blakemore, J. R.: *A Comparison of Scores of Negro and White Children on the Wechsler Intelligence Scale for Children.* Unpublished master's thesis, College of the Pacific, 1952.

254. Racheile, L. D.: *A Comparative Analysis of Ten Year Old Negro and White Performance on the Wechsler Intelligence Scale for Children.* Unpublished doctoral dissertation, University of Denver, 1953.

II. THE WIDE RANGE ACHIEVEMENT TEST, THE ORAL READING AND ARITHMETIC SUBTESTS

The requirement ot the Survey for an individually administered, brief, well-standardized, reliable, valid, and flexible school achievement test was filled by the selection of the Reading and Arithmetic subtests of the 1963 revision of the Wide Range Achievement Test. The 1963 WRAT, by J.F. Jastak, replaces the original 1946 edition by Jastak and S.W. Bijou and appears to be quite similar to the original in design and item content, except that the new edition is divided, for the convenience of users, into two levels (Level I covers ages 5 to 12 years; Level II, 12 years through adulthood), in contrast with the broad sweep of the original, from kindergarten through adulthood.

The principal difference between the two editions appears to be in the method of standardization. The 1946 norms were computed to conform to those of the New Stanford Achievement Test (Reading, to New Stanford Word and Paragraph Reading, and Arithmetic Computation, to New Stanford Arithmetic Computation), whereas the 1963 norms, in each age bracket, depend on "probability samplings based on IQ's . . . that would correspond to the achievement of mentally average groups with representative dispersions of scores above and below the mean" (301).

The purpose of this section is both to review the literature on the WRAT and to evaluate it in relation to its suitability for the objectives of the Survey. Unfortunately this must be done almost entirely on the basis of the tests, manuals, and research available on the 1946 edition, which is itself extremely limited. Appropriate data for critical evaluation of the 1963 edition are almost totally lacking. Although released for sale in 1963, the test manual for this edition was still incomplete in June 1964 (301), and no independent data on validity have been found.

EVALUATIVE CRITERIA

Measurement experts believe that in addition to the standard questions concerning such issues as reliability, validity, representativeness of standardization sample, and agreement of norms with criterion levels, some problems are inherent in the wide-range type of design. These are stated forthrightly by Chauncey and Dobbin (310), in a discussion of various defects of tests:

> The "wide-range" test . . . is the too-short test in disguise. There are only a few of them around. They are promoted as being suitable measures of ability (or achievement) for people of many ages—from third grade through second year of college, for example. Since only a small part of any such test can be material suitable in difficulty for one individual, the effective part of the test may amount to no more than half a dozen questions—making it a very short test, indeed.

These remarks, by the president and one of the project directors of the Educational Testing Service, in a book written expressly to defend educational testing at a time when it is under

attack from many sources, command attention and concern by users of wide-range tests such as the WRAT. The particular implication of the critique is that reliabilities, validities, and score levels must be evaluated at every level covered (or at least at *every* level at which the test is used) and that broad-band coefficients of reliability and concurrent validity are likely to be misleading.

The problem of selecting a suitable achievement test for the Survey is highly complex. Time restrictions favor short forms and short-cut methods (such as the wide-range approach), provided that they meet reasonable standards of acceptability. However, it is just as true in testing as in all other areas that "you cannot get more out than you put in." Compromises with reality in testing often mean less reliable measures and less adequate coverage of appropriate universes of content; sometimes they mean penalties in relation to validity and consequent generalizability of measures.

The application of these points to the WRAT is considered as judicially as possible in this review, and the reality demands are weighed against possible shortcomings of this wide-range test in relation to alternatives available in the situation. A brief review of the 1946 edition and the general conceptualization of the WRAT is followed by a review of the 1963 edition used in Cycle II.

1946 EDITION OF WRAT

The conceptualization and rationale of this test (302) could not help but appeal to clinical psychologists in schools and mental health services. Jastak made an extremely strong case for the clinical use of his test, and it is not surprising that the WRAT has enjoyed considerable popularity in clinical circles despite psychometricians' prejudice against wide-range tests.

Jastak's arguments are briefly as follows:

1. A thorough psychological examination should include tests of school fundamentals as well as intelligence tests. Intelligence tests account for only a portion of the variance in school achievement, and failure in school and life adjustment may result from factors other than low intelligence.
2. Reliable (and valid) school tests should be used to assess discrepancies between intellectual capacity and performance in basic school subjects as well as discrepancies in the organization of learning abilities. Wide range discrepancies in school achievement are the rule rather than the exception, and their discovery is important for the understanding of personality and school performance problems and for the institution of proper remedial programs.
3. Clinically recognized discrepancy patterns in children are illustrated by the tendency of neurotic and disorganized children to be more proficient in reading than in arithmetic. In addition, "if neurotic tendencies and special reading handicaps occur together the child may function far below the level of his true capacity in all school subjects." Of course, failure in reading and in arithmetic may also reflect unrelated processes.

Jastak's criteria of a satisfactory school achievement test for (individual) clinical use are (a) *low cost,* (b) *individual standardization,* (c) *ease and economy of administration,* (d) *suitability of contents,* (e) *relevance of the functions studied, and* (f) *comparability of results over the entire range of the skills in question.* It is apparent that these criteria do in effect exclude such standard school achievement batteries as the Stanford, Iowa, Cooperative, and other well-known and highly respected batteries that are designed for group administration within a narrow grade range and cover a large universe of content, requiring considerable time to administer and score. These criteria certainly appear to be "tailor made" for the Survey (as well as for clinical practice). However, in view of the testing conditions for individually selected members of the national sample, the question is, how well are they implemented in the WRAT?

Jastak's views on test content are of particular interest. The WRAT focuses entirely on three basic school study skills—reading, spelling, and arithmetic—"around which most school studies revolve." The range of the subtests for each is indeed wide, from kindergarten to college.

The test content is concerned principally with mastery of the mechanics of the subject

rather than with comprehension. Thus the reading test is in effect a test of reading as a motor skill; the spelling test focuses on words without sentence contexts; and the arithmetic test involves number facility with minimal dependence on reading.

This emphasis is a reflection of the author's conception of the WRAT as an *adjunct* to tests of intelligence and behavior adjustment. Information concerning the subject's ability to comprehend can be obtained from intelligence tests, but accurate measurement of mechanics in the basic tools chosen is essential because of the dependence of most other studies on them. Further, it is argued that correct answers can often be given in conventional reading, arithmetic, and other subject-matter achievement tests on the basis of general knowledge and intellectual ability, even when mastery of mechanics is poor; thus, important diagnostic cues are overlooked.

Although the WRAT Reading and Arithmetic tests were reported to correlate satisfactorily with other achievement tests, their limitations of content and intended use were clearly outlined in the manual.

As stated above, the 1946 edition of the WRAT was standardized by anchoring the WRAT norms to those of corresponding subtests of the New Stanford Achievement Test. The standardization sample consisted of the scores of 4,052 students for Spelling and Arithmetic (about 1,500 were individually tested; the remainder were tested in groups) and 1,429 students, individually tested, for Reading. Reliability coefficients (retest) were reported as 0.95 for Reading (N=110) and 0.90 for Arithmetic (N=120). The Reading section of the New Stanford Achievement Test was reported to have correlated 0.81 with Paragraph and Word Reading; the Arithmetic section of the Stanford test correlated 0.91 with Arithmetic Computation.

The detailed composition of the various samples was not reported in the 1946 manual, and the validation data were not specified by age level as would be required to conform with the evaluative criteria discussed above. This was not exceptional in 1946, however, when the professional demands for rigorous reporting of critical information by test publishers were less stringent than they are today.

Nevertheless, despite the absence of comprehensive statistical information, the WRAT became a favorite of a large number of clinicians, and its use was extensive in the United States and abroad within a short time of its publication. It may appear surprising that so popular a test generated so little research. However, it appears that the principal use of the test was by clinicians whose attitudes toward tests are usually validated more by clinical experience than by statistics and whose opportunities and motivations to conduct and publish research are generally limited.

RESEARCH ON THE 1946 WRAT

It is noteworthy that only seven research reports have been found dealing with the 1946 edition and that of these seven, two were unpublished mimeographed papers (303 and 306) furnished by Dr. Jastak. Reliability coefficients and correlations of the WRAT with other tests, abstracted from these reports and the two test manuals (301 and 302), are reported in tables 4 and 5.

Reading

Hopkins, Dobson, and Oldridge (304) quoted Sundberg (312), in a 1961 paper, to the effect that although the WRAT was the second most popular achievement test in clinics, Sundberg could not find a single empirical study of it. They administered the Reading subtest to 502 children in grades 1 to 5 and correlated the scores with teacher ratings and scores on the California Reading Test (CRT). The correlations with teacher ratings were high for grades 1 to 5—0.79, 0.74, 0.86, and 0.85, respectively. The correlations with the total score of the California Reading Test were 0.86 for grade 3 and 0.71 for grade 5. The mean grade placements on the WRAT, for the five grades in order, were 1.4, 2.4, 3.5, 4.1, and 4.7.

Wagner and McCoy (303) reported correlations of the WRAT Reading subtest with the Sangren-Woody Silent Reading Test (grade level) for two samples, one of 29 fifth graders and the other of 57 primary school juvenile offenders. The correlations were 0.78 and 0.74. In the first sample, the WRAT Reading correlated 0.78 with both teacher ratings and with rank order of midterm grades. The correlation with the Stanford Reading Test, in the second sample, was 0.80.

Table 4. Studies reporting reliability coefficients of the WRAT

Investigator	Year	Subjects	Type of coefficient	Age range	Subtest of WRAT	Number	Reliability coefficient	Subtest of WRAT	Number	Reliability coefficient
Jastak and Bijou (302).	1946	Normals[a]	Test-retest	N.R.	Reading	110	0.95	Arithmetic	120	0.90
Jastak (301)	1963	N.R.	Split-half		Reading, Level II.			Arithmetic		
				20+ years		200	0.99		200	0.97
				18-19 years		200	0.98		200	0.97
				16-17 years		200	0.99		200	0.95
				15 years		200	0.99		200	0.97
				14 years		200	0.99		200	0.96
				13 years		200	0.99		200	0.96
				12 years		200	0.99		200	0.94
					Reading, Level I.			Arithmetic		
				11 years		200	0.99		200	0.95
				10 years		200	0.99		200	0.95
				9 years		200	0.99		200	0.94
				8 years		200	0.99		200	0.95
				7 years		200	0.99		200	0.96
				6 years		200	0.99		200	0.96
				5 years		200	0.98		200	0.97
		Standardization population.	Form I with Form II.		Reading			Arithmetic		
				14-0 - 14-11		89	0.88		87	0.86
				13-0 - 13-11		224	0.90		194	0.87
				12-6 - 12-11		180	0.94		165	0.85
				12-0 - 12-5		179	0.92		164	0.86
				11-6 - 11-11		252	0.91		225	0.85
				11-0 - 11-5		197	0.91		191	0.82
				10-6 - 10-11		214	0.93		195	0.89
				10-0 - 10-5		207	0.90		190	0.84
				9-6 - 9-11		165	0.91		160	0.79
				9-0 - 9-5		81	0.90		78	0.88

[a]Level of subjects and time interval between tests not reported.

NOTES: All correlation coefficients are Pearson Product-Moment unless otherwise specified.

N.R.—Not reported.

Table 5. Studies reporting correlation between the WRAT and other measures

Investigator	Year	Test or criterion variable	Subjects[a]	Age range	Number			Correlation
					Σ	M	F	
		WRAT Reading Test						
Smith (126)	1961	Full Range Picture Vocabulary Test.	Normals, Grade 2.	6-11 - 8-10	100	51	49	0.42
Hopkins, Dobson, and Oldridge (304).	1962	California Achievement Test	Normals	N.R.	257	---	---	
		Reading Vocabulary	Grade 3	N.R.	171	---	---	0.83
			Grade 5	N.R.	86	---	---	0.67
		Reading Comprehension	Grade 3	N.R.	171	---	---	0.84
			Grade 5	N.R.	86	---	---	0.67
		Total Reading	Grade 3	N.R.	171	---	---	0.86
			Grade 5	N.R.	86	---	---	0.71
Smith (126)	1961	California Test of Mental Maturity	Normals, Grade 2.	N.R.	100	51	49	0.47
Lawson and Avila (305)	1952	Gray Standardized Oral Reading Paragraphs Test.	Mental defectives.	16-45 years	30	19	11	[b]0.94
Reger (307)	1962	Metropolitan Achievement Tests, Reading.	Retarded boys.	9-9 - 14-6	25	---	---	[b]0.76
Wagner and McCoy (303)	N.R.	Midterm grades	Normals, Grade 5.	N.R.	29	---	---	0.78 (rank order)
Jastak and Bijou (302)	1946	Stanford Achievement Test, Reading	Normals, Grades 7 and 8.	N.R.	389	---	---	
		Word Meaning		N.R.	389	---	---	0.84
		Paragraph Meaning		N.R.	389	---	---	0.81
Wagner and McCoy (303)	N.R.	Sangren-Woody Reading Test.		N.R.	86	---	---	
			Normals, Grade 5.	N.R.	29	---	---	0.78
			Juvenile offenders.	N.R.	57	---	---	0.74
		Stanford Reading Tests	Juvenile offenders.	N.R.	47	---	---	0.80
		Teacher rating of reading ability	Normals, Grade 5.	N.R.	29	---	---	0.78
Hopkins, Dobson, and Oldridge (304).	1962	Teacher rating of reading ability	Normals		502	---	---	
			Grade 1	N.R.	90	---	---	0.79
			Grade 2	N.R.	106	---	---	0.74
			Grade 3	N.R.	171	---	---	0.86
			Grade 4	N.R.	49	---	---	0.86
			Grade 5	N.R.	86	---	---	0.85
Smith (126)	1961	Wechsler Intelligence Scale for Children.	Normals, Grade 2.	N.R.	100	51	49	
		Verbal Score						0.55
		Performance Score						0.47
		Full Score						0.61

See footnotes at end of table.

Table 5. Studies reporting correlation between the WRAT and other measures—Con.

Investigator	Year	Test or criterion variable	Subjects[a]	Age range	Number			Correlation
					Σ	M	F	
		WRAT Arithmetic Test						
Holowinsky (309)	1961	California Reading Test	Normals and retarded.	12-17 years	600	---	---	0.61
Murphy (306)	N.R.	First-quarter grades	Normals		241	---	---	
			Grade 5	N.R.	135	---	---	0.64
			Grade 6	N.R.	106	---	---	0.56
Holowinsky (309)	1961	Grade placement	Normals and retarded.	12-17 years	600	---	---	0.31
Reger (307)	1962	Metropolitan Achievement Tests, Arithmetic.	Retarded boys.	9-9 - 14-6	25	---	---	[b]0.87
Jastak and Bijou (302)	1946	Stanford Achievement Tests, Arithmetic Computation.	Normals, Grades 7 and 8.	N.R.	140	---	---	0.91
Holowinsky (309)	1961	Otis Quick Scoring Mental Ability Tests.	Normals, retarded.	12-17 years	600	---	---	0.30
				12-13 years	N.R.	---	---	0.59
				13-14 years	N.R.	---	---	0.39
				14-15 years	N.R.	---	---	0.54
				15-16 years	N.R.	---	---	0.02
				16-17 years	N.R.	---	---	0.09
Murphy (306)	N.R.	Stanford Achievement Tests, Arithmetic, and school grades.	Normals		241	---	---	
			Grade 5	N.R.	135	---	---	0.59
			Grade 6	N.R.	106	---	---	0.35
		Stanford Achievement Tests, Arithmetic, and school grades.	Normals		241	---	---	
			Grade 5	N.R.	135	---	---	0.75 (Multiple r)
			Grade 6	N.R.	106	---	---	0.70 (Multiple r)

[a]Designation of subjects are always white Americans unless otherwise specified.
[b]Spurious correlation with age for small N.

NOTES: All correlation coefficients are Pearson Product-Moment unless otherwise specified.

Σ —Total population; M—male; F—female; N.R.—not reported; r—correlation.

The report by Lawson and Avila (305) of a correlation of 0.94 between the WRAT Reading subtest and the Gray Oral Reading Test, administered to a sample of retarded adults ranging widely in age and IQ, is probably inflated because of the nature of the sample. Similarly, Reger's (307) sample of 25 emotionally disturbed, retarded boys (age range 9-9 to 14-6) is also quite a diverse population. Reger reported a correlation of 0.76 between the WRAT Reading subtest and the Metropolitan Achievement Test.

Holowinsky (309) had an apparently well-designed sample of 600, including 75 children at each age from 12 to 16 years. Each group was divided into three categories on the basis of IQ scores. The categories were as follows: 80-89 IQ, 90-99 IQ, and 100-109 IQ. For the total sample of 600 children, the California Reading Test correlated 0.61 with the WRAT Arithmetic subtest. Students of lower intellectual ability tended to show better achievement in arithmetic than in reading. For the total sample of 600 children the WRAT had a correlation of 0.31 with grade placement.

These limited results tend to support the claims for the WRAT with regard to concurrent validity both with other reading tests and with grade placement. The evidence is far from sufficient to permit definitive evaluation, and the lack of information on many points is obvious. However, no contrary evidence was found and as far as these papers are concerned, the report for the WRAT Reading subtest is favorable.

Arithmetic

The most adequate independent study of the WRAT Arithmetic subtest is that of Murphy (306), who tested 135 fifth and sixth graders (with average IQ of 114) with the WRAT and the Stanford Achievement Test (SAT). The correlation of the two tests was 0.59 for grade 5 and 0.35 for grade 6. The correlations between Arithmetic grades and the WRAT were 0.64 for grade 5 and 0.56 for grade 6. Correlations between the SAT and Arithmetic grades were 0.68 for grade 5 and 0.59 for grade 6. In Reger's sample, noted above (307), the WRAT Arithmetic test had a correlation of 0.87 with the Metropolitan Achievement Test. Holowinsky's study mentions a correlation of 0.59 between the IQ scores of 12-year-olds and the WRAT Arithmetic subtest, as compared with 0.71 for the Reading subtest.

These results are less satisfactory than those for Reading in the respect that the correlations reported compare less favorably with those mentioned in the manual. This type of cross-validation is imperative and demonstrates the importance of independent reports to supplement the data provided in a test manual. To Dr. Jastak's credit, however, it should be noted that the Murphy report, in which the lower correlations appear, is an unpublished paper which he, Dr. Jastak, furnished unsolicited for this review. These studies are insufficient for an evaluation of the WRAT Arithmetic subtest, to be sure. As the only information available, they leave the case for the Arithmetic test without strong independent support.

1963 EDITION OF WRAT

Two major changes appear in the 1963 edition. One is the division of the test into two levels. Level I covers the age range of 5 to 12 years; Level II covers the age range 12 years through adulthood. It is pointed out in the mimeographed manual for this edition that this change not only has reduced the time of test administration, but also has increased the number of items at each level, thereby increasing "the already high reliability" of the test. Indeed, the test has been lengthened, and the reliabilities have been listed for samples of 200 each for ages 5 through 11 years (Level I). For Reading, all—with the exception of 5 years of age—correlate 0.99. (Age 5 correlates 0.98.) Similarly computed reliabilities for Arithmetic are listed at or above 0.94, with the highest correlation, 0.97, occurring at 5 years of age. Since these coefficients are based on correlations between two forms of the test, they are considered by the authors to be inflated. The text of the reliability section of the manual (301, p. 47) states that the reliability coefficients are more likely within the range 0.90 to 0.95 with a mean of 0.92. At this level, they do not seem perceptibly higher than the reliabilities reported in the 1946 manual.

The second major change is in method of standardization. The 1963 manual (301) describes

the development of norms and the normative population sample as follows:

> The revised WRAT was administered to school children and adults in a number of states: Delaware, Pennsylvania, New Jersey, Maryland, Florida, Washington, and California. *No attempt was made to obtain a representative national sampling.* Nor is such a sampling *considered essential for proper standardization.* (italics added)
>
> The groups of children were selected from schools of known socioeconomic levels. The IQ's of the children were also known from group tests such as the Lorge-Thorndike, the Kuhlmann-Anderson, and the California Mental Maturity Test, administered at the schools. Many of the cases (over 1,000) in the standardization group had been given individual tests such as the Stanford-Binet, Wechsler Intelligence Scale for Children, and others. *In each age bracket, probability samplings based on IQ's were studied to develop WRAT norms that would correspond to the achievement of mentally average groups with representative dispersions of scores above and below the mean.* (italics added)

From the standpoint of the Health Examination Survey, with particular reference to Cycle II (children aged 6-11 years), the first of the two mentioned changes is an advantage. The age range of Level I fits the age range of Cycle II perfectly, and the increased length of the test and more extensive reliability studies reported support the claim of excellent reliability. The second change, in standardization and norm development, does, however, present a potential problem which is accentuated by the absence of validity data. This is discussed below.

Validity and Norms

Although published in 1963, the validity section of the revised WRAT was not available for review until late in June 1964. The delay was explained by the author of the test as occasioned by comparison of the WRAT "with a number of other tests in order to determine the meaning and diagnostic value of the three subtests in relation to other abilities." In addition, his letter disclosed that "specific methods to identify, in individual cases, the size of the independent and separate variances will have to be developed. Since this is somewhat of a novel and pioneering venture, it takes more time than routine manual preparation." The latter quotation is discussed separately below.

The basis for the present evaluation is, then, a comparison of the content and structure of the 1946 and 1963 editions of the WRAT, supplemented by the limited independent literature on the 1946 edition, reviewed above, and the limited data on the 1963 edition provided in the manual furnished by the author. No independent studies of the 1963 edition were available.

Comparison of the Two Editions

Examination of the two booklets indicates close similarity in item content, format, administration, and scoring. The Reading test for Level I, in the revised edition, contains 55 words that were in the 1946 edition, and their rank order of sequential position in the two editions is about 0.99. It is presumed that the 20 new words were empirically calibrated to fit into the previously established word order. The arithmetic items of the new test are of the same general type as in the earlier test, although the format is slightly different and the number of items is increased.

In view of this similarity, it appears reasonable to expect that the network of correlations of the revised test with other measures would be approximately the same as that reported for the 1946 edition. In fact, the correlations might even be slightly higher as a result of the greater length of the revision. To the extent that concurrent validity could be accepted for the 1946 edition, therefore, there is no reason to doubt that it will be upheld with the 1963 edition. Although the data are quite inadequate, tentative acceptance on this point appears warranted, based on the authors' reputations and the statements in the manual. However, this is only part of the problem.

Validation of 1963 Edition

It is equally important to be able to meaningfully interpret the grade ratings, standard scores, and percentiles in relation to individual age and

grade placement and in relation to population parameters. In the absence of empirical information on this issue, nothing definite can be concluded. It is appropriate to raise some questions which have been generated by statements made in the 1963 manual.

In the first place, the reviewer would take issue with the test author's statement that a representative national sampling is not essential for proper standardization. A national sample is certainly necessary if national norms are to be promulgated. Although the 1946 edition was developed on a restricted (as opposed to national) sample, its norms were presumably keyed to the grade norms of the New Stanford Achievement Test, for which a more extensive base existed. Even though regional, ethnic, and other perturbing effects were not known, it was at least possible to invoke the Stanford norms in interpreting grade levels. With the 1963 edition, however, no such anchoring process was followed. The only indications concerning age-grade levels are, in fact, disquieting.

The manual goes on to say that intelligence quotients of a number of group and individual tests (which are generally known to vary in level among themselves) were used to select samples in each age bracket "that would correspond to the achievement of *mentally average groups with representative dispersions of scores above and below the mean.*" (italics added) It would indeed be remarkable if such a procedure could produce a standard reference sample of known characteristics for normative purposes. Therefore it is doubtful that the resulting norms could have dependable accuracy for individual assessment or for analysis of groups in the manner required for the national sample of the Health Examination Survey. Perhaps the test author's current concern with comparisons with other tests, referred to above, reflects realization of this problem.

Furthermore, in view of the professed clinical purposes of the WRAT, it is surprising that the standardization research is confined to "mentally average groups," and that no studies were undertaken of such groups as gifted pupils, students retarded in reading, arithmetic, and other school subjects, disturbed children, and subnormal children.

For the purposes of a national survey, problems of ethnic and regional variations in test performance are important, as are other sources of perturbation attributable to deviations of ability, personality, and physical and social factors. The absence of such data for the 1963 WRAT is certainly not the sole responsibility of the author-publisher; ordinarily test producers do not assume responsibility for all possible research of interest to all possible users. If a test attracts interest, information about it in various situations gradually accumulates in the literature. However, in the present case it appears fair to say that the author's confidence in his test led him to publish the revision before he had completed his own research and before research on it by any users could be reported. The test was issued without a formal designation of the norms as "tentative" and without any qualifications.

Validity Variances

Instead, the 1963 manual (301, p. 2) concludes its introductory section with the following paragraph:

> In addition to the three operational aspects (of mechanics and comprehension in relation to each skill test) the basic skills have several unique validities which will be explained later by reference to appropriate research. The validity variances will not only support the empirical distinctness of mechanics and comprehension, but will provide the degrees to which each is important in learning to read, spell and figure and the impact the relationship between them has on the total learning process.

The burden of proof is on the author. The development of such an analytic scheme for interpretation of test scores is indeed both novel and ambitious and deserves all the time required to complete it. It seems regrettable, however, that the test was released before critical users could evaluate not only these devices, but even the grade ratings, percentiles, and standard scores included in the manual.

Validity Data in 1963 Manual

The section of the manual entitled "Validity of the WRAT" (301, p. 51), contains a table of means and standard deviations of raw scores for

the Reading, Spelling, and Arithmetic subtests, which indicates considerable need for refinement of the tests in order to produce an even progression of scores from grade to grade. The difficulties are considerable at some levels (8.0 to 8.5, 9.5 to 10.0, and 10.5 to 11.0, on the Reading test, for example), to say nothing of the fact that the basic difficulties reported about the standardization sample are not only not clarified, but are not even referred to in this section of the manual.

Two paragraphs on the validity of the Reading test (301, p. 50) refer only to the studies cited above, which involve the 1946 edition of the WRAT. No validity data on the 1963 edition are presented. Similarly, data are presented (301, p. 52) on correlations of the WRAT with achievement tests and on the validity of the Arithmetic subtest, but these are also identified as relating to the 1946 edition.

Internal consistency data cited by the author (301, p. 53) involve intercorrelations among the three WRAT subtests and *not validity*, despite the author's assertion that "criteria of internal consistency, if properly interpreted, are usually more valid than are external criteria of comparison." These data are also presented as "one method of cross-validation."

Correlations of the Wide Range Achievement Test with the California Test of Mental Maturity are given (301, p. 54) for a sample of 74 children spanning the age range of 5 to 15 years. They range from 0.74 to 0.84 and may be spuriously high in view of the heterogeneity of the sample. Similarly structured comparisons with the WISC for 300 boys (aged 5 to 15 years) and 244 girls (aged 5 to 15 years) are reported which indicate correlations as follows:

Sex and test	Reading	Arithmetic
Boys		
Vocabulary[1]	0.65	0.56
Block Design	0.41	0.41
Girls		
Vocabulary[1]	0.56	0.56
Block Design	0.39	0.50

[1] Based on Jastak's short-form revision (311).

In view of the composition of the sample, these are surprisingly low.

The manual also reports (301, p. 55) correlations of WISC Verbal Scale, Performance Scale, and Full Scale with the WRAT (1963), with samples covering narrower age ranges of 5 to 7 years and 8 through 11 years. The results here are the most impressive concurrent validity data in the manual, although they indicate correlations in the 0.6 to 0.7 range with intelligence rather than achievement criteria, for which they are intended.

As stated several times earlier, the accuracy of score levels in the WRAT norms is regarded as a more pressing problem for empirical demonstration than the concurrent validity (covariation with related measures) of the test. On this point the validity section of the manual is silent.

Grade Equivalents

The 1963 manual (301, p. 22) states that grade norms were derived from "the actual mean grade levels of the children in each grade group." Despite variations in school grade-placement practices over time, grade rating is characterized as "rather stable." The manual further asserts "striking comparability" of grade ratings of the old and the new WRAT's "through nearly all educational levels except the upper ranges." Grade ratings below 14 years of age are said to be less arbitrary than grade ratings over 14 years of age. The grade scores are intended to be comparable to mental ages.

Standard Scores

The WRAT standard scores can be converted from raw scores by age group in a table provided in the manual. The standard score has a mean of 100 and a standard deviation of 15 and is intended to be equivalent to an IQ from the WAIS, WISC, Stanford-Binet (Form L-M) or any of the major intelligence scales. Although these scales are not comparable themselves (as developed in some detail in section I of this report), the manual states that "the results from the WRAT test can thus be directly compared with the major individual intelligence scales."

The standard score is asserted to be the "most precise and most meaningful score." It is the only score that is comparable between sub-

tests and that provides for uniform differences between scores.

Percentiles

Percentiles are included "because of their present popularity and convenience," but the manual appropriately downgrades them and discourages their use.

SUMMARY AND CONCLUSIONS

The foregoing review of the WRAT is necessarily incomplete because of lack of adequate information on which to base a technical evaluation. The test is well conceptualized and has much face validity, but standardization information on the 1946 edition was inadequate, and on the 1963 edition it is thus far insufficient.

Published research on the 1946 WRAT has been extremely limited and fails to answer most of the questions left unanswered by the authors' manual. Moreover, analysis of the available information on the 1963 edition raises doubts about normative score levels.

The selection of the WRAT over other available school achievement tests may be defended on the grounds of administrative expediency and suitability of the material for the purposes of the Survey, in spite of the fact that inadequate data exist to support the author's claims of validity. It is possible that such data may be produced, and every effort should be made to obtain them. However, unless these results are convincing—and reason to doubt that they will be has been expressed—it is recommended that serious consideration be given to carrying out a complete restandardization of the Reading and Arithmetic subtests on the entire national sample. Unless this is done, projections of estimates to population may be seriously in error.

BIBLIOGRAPHY

Research References and Manuals

301. Jastak, J. F.: *Wide Range Achievement Test,* rev. ed. Wilmington, Del. Guidance Associates, 1963.
302. Jastak, J. F., and Bijou, S. W.: *The Wide Range Achievement Test.* Wilmington, Del. C. L. Story Co., 1946.
303. Wagner, R. F., and McCoy, F.: Two validity studies of the Wide Range Achievement Reading Test. Personal communication.
304. Hopkins, K. D., Dobson, J. C., and Oldridge, O. A.: The concurrent and congruent validities of the Wide Range Achievement Test. *Educ.Psychol.Measur.* 22:791-793, 1962.
305. Lawson, J. R., and Avila, D.: Comparison of Wide Range Achievement Test and Gray Oral Reading paragraphs reading scores of mentally retarded adults. *Percept.Mot. Skills* 14:474, 1962.
306. Murphy, G. M.: An investigation of the utility of mathematics sub-test from the Wide Range Achievement Test, as applied to intermediate level groups. Personal communication.
307. Reger, R.: Brief tests of intelligence and academic achievement. *Psychol.Rep.* 11:82, 1962.
308. Warren, S. A.: Academic achievement of trainable pupils with five or more years of schooling. *Train.Sch.Bull.* 60:75-88, 1963.
309. Holowinsky, I.: The relationship between intelligence (80-110 I.Q.) and achievement in basic educational skills. *Train.Sch.Bull.* 58:14-22, 1961.

Other References

310. Chauncey, H., and Dobbin, J. E.: *Testing, Its Place in Education Today.* New York. Harper and Row, 1963.
311. Jastak, J. F., and Jastak, S. R.: Short forms of the WAIS and WISC vocabulary subtests. *J.Clin.Psychol.* 20:167-199, 1964.
312. Sundberg, N. D.: The practice of psychological testing in clinical services in the United States. *Am.Psychologist* 16:79-83, 1961.

III. THE GOODENOUGH DRAW-A-MAN TEST

BACKGROUND AND DEVELOPMENT

A comprehensive historical survey of the study of children's drawings appeared recently in an important new book by Dale B. Harris (522), a former colleague of Florence Goodenough and apparent successor to her in the leadership role in the measurement of children's intelligence by point scales based on drawings of the human figure. The present review does not duplicate Harris' scholarly survey, but focuses more specifically on the problems of the Goodenough Test as used in the Health Examination Survey.

The first formal intelligence test based on the analysis of children's drawings was published by Florence Goodenough (595) in 1926, but the literature on this subject goes back at least to 1885 (595, ch. I). Some of the early papers are summarized in this study, but the major emphasis has been placed on recent critical research on the Draw-A-Man Test and its variants. Nevertheless, it is of interest that in 1893 Herrick (501) demonstrated the developmental significance of profile drawings and that in the same year Barnes (502) recognized that drawings are used by young children as a means of expressing their ideas. Meanwhile, Lukens (503), in 1896, outlined many details of human figure drawings which were later incorporated in the point-scoring systems of Goodenough (595) and of Harris (522).

The Goodenough Test is referred to in this discussion as the Draw-A-Man Test although the specific instructions in Cycle II of the Survey are to "make a picture of a person." However, the instructions go on to state that "when a bust picture has been drawn intentionally, the child is given another sheet of paper with the instruction 'Now make a picture of a whole person.'" Only one picture is used.

Rationale

In this procedure emphasis is placed on the representation of details in the drawing to measure conceptual maturity. Drawing technique is minimized, and distortions potentially usable as cues for personality evaluation are not scored. Recent drawing tests focused on personality study have used two or more drawings. For example, Machover (596) instructs the subject to "draw a person" and then to draw a person of the sex opposite to the one previously drawn, while Buck (594) uses drawings of a house, a tree, and a person. In general, the cues and signs interpreted in personality study of drawings are different from those employed for the measurement of intelligence.

Point-Scoring System

The point system developed by Goodenough (595) for drawings which can be recognized as attempts to represent the human figure—no matter how crude—involves the presence or absence of 51 detailed points, which are listed as follows:

1-4a	Head, legs, arms, trunk present
4b	Length of trunk greater than breadth
4c	Shoulders definitely indicated
5a	Attachment of arms and legs
5b	Legs attached to trunk; arms attached to trunk at correct point
6a	Neck present
6b	Outline of neck continuous with that of the head, of trunk, or both
7a-c	Eyes, nose, mouth present
7d	Both nose and mouth shown in two dimensions; two lips shown
7e	Nostrils shown
8a	Hair shown
8b	Hair on more than circumference of head; nontransparent
9a	Clothing present
9b	At least two clothing items nontransparent
9c	Entire drawing free from transparencies of any sort; sleeves and trousers shown
9d	At least four clothing items definitely indicated
9e	Costume complete without incongruities
10a	Fingers present
10b	Correct number of fingers shown
10c	Detail of fingers correct

10d Opposition of thumb shown
10e Hand shown as distinct from fingers or arm

11a Arm joint shown (elbow, shoulder, or both)
11b Leg joint shown (knee, hip, or both)

12a-e Proportion: head, arms, legs, feet, two dimensions

13 Heel shown

14a-f Motor coordination
- a Lines reasonably firm and joining usually accurate
- b Increased firmness of lines and increased accuracy of line junctions
- c Head outline free from unintentional irregularity
- d Trunk outline free from unintentional irregularity
- e Arms and legs without irregularities, narrowing at point of body junction
- f Features symmetrical

15a Ears present
15b Ears in correct position and proportion

16a-d Eye detail, brow, lashes, or both shown; pupil shown; proportion; glance

17a Both chin and forehead shown
17b Projection of chin shown; chin clearly differentiated from lower lip

18a-b Profile drawings

Standardization

In Goodenough's original research, point scores based on these items were equated to age norms from which intelligence quotients could be computed in the same manner as in the Stanford-Binet test. Data on reliability and validity were reported in the 1926 book (595) and also in a monograph (504) published the same year. Using a basic standardization sample of 5,627 school children from kindergarten to the sixth grade aged 4 to 12 years, split-half and retest reliabilities were computed. A split-half reliability of 0.77 (corrected) was found to be constant from 5 to 10 years of age, and a retest reliability coefficient of 0.94 was reported for 194 first-grade children. Correlations with Stanford-Binet were 0.76 for mental ages and 0.74 for intelligence quotients. The experimental work, analysis, and reporting which characterized this undertaking would be regarded as impressive today, and the critical reader of Goodenough's book can well appreciate Lewis M. Terman's description of it (in the foreword) as "a notable accomplishment."

Perspective

In 1950, a quarter of a century after the publication of her book, Goodenough collaborated with Dale Harris in a review (510) of the extensive literature generated by her test. This review was critical of many studies of graphic expression that lacked quantification, but it acknowledged the value of drawings used projectively as a source of diagnostic cues. Goodenough and Harris made special note of some writers' attempts to attribute discrepancies between the Draw-A-Man Test and the Stanford-Binet (in which Draw-A-Man IQ's are markedly lower) as possible diagnostic cues of emotional or nervous instability or of brain damage. They also cautioned about the use of the Draw-A-Man Test in cross-cultural comparisons, pointing out that the Draw-A-Man is not a *culture-free* test, as many users have incorrectly assumed. This point is most dramatically illustrated by the Near Eastern study of Dennis (555).

In the *Fourth Mental Measurement Yearbook*, 1953, Stewart (514), while presenting a very favorable evaluation, suggested that the Goodenough norms might require revision due to social changes which have occurred since the original standardization. Such a revision was apparently justified, and the new Goodenough-Harris Drawing Test (552), published in 1963, fills an important need. This modified procedure consists of three drawings: a man, a woman, and "yourself." Separate point scales are provided for drawings of men and drawings of women; separate norms are also provided for drawings made by boys (men) and drawings made by girls (women).

An empirical study on a sample of 195 drawings taken from the Health Examination Survey population, in which the Harris scoring and norms were compared with the original Goodenough scoring and norms, is reported below. This study supports a recommendation that the Harris revi-

sion be adopted for scoring the Goodenough test in this Survey.

EVALUATION OF INTELLIGENCE BY HUMAN FIGURE DRAWINGS

Effective Range

Barnes' (502) early observation that children draw candidly up to about 14 years of age and then more abstractly is supported by Barnhart (507), who described three types of drawings—*schematic* (graphic representation), predominating in the age range 5 to 9 years; *mixed*, in the range 8 to 13 years; and *visual realistic* (abstracted, esthetic, nonspecific as to factual details), principally in the range 10 to 16 years. This apparently explains why the point scores cannot be validly extended above 14 years of age (522).

The increase in point scores with age, up to 14 years of age, apparently reflects mental maturity and not chronological age. This was noted by Smith (506) and by McElwee (524), who reported a correlation of 0.72 between the Draw-A-Man and the Stanford-Binet mental ages for a sample of 45 subnormal 14-year-old children. Israelite (562) found a correlation of 0.71 between the Draw-A-Man and the Stanford-Binet for 256 mental defectives. Others have also successfully tested mentally defective adults with the Draw-A-Man Test.

Relation to Artistic Ability

An area of special interest in the interpretation of children's drawings has been the relation of drawing "maturity," as reflected in point score, and artistic ability. Goodenough acknowledged that drawings could be influenced by special coaching (as can most human responses) but that ordinary art instruction in school has little effect on the Draw-A-Man score. She reported a correlation of 0.44 between the Draw-A-Man and teacher ratings of drawing ability (504).

Perturbing Factors

Intelligence scores based on drawings are relatively independent of artistic ability. However, there is evidence that both internal factors, such as health, emotions, and attitudes, and external environmental factors affect the drawing content. In the present review, studies have been found which demonstrate the influence on drawings of factors such as height and weight (543), sex and body image (512, 537-539, and 541), physical handicaps (571 and 572), mental age (521), affective states experienced and experimentally induced (529, 530, and 532), institutionalization (540), teacher attitude (533), sociometric popularity (534), social acceptance (531), and social class (536).

Although size of drawings appears to increase with mental age over the effective range of the Draw-A-Man, size standards have not been incorporated in any of the published point scores. In general, the studies referred to in the preceding paragraph may be viewed as minor perturbing influences within a homogeneous cultural framework. Variability among drawings attributable to perturbing factors of the types enumerated within the social boundaries of the American culture appears to have significance for the study of personality and social behavior, but it does not appear to influence measures of intelligence derived from children's drawings in the age range 5 to 12 years.

Culture

The factors which influence children's drawings of the human figure most are those that reflect the effects of a culture's customs and values, since these determine the way in which children are exposed to different representations of the human figure in dress, art, photographs, religious practices, and sex roles and attitudes. Hunkin (554) found the Goodenough norms inapplicable to Bantu school children, and Dennis (555) attributed the steady decline in mean Draw-A-Man IQ from 5 to 10 years of age (among Egyptian and Lebanese children in the Near East) to the Arab culture, which restricts access to representations of the human figure. Studies of the Draw-A-Man with children of various American Indian tribes on reservations (558-560) have produced varying results which may perhaps be understood only in the context of their respective culture patterns.

On the other hand, Anastasi and DeJesus (556) found sex differences in agreement with

Harris, discussed below, but found no ethnic differences in a comparison of Draw-A-Man scores of 50 Puerto Rican children of low socioeconomic class in New York City with those of Negro and white children of similar status which were reported by other investigators. Similarly, Levinson (243) found that the Draw-A-Man, as well as WISC Block Design, is culturally "fair" for native-born Jewish bilingual children in New York City.

The importance of taking into account cultural variations when dealing with a heterogeneous population such as that sampled by the Health Examination Survey is illustrated by the following quotations from Harris (522, pp. 131 and 132). These quotations have been exerpted to illustrate how the customary dress of Eskimo children affects point scores on drawings of the human figure.

> Eskimo children are less likely to depict the neck, the ears, and to correctly place the ears. These facts seem to reflect the greater prevalence of parkas in the Eskimo group's drawings and [this] is thus an artifact of the drawing situation. Due to the voluminous parka garments, elbow joints, knee joints and modeling of the hips are less likely [to be] shown, resulting in greater stiffness of figures portrayed.
>
> Since the Eskimo boot does not have a heel, Eskimo children are less likely to indicate heels in their drawings. [Several instances], however, show that when the garb is appropriate, the heel is shown. The children do have the concept of heels; their drawings are quite appropriate to the type of figure they are representing at the time. Eskimo children are also less likely to portray the arm and shoulder performing some type of movement, probably due to the loose parka, though this is not invariably the case.
>
> On the other hand, Eskimo children are more likely to portray with exactness the nostrils, the bridge of the nose, and, when portrayed at all, the thumb or fingers. The characteristic tendency of the Eskimo children to show a mittened hand earns for them a greater credit on the thumb opposition point and on the hand as distinct from fingers or arm in the age group ten to thirteen inclusive. In this age group also the Eskimo is more likely to draw the arms down at the side than held out stiffly from the body. The Eskimo child is more likely to show the feet with a wide stance, that is, with toes pointing apart, or in perspective in either full-face or profile drawings. The Eskimo drawings include fewer transparencies in these age groups, and a larger percentage of them earn credit for showing a distinct costume, which of course follows from the tendency to draw the parka—the everyday costume in this part of Alaska.
>
> Aspects of the Eskimo drawings that are distinctive and that are not apparent in the detailed scoring technique of the Goodenough method include: a greater emphasis on the eyebrow, on the nostrils and nose (as indicated above), and on general detail of facial features. There is some evidence of a general decrease in quality of the drawing in adolescence. This is not sufficiently great, however, to reveal itself markedly in the trend of median scores as in the normative group. It is most noticeable in the increased tendency to draw the facial features and hands "sketchily." Particularly among young Eskimo children there is a very distinct tendency to draw shorter arms and legs than in the norm group. Here again there is the possibility that the proportions of the body are distorted somewhat by so many children depicting the figures in parkas.

Cultural factors influence drawings in many obvious ways such as type of garb, vehicles, implements, and actions portrayed, but the nature of the influence on a Goodenough-type point score is subtle, as illustrated in the preceding quotations from Harris. Because such variations are often inconsequential within the mainstream of American culture, there has been a wide temptation to use the Draw-A-Man as a culture-free intelligence test. Nevertheless, as Harris properly insisted (522, p. 133), "the data . . . suggest that the child's drawing of certain body features or parts is influenced by garb, and possibly by other conditions of living that call attention to particular parts or their functions. *Allowance would have to be made, both in scoring and in*

the norms, for parts omitted in one of these cultures included in the present scoring system. Such allowance would have to be worked out empirically within each culture group." (italics added)

Goodenough and Harris (510), in their 1950 review, affirmed that although the test may be unsuited to comparing children *across* cultures, it may still rank children *within* a culture according to relative intellectual maturity. In his 1963 publication (522, p. 133) Harris has further amended this position to state that "for the most valid results, the points of the scale should be restandardized for every group having a distinctly different pattern of dress, mode of living, and quality or level of academic education." In Harris' judgment, "This conclusion virtually rules out the scale for cross-cultural comparisons; indeed, psychologists increasingly believe that mean differences among large, representative samples drawn from varying cultures express the gross differences in conceptual experience and training these groups have had. Further work, to determine exactly which aspects of intellectual or conceptual maturity the drawing task expresses, will be necessary to explain scientifically these observed cultural differences."

No systematic research such as Harris delineated with respect to Eskimo children has been done on the detailed effects of microvariations within the American culture. Yet there is little reason to doubt that subtle differences between urban and rural, industrial and suburban, warm climate and cold, eastern and western, and other prominent contrasting situations within the continental United States (to say nothing of Alaska and Hawaii) might produce some significant variations. Undoubtedly, some of these subcultural variations reflect ethnic factors, such as the superstitious reluctance of some southwestern children of Mexican origin to draw eyes because of fear of the "evil eye."

It is also possible that secular trends, which are revealed in the comparison of the 1926 and 1963 norms, may be occurring at differential rates in different localities and segments of the culture and that these also may subtly affect point scores. For example, the high-fashion announcements of transparent garments for females not only aroused different reactions among different segments of the population but also received widely varying prominence in different localities. Although this is an extreme example, it is nevertheless possible that some children might draw the female figure appropriately reflecting a sophisticated transparent garment and be penalized on the point score for what could be considered a "bright" response.

Sex Differences

Both Goodenough (504) and Harris (522) have reported qualitative and quantitative differences in drawings which are related to the sex of the person doing the drawing. Harris' more recent work is of greater relevance. He believes that these sex differences cannot be attributed to differential selection of boys and girls according to intellect. Harris' recent data show that sex differences in total point scores appear at an early age and are considerably greater than those reported by Goodenough. Harris found that for the drawing of a man, the mean score difference favors girls by about one-half year of growth at each year of age, while for the drawing of a woman, this difference is roughly equal to a full year of growth. The Harris point scale, applied differentially to Man and Woman drawings by boys and by girls, appears to reduce mean differences.

Sex differences in drawing point scores reflect differences in maturation, cultural factors—including sex role and awareness—and perhaps some degree of difference in drawing proficiency. However, it is believed that these will be minimized by the adoption of the Harris norms and scoring system and that the remaining residual error probably will be inconsequential. Without doubt, the error will be smaller than that which would result from the blanket use of one uniform scoring system for the entire population.

PERSONALITY STUDY BY CHILDREN'S DRAWINGS

Although personality evaluation is not the primary reason for including the Draw-A-Man Test in the Survey, a review of the potentialities for such analysis is relevant. Since this topic has been covered more extensively by Harris in his recent publication than in this review, the following

discussion is organized in relation to Harris' summary. Below are eight widely accepted but not necessarily established generalizations concerning personality measurement by children's drawings. These were evaluated by Harris in his recent book (522, p. 52). As will be noted, several of the generalizations are rejected.

1. *Drawing interpretation is more valid when based on a series of a subject's protocols than when based on one drawing.* Despite the lack of clear-cut empirical evidence on this issue, Harris equates additional pictures as having the effect of increasing the length and therefore the reliability of the test. From this logical viewpoint, he considers it justified.
2. *Drawings are most useful for psychological analysis when teamed with other available information about the child.* This, too, is a logically sound principle, "especially when it is the content of drawings alone that is being used for psychological interpretation."
3. *Free drawings are more meaningful psychologically than drawings of assigned topics.* This is probably true for certain purposes, such as exploration of interests, but systematic comparison of individuals, as in a national survey, requires control of the task.
4. *When a human figure drawing is assigned, the sex of the figure first drawn relates to the image the drawer holds of his own sex role.* Of the studies summarized in Appendix III, those most relevant to the study of children ages 6 to 12 years are as follows: 512, 537-539, 541, and 542. According to Brown and Tolor (541), normal individuals of both sexes tend to draw their own sex first, while persons with behavior disorders draw the opposite sex first. Harris agrees that most children of either sex will draw their own sex first when asked to "draw a person." He further elaborates that as girls grow older there is an increasing tendency for them to draw a male figure. This, he feels, reflects both the cultural preference given to the male role and an increasing dissatisfaction with the female role.

   Harris also hypothesizes that the male figure is more culturally stereotyped and easier to draw than is the female figure. He considers *deviates* from this norm to be psychologically different from non-deviates. He also feels that the deviation has different meanings for the two sexes and has unique, idiosyncratic meanings to individuals. Since many deviations from the norm occur and since the meaning of such deviations is as yet unknown, it is unlikely that the principle (the figure drawn first relates to the image the drawer holds of his own sex role) is universally valid. Therefore, even though about 86 percent of boys and 65 percent of girls have been reported to draw their own sex first, it is not possible to formulate any reliable interpretation for those who do not.
5. *A child adopts a schema or style of drawing which is peculiar to him and which becomes highly significant psychologically.* Most of the evidence is opposed to this and suggests rather that developmental patterns do exist among children's drawings.
6. *The manner in which certain elements are portrayed in drawings may be used as signs of certain psychological states or conditions in the artist.* In agreement with Harris, the present writer regards this statement as one of the eternal, unfulfilled wishful myths of the "depth psychologist." Two particular statements by Harris are relevant to possible further research in this frustrating area. First, "whether or not 'signs' are selected by an empirical or deductive procedure, there is still the question whether form or content will provide the cues. Size, quality or texture of line, degree of angularity, pattern or shape, and placement on the page are often thought to be highly significant avenues for 'projecting' unconscious motives or needs." References 512, 521, 537, 540, 543, 564, and 566 support this view, but neither form nor content signs of unequivocal value have thus far been validated. Thus, Harris' second statement, that "useful and valid signs leading to dependable conclusions are, for the

most part, still to be ascertained," disposes of this generalization.

7. *Drawings must be interpreted as wholes rather than segmentally or analytically.* This, too, has been a strong sentimental favorite, but the evidence is mostly the other way, particularly in personality assessment. In fact, the history of psychometric progress has been away from global analysis toward specific analysis, has favored linear over curvilinear relations, and generally has demonstrated that quantitative procedures are more valid, even if less spectacular, than those based on scorer judgment.

   Harris has cited analytic studies of component qualities of children's drawings, by Martin and Damrin and by Stewart (522, p. 56), which suggest that "drawings are actually appraised in terms of a few general dimensions, although they may be rated on a number of specifically defined elements or qualities." Harris believes that these studies lend credence to the belief that broad, dimensional evaluations (rather than highly particularistic ones), based on such analytic results, may be made more readily and more reliably. He also believes that they suggest the direction these quantitatively and factorially defined "global" ratings may take. "Their findings in relation to personality qualities, however, are not of such magnitude as to support the use of drawings in diagnosing individual cases."

8. *The use of color in drawings can be significant for studying personality.* This is another popular clinical belief, on which the empirical evidence is equivocal.

RESEARCH ON THE GOODENOUGH TEST

Reliability Studies

Table 6 summarizes the reliability coefficients reported for the Draw-A-Man Test in the studies included in this review (523-528). In general, the reliabilities obtained by independent investigators have confirmed those reported by Goodenough. The reliability of the point scale holds up in the mentally retarded range (523 and 524), and scorer agreement is high (526).

One problem observed in interscorer comparisons by the reviewer which is mentioned in connection with the Goodenough vs. the Goodenough-Harris comparison is that while the results of two scorers may show a very high correlation, there may nevertheless be a constant difference in score levels between them, reflecting individual idiosyncrasies of their interpretations. The safest method of coping with such constant errors, in a survey in which a number of scorers may be used for different segments of the total sample, would be to have at least two people score every test and to use the average of the two for record.

Correlations With Other Tests

Correlations of the Draw-A-Man with the Stanford-Binet are summarized in table 7, and its correlations with other tests, in table 8. Similar tables appear in Harris (522, pp. 96 and 97). With few exceptions, correlations of the Draw-A-Man with the Stanford-Binet (in which coefficients are based on IQ's) reported by other investigators have averaged lower than those reported by Goodenough in 1926 (504). The exceptions found are Williams (505), Israelite (562), White (565), and Ellis (unpublished master's colloquim paper, University of Minnesota, 1953), whose data agree substantially with those of Goodenough.

Unfortunately, most of the publications cited which involve correlations of the Draw-A-Man with the Stanford-Binet and a number of other tests are based on very small samples (rarely more than 100), are usually not representative of their respective subuniverses, and do not always present assurance of testing under standard conditions. As a result, the collection of correlation coefficients can only be interpreted very generally.

These results indicate a considerable association between the Draw-A-Man Test and general intelligence tests, such as the Stanford-Binet and the WISC, which measure mental maturity. The common variance is probably about 50 percent. Maturationally, the original rationale presented by Goodenough—that drawing point

Table 6. Studies reporting reliability coefficients of human figure drawing tests

Investigator	Year	Test and scoring method	Subjects[a]	Age range	Number			Type of coefficient	Reliability coefficient
					Σ	M	F		
Yepsen (523)----	1929	Goodenough-----	Feebleminded-----	9.0 - 18.2	37	37	-	Test-retest	
								Administration 1-2-----	0.89
								Administration 2-3-----	0.91
								Administration 1-3-----	0.91
Brill (525)-----	1935	Goodenough-----	Feebleminded-----	N.R.	N.R.	---	---	Test-retest	
					71	71	-	Administration 1-2-----	0.77
					65	65	-	Administration 2-3-----	0.80
					67	67	-	Administration 1-3-----	0.68
Albee and Hamlin (579).	1949	Human Figure Drawing, Paired Comparisons.	VA Mental Hygiene Clinic. Range—normals to psychotics.	N.R.	N.R.	---	---	Interjudge----------	0.95
								Spearman-Brown----------	0.98
Albee and Hamlin (581).	1950	Machover-------	Neurotic, schizophrenic, normal.	N.R.	72	---	---	Interjudge----------	0.89
Hinrichs (586)--	1935	Goodenough-----	Normals----------	10-18 years	81	---	---	Split-half, Spearman-Brown.	0.88-0.90
Herron (532)----	1957	Goodenough-----	Normals, Grades 3 and 4.	113 months (mean)	16	16	-	Test-retest, group A,[b]	
								Administration 1-2-----	0.52
								Administration 2-3-----	0.51
								Administration 1-3-----	0.27
					28	-	28	Test-retest, group A[b]	
								Administration 1-2-----	0.79
								Administration 2-3-----	0.69
								Administration 1-3-----	0.85
					24	24	-	Test-retest, group B[b]	
								Administration 1-2-----	0.92
								Administration 2-3-----	0.40
								Administration 1-3-----	0.86
					15	-	15	Test-retest, group B[b]	
								Administration 1-2-----	0.85
								Administration 2-3-----	0.73
								Administration 1-3-----	0.63
McCurdy (527)---	1947	Goodenough-----	Normals----------	83.9 months (mean)	59	59	-	Test-retest----------	0.69
Buhrer, de Navarro, and Velasco (511).	1951	Goodenough-----	Normals, Spanish-speaking.	7-14 years	1,936	---	---	N.R----------	0.97
Frankiel (518)--	1957	Goodenough and Frankiel.	Normals----------	----------	200	100	100	----------	----------
				7 years	100	50	50	Intrajudge----------	0.83
				7 years	100	50	50	Interjudge----------	0.71-0.84
				12 years	100	50	50	Intrajudge----------	0.89
				12 years	100	50	50	Interjudge----------	0.81-0.86
McHugh (508)----	1945	Goodenough-----	Normals, pre-school.	62.0 months (mean)	83	---	---	Test-retest----------	0.46 (IQ) 0.51 (MA)
Goodenough (504).	1926	Goodenough-----	Normals----------	4-12 years	5,627	---	---	----------	----------
								Split-half, Spearman-Brown.	0.77
								Test-retest, Grade 1 only-	0.94

See footnotes at end of table.

Table 6. Studies reporting reliability coefficients of human figure drawing tests—Con.

Investigator	Year	Test and scoring method	Subjects[a]	Age range	Number			Type of coefficient	Reliability coefficient
					Σ	M	F		
Williams (505)--	1935	Goodenough-----	Normals---------	3-15 years	100	50	50	Interrater----------	0.80-0.96
Smith (506)-----	1937	Goodenough-----	Normals---------	----------	1,000	---	---	Test-retest----------	----------
				6 years	100	---	---	----------	0.91
				7 years	100	---	---	----------	0.91
				8 years	100	---	---	----------	0.95
				9 years	100	---	---	----------	0.96
				10 years	100	---	---	----------	0.93
				11 years	100	---	---	----------	0.95
				12 years	100	---	---	----------	0.92
				13 years	100	---	---	----------	0.92
				14 years	100	---	---	----------	0.94
				15-16 years	100	---	---	----------	0.84
McCarthy (526)--	1944	Goodenough-----	Normals, Grades 3 and 4.	N.R.	386	---	---	----------	----------
								Intrascorer----------	0.94
								Interscorer----------	0.90
								Test-retest----------	0.68
								Odd-even, Spearman-Brown.	0.89
McHugh (529)----	1952	Goodenough-----	Normals, Grade 3.	N.R.	118	58	60	----------	----------
								Intrajudge----------	0.98
								Interjudge----------	0.97
Stone (582)-----	1952	Machover-------	Normals, Grade 6.	N.R.	492	---	---	----------	----------
								Split-half	
								First drawing---------	0.82
								Second drawing--------	0.76
								Test-retest	
								Drawings 1 and 2, males----------	0.56
								Drawings 1 and 2, females----------	0.39
								Drawings 1 and 2, total----------	0.50

[a]Designations of subjects are always white Americans unless otherwise specified.
[b]Indicates conditions preceding Draw-A-Man testing.

Group	Initial test	Second test	Third test
A	Satisfying activity	Satisfying activity	Frustrating activity
B	Frustrating activity	Frustrating activity	Satisfying activity

NOTES: Unless otherwise indicated, it is assumed that reliability coefficients were Pearson Product-Moment and were computed from raw scores.

Σ —Total population; M—male; F—female; N.R.—not reported; IQ—intelligence quotient; MA—mental age.

Table 7. Studies reporting correlations between the Goodenough and Stanford-Binet

Investigator	Year	Subjects[a]	Age range	Number			Correlations	
				Σ	M	F	IQ	MA
McElwee (524)	1932	Retarded	14 years	45	---	---	N.R.	0.72
Rohrs and Haworth (569)	1962	Retarded		46	23	23	0.28	N.R. (Form L-M)
		Familial	12.57 years (mean)	20	10	10	N.R.	N.R.
		Organic	9.2 years (mean)	26	13	13	N.R.	N.R.
Birch (550)	1949	Retarded	10-6 - 16-3	68	43	25	0.62	0.69
Israelite (562)	1936	Feebleminded	6-3 - 40 years	256	162	94	N.R.	0.71
Johnson, Ellerd, and Lahey (592)	1950	State hospital population	6-9 - 17 years	209	---	---	0.48	N.R.
White (565)	1945			141	---	---		
		Feebleminded	8-0 - 19-4	47	---	---	0.63	N.R.
		Epileptic	8-0 - 19-4	47	---	---	0.52	N.R.
		Normal	4-8 - 10-6	47	---	---	0.71	N.R.
Havighurst and Janke (544)	1944	Normals	10 years	114	---	---	0.50	N.R.
Fowler (531)	1953	Normals	9-2 - 12-1	41	19	22	0.41	N.R.
Lessing (551)	1961	Normals	8-9 years	23	21	2	0.51	N.R.
McHugh (549)	1945	Normals	64 months (mean)	90	43	47	0.41	0.45
Thompson and Finley (552)	1963	Guidance clinic referrals	5-9 years	164	81	83	0.67	N.R. (Form L-M)
Goodenough (504)	1926	Normals	4-12 years	5,627	---	---	0.74	0.76
Williams (505)	1935	Normals	3-15 years	100	50	50	0.65	0.80

[a]Designations of subjects are always white Americans unless otherwise specified.

NOTES: Unless otherwise indicated all correlations are Pearson Product-Moment, with the Stanford-Binet, Form L.
Σ —Total population; M—male; F—female; IQ—intelligence quotient; MA—mental age; N.R.—not reported.

scores largely reflect the ability to form concepts—is supported by the network of correlations compiled from a variety of tests and by studies such as that of McHugh (549), which analyzed Draw-A-Man items. McHugh computed biserial correlations of Goodenough items with the Stanford-Binet and reported positive correlations for 29 items; the remainder were zero or slightly negative. The highest correlations, which support the conceptual interpretation stated, were the following:

	Item	Correlation
2	(legs present)	0.48
7a	(eyes present)	0.47
9a	(clothing present)	0.40
11b	(leg joint shown)	0.35
12e	(proportion, two dimensions)	0.54
13	(heel shown)	0.35

Table 8. Studies reporting correlations between the Goodenough and other measures

Investigator	Year	Test or criterion variable	Subjects[a]	Age range	Number			Correlation
					Σ	M	F	
Havighurst, Gunther, and Pratt (558).	1946	Arthur Point Scale of Performance Tests (IQ).	American Indians.	6-11 years	294	---	---	---
			Zuni		42	---	---	0.10
			Hopi		78	---	---	0.21
			Navaho		47	---	---	0.23
			Sioux		53	---	---	0.33
			Papago		74	---	---	0.64
Albee and Hamlin (579)	1949	Clinical ratings of adjustments	VA Mental Hygiene Clinic. Range—normals to psychotics.	N.R.	N.R.	---	---	0.62 (rank order) 0.64 (product moment)
Havighurst and Janke (544).	1944	Cornell-Coxe Performance Ability Scale.	Normals	10 years	114	---	---	0.63
Havighurst, Gunther, and Pratt (558).	1946	Cornell-Coxe Performance Ability Scale.	Normals	6-11 years	66	28	38	0.63
Hinrichs (586)	1935	Furfey Revised Scale for Measuring Developmental Age in Boys.	Delinquents	9-18 years	425	---	---	0.35
Johnson (557)	1953	Hoffman Bilingual Schedule	Spanish bilinguals (U.S.).	N.R.	30	---	---	0.05
Boehncke (546)	1938	Leiter International Performance Scale.	Normals	5-12 years	257	---	---	0.83
Ansbacher (553)	1952	MacQuarrie Test for Mechanical Ability.	Normals	10 years	100	---	---	---
		Tracing						0.34
		Tapping						0.23
		Dotting						0.16
Brenner and Morse (517)	1956	Metropolitan Readiness Tests, Number Readiness (IQ).	Normals	4-7 - 5-11	16	7	9	0.58 (rank order)
Havighurst and Janke (544).	1944	Revised Minnesota Paper Form Board Test, Form AR.	Normals	10 years	110	---	---	0.48
Brenner and Morse (517)	1956	Monroe Visual subtest (IQ)	Normals	4-7 - 5-11	16	7	9	0.64 (rank order)
Hornowski (547)	1961	Moray House Picture Intelligence Test.	Normals (Scotland).	N.R.	N.R.	---	---	0.34 (M) 0.49 (F)
Johnson (557)	1953	Otis Self-Administering Tests of Mental Ability.	Spanish bilinguals (U.S.).	N.R.	30	---	---	-0.02
Brenner and Morse (517)	1956	Picture Judgment of Maturity (IQ)	Normals	4-7 - 5-11	16	7	9	0.64 (rank order)
		Pintner-Cunningham Primary Mental Test (MA).						0.66 (rank order)
Shirley and Goodenough (575).	1932	Pintner Non-Language Primary Mental Test (IQ).	Deaf	5+ years	229	---	---	0.33
Norman and Midkiff (559)	1955	Progressive Matrices	Normals, American Indian.	6-6 - 15-6	96	---	---	0.24 (IQ) 0.35 (MA)
Harris (548)	1959	Progressive Matrices	Normals	5-1 - 6-1	98	45	53	0.22
Johnson (557)	1953	Reaction time	Spanish bilinguals (U.S.).	N.R.	30	---	---	0.43
Brenner and Morse (517)	1956	Sangren Information Mental Age	Normals	4-7 - 5-11	16	7	9	0.67 (rank order)

See footnotes at end of table.

Table 8. Studies reporting correlations between the Goodenough and other measures—Con.

Investigator	Year	Test or criterion variable	Subjects[a]	Age range	Number			Correlation
					Σ	M	F	
Buhrer, de Navarro, and Velasco (511).	1951	School grades	Normals, Spanish-speaking.	7-14 years	1,936	---	---	
		Mathematics						-0.04
		Language						-0.10
		Language and Mathematics						-0.01
		Drawing						0.27
Fowler (531)	1953	Social Distance Scale (Fowler)	Normals	9-2 - 12-1	41	19	22	0.40
Shirley and Goodenough (575).	1932	Stanford Achievement, Education (quotient).	Deaf	5+ years	41	---	---	0.34
Ansbacher (553)	1952	SRA Primary Mental Abilities	Normals	10 years	100	---	---	
		Word Vocabulary						0.23
		Picture Vocabulary						0.19
		Total Verbal Meaning						0.26
		Space						0.38
		Word Grouping						0.28
		Figure Grouping						0.34
		Total Reasoning						0.40
		Perception						0.37
		Number						0.24
		Total Nonreading						0.45
		Total Score						0.41
		S+R+P						0.48
Harris (548)	1959	SRA Primary Mental Abilities	Normals	5-1 - 6-1	98	45	53	
		Verbal						0.50
		Perception						0.44
		Quantitative						0.54
		Motor						0.40
		Space						0.51
Brenner and Morse (517)	1956	Teacher rank of school readiness	Normals	4-7 - 5-11	16	7	9	0.69 (rho)
Britton (536)	1954	Warner's Index of Status Characteristics.	Normals	9-11 years	232	102	130	0.11
Hanvik (593)	1953	WISC Full Scale (IQ)	Psychiatric patients.	5-12 years	25	---	---	0.18 (rank order)
Rohrs and Haworth (569)	1962	Wechsler Intelligence Scale for Children (IQ).	Retarded, familial and organic.	N.R.	46	23	23	
		Verbal Scale						0.28
		Performance Scale						0.53
		Full Scale						0.46

[a]Designations of subjects are always white Americans unless otherwise specified.

NOTES: All correlation coefficients are Pearson Product-Moment unless otherwise specified.

Σ —Total population; M—male; F—female; IQ—intelligence quotient; N.R.—not reported; MA—mental age.

It is of interest that a careful survey of the literature spanning a period of over 40 years fails to disclose any definitive pattern of the particular components of mental maturity measured by the Goodenough test. Harris believes that this may be attributed to the fact that such components are themselves not clearly differentiated in young children. The correlational results do, however, suggest strongly that the Draw-A-Man is more highly associated with factors measured by performance tests than with verbal abilities.

In the Health Examination Survey, correlations of the Draw-A-Man with WISC and, more particularly, with the short form composed of WISC Vocabulary and Block Design would be most relevant. Table 3 includes three reports (115, 130, and 224) which mention correlations between the Draw-A-Man Test and the Full Scale IQ of the WISC. Of these, none mentions correlations between the Draw-A-Man and the short form of the WISC. Harris' summary also cites the following unpublished data by Ellis.

Age	Number	Correlation with:		
		FS	VS	PS
8 years-------	16	0.70	0.77	0.67
9 years-------	34	0.67	0.63	0.59
10 years-------	20	0.24	0.17	0.26
11 years-------	17	0.50	0.45	0.46
12 years-------	19	0.62	0.50	0.68
13 years-------	17	0.13	0.05	0.15

Disregarding the 13-year-old group, since it is outside the effective range of the test as well as outside the age range of the Survey, Ellis' results for the total sample of 106 have an average correlation with the WISC Full Scale IQ of 0.57. Again, this is higher than the correlations reported by others.

In summary, it appears that the WISC correlations with the Draw-A-Man Test are substantial but lower than those of the Stanford-Binet. They are, however, higher with the Performance Scale than with the Verbal Scale (except in Ellis' two lowest grades).

In comparing Draw-A-Man scores with WISC Full Scale estimates, there is no reason to assume any systematic differences in mean levels across the entire population. However, for statistical estimation as well as analytic purposes, it is most appropriate to compute the regression of Draw-A-Man on Voc., BD, and Total Score and then to work with differences between regressed and actual scores for discrepancy analysis, rather than with differences between scaled scores.

In view of the Draw-A-Man's sensitivity to cultural variations, cases in which there are large discrepancies between the Draw-A-Man and the WISC should be thoroughly evaluated in the light of the WRAT scores and other information from the Health Examination Survey. Although Harris' summary and the reports consulted in this review have suggested a number of promising diagnostic score patterns, none of them seem well enough established to be adopted.

THE HARRIS REVISION OF THE GOODENOUGH TEST

Dale Harris' 1963 publication (522), which he has named the Goodenough-Harris Drawing Test, is a thorough revision and extension of Goodenough's test. As already mentioned, it bases the lengthier point-score scales on both drawings of the male figure and drawings of the female figure, for which it provides separate norms for boys and for girls. A third picture, in which the child draws a representation of himself, has not been empirically standardized.

Standardization of the Harris revision was completed on a total sample of 2,965 children, representative of four major geographic areas of the country. The sample was also representative of the 1960 census distribution of fathers' occupations. Total point scores are converted to standard scores with a mean of 100 and a standard deviation of 15. Conceptually, these are equivalent to the WISC deviation IQ's. The new scales overlap extensively with the original point scales, and Harris found that children now earn substantially

higher scores when the 1963 norms, rather than the 1926 ones, are utilized. The explanation for this phenomenon is not clear. The new norms do appear to take into account technical and social changes which have occurred between 1926 and 1963. They also offer the advantages of greater length (hence, higher reliability) and more adequate provision for sex differences.

Comparison of Goodenough and Goodenough-Harris Scores

It seems desirable to inquire whether the Harris scales and norms could be used to score human figure drawing obtained in the Health Examination Survey. As noted above, in this Survey only one picture is drawn by each child, who is instructed, "Make a picture of a person. Make the very best person that you can." To use the Harris scales in the Survey it would be necessary for the scorer to decide whether each drawing was of a "Man" or of a "Woman."

A sample of 200 drawings, 100 drawn by boys and the other 100 drawn by girls, was taken at random from the Survey files. These drawings were then carefully scored using Harris' norms, and the scores obtained were compared with the scores the drawings had already received on the 1926 Goodenough scale. (Scoring by the 1926 method is completed in the field by Survey staff psychologists.)

Of the 200 cases, 195 were usable. Three drawings were rejected because they contained a face only, and for two cases age had been inadvertently omitted, precluding the computation of standard scores. For the remaining drawings, neither scorer reported any difficulty in identifying the sex represented, and their agreement on this was perfect.

Table 9. Means of Goodenough-Harris and Goodenough variables and correlations between scorers and between methods for total sample and six subsamples

Variable	Total group	Drawings of a woman	Drawings of a man	Drawings of a woman		Drawings of a man	
				By boys	By girls	By boys	By girls
	N=195	N=94	N=101	N=17	N=77	N=83	N=18
1. Goodenough-Harris point (A)	30.75	31.41	30.13	28.12	32.14	30.20	29.78
2. Goodenough-Harris SS (A)	96.59	95.89	97.24	93.06	96.52	97.29	97.00
3. Goodenough-Harris point (B)	36.02	36.62	35.47	34.71	37.04	35.54	35.11
4. Goodenough-Harris SS (B)	105.97	105.15	106.73	104.06	105.39	106.63	107.22
1+3. Average Goodenough-Harris point (A,B)	33.39	34.02	32.80	31.42	34.59	32.87	32.45
2+4. Average Goodenough-Harris SS (A,B)	101.28	100.52	101.99	98.56	100.96	101.96	102.11
5. Goodenough point	26.38	25.57	27.14	24.29	25.86	27.20	26.83
6. Subject's CA	115.01	111.89	117.92	118.35	110.47	118.10	117.11
7. Goodenough MA	114.61	112.48	116.59	108.88	113.27	116.71	116.06
8. Goodenough IQ	101.23	102.27	100.27	92.59	104.42	100.10	101.06
r_{13}	0.90	0.89	0.91	0.82	0.91	0.90	0.95
r_{24}	0.90	0.88	0.91	0.79	0.89	0.92	0.83
r_{28}	0.78	0.76	0.81	0.60	0.78	0.87	0.47
r_{48}	0.81	0.78	0.84	0.58	0.82	0.89	0.48

NOTE: N—number; A—scorer A; B—scorer B; SS—standard score; CA—chronological age; MA—mental age; r—correlation.

The usable sample of 195 cases consisted of 100 boys and 95 girls. Of these, 17 boys drew a Woman figure and 18 girls drew a Man figure. The remaining 82 percent of the total group (83 percent of the boys and 81 percent of the girls) drew their own sex.

The following eight variables were recorded for all 195 cases:

1. Harris method, point score, scorer A
2. Harris method, standard score, scorer A
3. Harris method, point score, scorer B
4. Harris method, standard score, scorer B
5. Goodenough point score
6. Subject's chronological age in months
7. Goodenough mental age
8. Goodenough IQ

Means, standard deviations, and intercorrelations were computed for the total sample and for the following six subsamples: (1) Woman drawings (N=94), (2) Man drawings (N=101), (3) Woman drawings by boys (N=17), (4) Woman drawings by girls (N=77), (5) Man drawings by boys (N=83), and (6) Man drawings by girls (N=18). A summary of the most relevant results, for all seven sample combinations, appears in table 9.

The correlations between the two scorers (r_{13} and r_{24}) are high despite a systematic tendency for scorer B's results to exceed those of scorer A (they average 5.25 above scorer A on point score and 9.38 higher on standard score). As a more stable estimate of the Harris scores for comparison with the Goodenough, average mean scores for the two scorers were computed. These appear in table 9 between variables 4 and 5.

Although agreement between the two scorers is generally high, the lowest correlations were found for the 17 boys who elected to draw a female figure (subsample 3). The standard score correlations for the 18 girls who elected to draw a male figure (subsample 6) are also comparatively low. These opposite-sex drawings also reflect the lowest correlations between Harris and Goodenough IQ's for both scorers (r_{28} and r_{48}). Thus scorer agreement is lowest on opposite-sex drawings, and the results for these show the poorest agreement, correlationwise, between the Goodenough-Harris and Goodenough IQ's. It is possible that these differences could be eliminated by further training of scorers. Certainly these results illustrate the importance of quality control of scoring. The averaging process is also highly recommended if systematic scorer differences cannot be eliminated.

The principal support, indicating an advantage of the Goodenough-Harris scale, appears in the comparison of mean scores for boys and girls on Woman and Man drawings as abstracted in table 10. In accordance with Harris' own findings, girls score higher than boys, but the differences are greater on the Goodenough scale than on the Goodenough-Harris scales and are greater on the Woman drawings than on the Man drawings. The greatest discrepancy and resulting scoring penalty by the Goodenough scale occurs in the case of the 17 percent of boys (subsample 3) who elected to draw a Woman. At the same time, the 81 percent of girls (subsample 4) who elected to draw their own sex received disproportionately high scores on the Goodenough, in comparison with the mean levels on the Goodenough-Harris. The Goodenough-Harris scores are higher than the Goodenough for both sexes on the Man drawing.

The problems with the Woman drawing clearly support the observation, first pointed out by Goodenough and strongly reiterated by Harris, that the female figure is more culture-bound than the male, is less stereotyped, and is more susceptible to individual interpretation. Although the data on which the present analysis is based are limited, they do suggest that the Harris revision does less violence to the female figure than does the Goodenough scoring and that, in general, the Harris revision is more adequate for opposite-sex drawings.

These data, which indicate a superiority of girls over boys in drawing scores, a tendency for the Goodenough-Harris scores to be higher than the Goodenough scores, and a tendency for girls who draw male figures to be older than girls who draw their own sex (while no such differentiation occurs among boys), are all consistent with trends reported elsewhere in the literature. However, the most important argument in favor of using the Goodenough-Harris scoring system is that the variation of mean scores among the four subsamples is thereby greatly reduced around a mean of 100. This range is from 92.59 to 104.42 (11.83) on the Goodenough and from 98.56 to 102.11 (3.55) on the Goodenough-Harris. Although the

Table 10. Comparison of Goodenough-Harris and Goodenough mean IQ's for boys and girls on same-sex and opposite-sex drawings

Sex	Drawing of a woman			Drawing of a man		
	Goodenough IQ	Goodenough-Harris IQ	Difference	Goodenough IQ	Goodenough-Harris IQ	Difference
Boys-------	92.59	98.56	+5.97	100.10	101.96	+1.86
Girls------	104.42	100.96	-3.46	101.06	102.11	+1.05
Difference-	11.83	2.40	----------	0.96	0.15	----------

Table 11. Coefficients of variation of Harris and Goodenough IQ's for total sample and six subsamples

Item	Total group	Drawings of a woman	Drawings of a man	Drawings of a woman		Drawings of a man	
				By boys	By girls	By boys	By girls
Harris standard score----------	0.16	0.15	0.15	0.10	0.16	0.17	0.13
Goodenough IQ----------	0.19	0.18	0.19	0.14	0.18	0.21	0.18

standard deviations of the Goodenough-Harris and Goodenough scores were not shown in table 9, the relative variability of scores based on the two systems is indicated in table 11, which reports coefficients of variation $\left(\frac{\text{standard deviation}}{\text{mean}}\right)$ for Goodenough-Harris standard scores and for Goodenough IQ's for each of the subsamples. It is apparent that in every case variance is lower for the Harris scores.

Recommendation

On the basis of this analysis it is recommended that the following steps be adopted in relation to the Draw-A-Man Test in the Survey: (1) the Goodenough-Harris system should be used; (2) the entire sample should be scored centrally by uniform standards, with adequate training of scorers and quality control procedures routinely followed; and (3) if scorer variations cannot be eliminated by training, the procedure of averaging the results of two or more scorers should be adopted.

SUMMARY AND CONCLUSIONS

The foregoing review of the Draw-A-Man Test supports the view that it is a reliable and valid nonlanguage measure of mental maturity, although highly sensitive to cultural influences on the child's conceptual representation of the human figure. Its use in a national survey in the 6 to 12 age range, in conjunction with the WISC and WRAT, is logical and desirable—particularly as a means of assessing intellectual development in cases in which there is impairment of verbal development or verbal performance.

Personality assessment by means of thematic and qualitative assessment of children's drawings would probably be unrewarding. Some indications justifying further research have been noted; however, such research is not sufficiently promising to warrant the expenditure of Survey funds. On the other hand, several lines of empirical work appear worthwhile. These are enumerated below.

As discussed in the final portion of the review of the Draw-A-Man, there is strong evidence for the adoption of the Harris revision of the Draw-A-Man with central scoring by trained scorers, and

averaging of scores of two or more scorers, if scorer variations cannot be eliminated in training. This procedure need not be regarded as expensive, since it could leave the field psychologists free to test more children while the scoring is done centrally by lower paid workers.

Although research on personality-assessment uses of the drawings within the Survey program is not recommended, the following lines of empirical study and analysis are regarded as useful and even important:

1. A systematic study of cultural variations related to the principal geographic areas in which Survey data were collected to evaluate the effects of factors such as customs, attitudes, dress, art, and social roles in relation to the items in the point scales by which the Draw-A-Man is scored. Even if the results of such an analytic study should be negative, they would be very reassuring in relation to the use of the Draw-A-Man scores in the Survey.
2. Regression studies of Draw-A-Man scores with other psychometric variables in the Survey so that comparisons can be made on the basis of differences between regressed and actual scores rather than directly between raw scores.
3. Further restandardization of the Goodenough-Harris norms on a national sample would be a valuable contribution to psychological measurement of children that could only reflect credit on the Survey and would be of major importance for future use of this well-established and useful intelligence test. This significant undertaking, if approved, should include a complete item analysis as well as recomputation of norms.

Some additional suggestions regarding cross-disciplinary studies with reference to the Draw-A-Man Test are presented in a later section of this report.

BIBLIOGRAPHY

General References to Draw-A-Man

501. Herrick, M. A.: Children's drawings. *Ped.Sem.* 3:338-339, 1893.
502. Barnes, E.: A study of children's drawings. *Ped.Sem.* 2:455-463, 1893.
503. Lukens, H.: A study of children's drawings in the early years. *Ped.Sem.* 4:79-110, 1896.
504. Goodenough, F. L.: A new approach to the measurement of the intelligence of young children. *J.Genet.Psychol.* 33:185-211, 1926.
505. Williams, J. H.: Validity and reliability of the Goodenough intelligence test. *Sch.&Soc.* 41:653-656, 1935.
506. Smith, F. O.: What the Goodenough intelligence test measures. *Psychological Bull.* 34:760-761, 1937.
507. Barnhart, E. N.: Developmental stages in compositional construction in children's drawings. *J.Exp.Educ.* 11:156-184, 1942.
508. McHugh, G.: Changes in Goodenough IQ at the public school kindergarten level. *J.Educ.Psychol.* 36:17-30, 1945.
509. Lehner, G. F. J., and Silver, H.: Some relations between own age and ages assigned on the Draw-a-Person Test; abstracted, *Am.Psychologist* 3:341, 1948.
510. Goodenough, F. L., and Harris, D. B.: Studies in the psychology of children's drawings, II, 1928-1949. *Psychological Bull.* 47:369-433, 1950.
511. Buhrer, L., de Navarro, R., and Velasco, E. S.: Ensayo de tipificacion de la prueba mental "Dibujo de un hombre" de F. Goodenough. *Publ.Inst.Biotipol.Exp.U.Cuyo,* 2:113, 1951.
512. Weider, A., and Noller, P. A.: Objective studies of children's drawings of human figures, II, Sex, age, intelligence. *J.Clin.Psychol.* 9:20-23, 1953.
513. Tuska, S. A.: *Developmental Concepts With the Draw-a-Person Test at Different Grade Levels.* Unpublished master's thesis, Ohio University, 1953.
514. Stewart, N.: Review of Goodenough Draw-A-Man Test, in O.K. Buros, ed., *The Fourth Mental Measurements Yearbook.* Highland Park, N.J. The Gryphon Press, 1953.
515. Woods, W. A., and Cook, W. E.: Proficiency in drawing and placement of hands in drawings of the human figure. *J.Consult.Psychol.* 18:119-121, 1954.
516. Bliss, M., and Berger, A.: Measurement of mental age as indicated by the male figure drawings of the mentally subnormal using Goodenough and Machover instructions. *Am.J.Ment.Deficiency* 59:73-79, 1954.
517. Brenner, A., and Morse, N. C.: The measurement of children's readiness for school. *Pap.Mich.Acad.Sci.* 41: 333-340, 1956.
518. Frankiel, R. V.: *A Quality Scale for the Goodenough Draw-a-Man Test.* Unpublished master's thesis, University of Minnesota, 1957.

519. Zuk, G. H.: Children's spontaneous object elaborations on a visual-motor test. *J.Clin.Psychol.* 16:280-283, 1960.

520. Stoltz, R. E., and Coltharp, F. C.: Clinical judgments and the Draw-A-Person Test. *J.Consult.Psychol.* 25: 43-45, 1961.

521. Zuk, G. H.: Relation of mental age to size of figure on the Draw-A-Person Test. *Percept.Mot.Skills* 14:410, 1962.

522. Harris, D. B.: *Children's Drawings as Measures of Intellectual Maturity*. New York. Harcourt, Brace & World, Inc., 1963.

Goodenough: Reliability Studies

523. Yepsen, L. N.: The reliability of the Goodenough drawing test with feebleminded subjects, *J.Educ.Psychol.* 20:448-451, 1929.

524. McElwee, E. W.: The reliability of the Goodenough intelligence test used with sub-normal children fourteen years of age. *J.Appl.Psychol.* 16:217-218, 1932.

525. Brill, M.: The reliability of the Goodenough Draw-a-Man Test and the validity and reliability of an abbreviated scoring method. *J.Educ.Psychol.* 26:701-708, 1935.

526. McCarthy, D.: A study of the reliability of the Goodenough Drawing Test of Intelligence. *J.Psychol.* 18: 201-206, 1944.

527. McCurdy, H. G.: Group and individual variability on the Goodenough Draw-A-Man Test. *J.Educ.Psychol.* 38:428-436, 1947.

528. Harris, D. B.: Intra-individual vs. inter-individual consistency in children's drawings of a man; abstracted, *Am.Psychologist* 5:293, 1950.

Goodenough: Factors Affecting Drawing Productions

529. McHugh, A. F.: *The Effect of Preceding Affective States on the Goodenough Draw-A-Man Test of Intelligence*. Unpublished master's thesis, Fordham University, 1952.

530. Reichenberg-Hackett, W.: Changes in Goodenough Drawings after a gratifying experience. *Am.J.Orthopsychiat.* 23:501-517, 1953.

531. Fowler, R. D.: *The Relationship of Social Acceptance to Discrepancies Between the IQ Scores on the Stanford-Binet Intelligence Scale and the Goodenough Draw-a-Man Test*. Unpublished master's thesis, University of Alabama, 1953.

532. Herron, W. G.: *The Effect of Preceding Affective States on the Goodenough Drawing Test of Intelligence*. Unpublished master's thesis, Fordham University, 1957.

533. Koppitz, E. M.: Teacher's attitude and children's performance on the Bender-Gestalt Test and Human Figure Drawings. *J.Clin.Psychol.* 16:204-208, 1960.

534. Richey, M. H., and Spotts, J. V.: The relationship of popularity to performance on the Goodenough Draw-A-Man Test. *J.Consult.Psychol.* 23:147-150, 1959.

535. Tolor, A.: Teachers' judgments of the popularity of children from their human figure drawings. *J.Clin.Psychol.* 11:158-162, 1955.

536. Britton, J. H.: Influence of social class upon performance on the Draw-A-Man Test. *J.Educ.Psychol.* 45:44-51, 1954.

Goodenough: Body Image, Sexual Identification

537. Weider, A., and Noller, P. A.: Objective studies of children's drawings of human figures. I, Sex awareness and socioeconomic level. *J.Clin.Psychol.* 6:319-325, 1950.

538. Knopf, I. J., and Richards, T. W.: The child's differentiation of sex as reflected in drawings of the human figure. *J.Genet.Psychol.* 81:99-112, 1952.

539. Swenson, C. H., and Newton, K. R.: The development of sexual differentiation on the Draw-a-Person Test. *J. Clin.Psychol.* 11:417-419, 1955.

540. Lakin, M.: Certain formal characteristics of human figure drawings by institutionalized aged and by normal children. *J.Consult.Psychol.* 20:471-474, 1956.

541. Brown, D. G., and Tolor, A.: Human figure drawings as indicators of sexual identification and inversion. *Percept.Mot. Skills* 7:199-211, 1957.

542. Fisher, G. M.: Sexual identification in mentally subnormal females. *Am.J.Ment.Deficiency* 66:266-269, 1961.

543. Silverstein, A. B., and Robinson, H. A.: The representation of physique in children's figure drawing. *J.Consult.Psychol.* 25:146-148, 1961.

Goodenough: Relation to Other Tests

544. Havighurst, R. J., and Janke, L. L.: Relations between ability and social status in a midwestern community. I, Ten-year-old children. *J.Educ.Psychol.* 35:357-368, 1944.

545. Condell, J. F.: Note on the use of the Ammons Full-Range Picture Vocabulary Test with retarded children. *Psychol.Rep.* 5:150, 1959.

546. Boehncke, C. F.: *A Comparative Study of the Goodenough Drawing Test and the Leiter International Performance Scale*. Unpublished master's thesis, University of Southern California, 1938.

547. Hornowski, B.: Interpretation psychologique des differences entre sexes dans le dessin due bonhomme chez les jeunes adolescents (Psychological interpretation of sex differences in the Draw-a-Man Test among young adolescents). *Revue Psychol.Appl.* 11:7-9, 1961.

548. Harris, D. B.: A note on some ability correlates of the Raven Progressive Matrices (1947) in the kindergarten. *J.Educ.Psychol.* 50:228-229, 1959.

549. McHugh, G.: Relationship between the Goodenough Draw-a-Man Test and the 1937 revision of the Stanford-Binet Test. *J.Educ.Psychol.* 36:119-124, 1945.

550. Birch, J. W.: The Goodenough Drawing Test and older mentally retarded children. *Am.J.Ment.Deficiency* 54: 218-224, 1949.

551. Lessing, E. E.: A note on the significance of discrepancies between Goodenough and Binet IQ scores. *J. Consult.Psychol.* 25:456-457, 1961.

552. Thompson, J. M., and Finley, C. J.: The relationship between the Goodenough Draw-a-Man Test and the Stanford-Binet Form L-M in children referred for school guidance services. *Calif.J.Educ.Res.* 14:19-22, 1963.

553. Ansbacher, H. L.: The Goodenough Draw-A.Man Test and primary mental abilities. *J.Consult.Psychol.* 16: 176-180, 1952.

Goodenough: Cultural Variations, Bilingualism

554. Hunkin, V.: Validation of the Goodenough Draw-a-Man Test for African children. *J.Soc.Res.Pretoria* 1:52-63, 1950.

555. Dennis, W.: Performance of Near Eastern children on the Draw-a-Man Test. *Child Development* 28:427-430, 1957.

556. Anastasi, A., and DeJesus, C.: Language development and nonverbal IQ of Puerto Rican preschool children in New York City. *J.Abnorm.&Social Psychol.* 48:357-366, 1953.

557. Johnson, G. B., Jr.: Bilingualism as measured by a reaction-time technique and the relationship between a language and a non-language intelligence quotient. *J. Genet.Psychol.* 82:3-9, 1953.

558. Havighurst, R. J., Gunther, M. X., and Pratt, I. E.: Environment and the Draw-A-Man Test, the performance of Indian children. *J.Abnorm.&Social Psychol.* 41:50-63, 1946.

559. Norman, R. D., and Midkiff, K. L.: Navaho children on Raven Progressive Matrices and Goodenough Draw-A-Man Tests. *SWest.J.Anthrop.* 11:129-136, 1955.

560. Carney, R. E., and Trowbridge, N.: Intelligence test performance of Indian children as a function of type of test and age. *Percept.Mot. Skills* 14:511-514, 1962.

Goodenough: With Subnormal, Retarded, and Mentally Defective Children

561. McElwee, E. W.: Profile drawings of normal and subnormal children. *J.Appl.Psychol.* 18:599-603, 1934.

562. Israelite, J.: A comparison of the difficulty of items for intellectually normal children and mental defectives on the Goodenough drawing test. *Am.J.Orthopsychiat.* 6: 494-503, 1936.

563. Spoerl, D. T.: Personality and drawing in retarded children. *Character. Pers.* 8:227-239, 1940.

564. Spoerl, D. T.: The drawing ability of mentally retarded children. *J.Genet.Psychol.* 57:259-277, 1940.

565. White, M. R.: *The Performance of Epileptic, Feebleminded and Normal Children on the Goodenough Test of Intelligence.* Unpublished master's thesis, State University of Iowa, 1945.

566. Gunzburg, H. C.: The significance of various aspects in drawings by educationally subnormal children. *J.Ment.. Sc.* 96:951-975, 1950.

567. Fabian, A. A.: Clinical and experimental studies of school children who are retarded in reading. *Quart.J. Child Behav.* 3:15-18, 1951.

568. Hunt, B., and Patterson, R. M.: Performance of familial mentally deficient children in response to motivation on the Goodenough Draw-A-Man Test. *Am.J.Ment.Deficiency* 62:326-329, 1957.

569. Rohrs, F. W., and Haworth, M. R.: The 1960 Stanford-Binet, WISC, and Goodenough Tests with mentally retarded children. *Am.J.Ment.Deficiency* 66:853-859, 1962.

Goodenough: Chronic Encephalitis

570. Bender, L.: The Goodenough Test (Drawing a Man) in chronic encephalitis in children. *J.Child Psychiat.* 3: 449-459, 1951.

Goodenough: Physically Handicapped

571. Martorana, A. A.: *A Comparison of the Personal, Emotional, and Family Adjustment of Crippled and Normal Children.* Unpublished doctoral dissertation, University of Minnesota, 1954.

572. Silverstein, A. B., and Robinson, H. A.: The representation of orthopedic disability in children's figure drawings. *J.Consult.Psychol.* 20:333-341, 1956.

573. Johnson, O. G., and Wawrzasek, F.: Psychologists' judgments of physical handicap from H-T-P drawings. *J.Consult.Psychol.* 25:284-287, 1961.

Goodenough: Intelligence of Deaf Children

574. Peterson, E. G., and Williams, J. M.: Intelligence of deaf children as measured by drawings. *Am.Ann.Deaf* 75:273-290, 1930.

575. Shirley, M., and Goodenough, F. L.: A survey of the intelligence of deaf children in Minnesota schools. *Am. Ann.Deaf* 77:238-247, 1932.

576. Springer, N. N.: A comparative study of the intelligence of deaf and hearing children. *Am.Ann.Deaf* 83:138-152, 1938.

Goodenough: Measurement of Adjustment

577. Brill, M.: A study of instability using the Goodenough drawing scale. *J.Abnorm.&Social Psychol.* 32:288-302, 1937.

578. Springer, N. N.: A study of drawings of maladjusted and adjusted children. *J.Genet.Psychol.* 58:131-138, 1941.

579. Albee, G. W., and Hamlin, R. M.: An investigation of the reliability and validity of judgments of adjustment inferred from drawings. *J.Clin.Psychol.* 5:389-392, 1949.

580. Ochs, E.: Changes in Goodenough drawings associated with changes in social adjustment. *J.Clin.Psychol.* 3: 282-284, 1950.

581. Albee, G. W., and Hamlin, R. M.: Judgment of adjustment from drawings; the applicability of rating scale methods. *J.Clin.Psychol.* 6:363-365, 1950.

582. Stone, P. M.: *A Study of Objectively Scored Drawings of Human Figures in Relation to the Emotional Adjustment of 6th Grade Pupils.* Unpublished doctoral dissertation, Yeshiva University, 1952.

583. Palmer, H. R.: *The Relationship of Differences Between Stanford-Binet and Goodenough IQ's to Personal Adjustment as Indicated by the California Test of Personality.* Unpublished master's thesis, University of Alabama, 1953.

584. Popplestone, J. A.: *Male Human Figure Drawing in Normal and Emotionally Disturbed Children.* Unpublished doctoral dissertation, Washington University, 1958.

585. Feldman, M. J., and Hunt, R. G.: The relation of difficulty in drawing to ratings of adjustment based on human figure drawings. *J.Consult.Psychol.* 22:217-219, 1958.

Goodenough: With Delinquents

586. Hinrichs, W. E.: The Goodenough drawing test in relation to delinquency and problem behavior. *Archs.Psychol., N.Y.* No. 175, 1935.

587. Starke, P.: *An Attempt To Differentiate Delinquents From Non-delinquents by Tests of Dominance Behavior, Dominance Feeling and the Goodenough Drawing of a Man.* Unpublished master's thesis, University of Minnesota. 1950.

Goodenough: With Disturbed Persons

588. Berrien, F. K.: A study of the drawings of abnormal children. *J.Educ.Psychol.* 26:143-150, 1935.

589. Despert, J. L.: *Emotional Problems in Children.* Utica. State Hospitals Press, 1938.

590. Des Lauriers, A., and Halpern, F.: Psychological tests in childhood schizophrenia. *Am.J.Orthopsychiat.* 17:57-67, 1947.

591. Holzberg, J. D., and Wexler, M.: The validity of human form drawings as a measure of personality deviation. *J. Project.Tech.* 14:343-361, 1950.

592. Johnson, A. P., Ellerd, A. A., and Lahey, T. H.: The Goodenough Test as an aid to interpretation of children's school behavior. *Am.J.Ment.Deficiency* 54:516-520, 1950.

593. Hanvik, L. J.: The Goodenough Test as a measure of intelligence in child psychiatric patients. *J.Clin.Psychol.* 9:71-72, 1953.

Goodenough: Other References Cited in Text

594. Buck, J. N.: The H-T-P technique, a qualitative and quantitative scoring manual. *J.Clin.Psychol.* 4:317-396, 1948.

595. Goodenough, F.: *Measurement of Intelligence by Drawings.* New York. Harcourt, Brace and World, Inc., 1926.

596. Machover, K.: *Personality Projection in the Drawing of the Human Figure.* Springfield, Ill. Charles C. Thomas, 1949.

IV. THE THEMATIC APPERCEPTION TEST

The technology of personality measurement lags far behind that of ability and achievement measurement. This lag makes it difficult for organizations (such as the Division of Health Examination Statistics) which seek to estimate population parameters on the basis of definitive test scores. At present there is not a single personality test for children that could be recommended without qualification. In view of the extensive use of personality tests in clinical practices and in school situations, this sweeping statement may appear extreme. It is, nevertheless, regrettably true. Perhaps clinical psychologists can justify their use of various personality measures on the basis of intensive individual case study in which test responses and scores are interpreted, by the clinician, in relation to consistent patterns of performance in the context of a total life record. The clinician usually feels free to accept or disregard information in this frame of reference, and he often employs informal, unstandardized "tests" as well as published procedures without regard for formal considerations of reliability and validity. Furthermore, since clinical judgments are confined to individual cases, they are not subject to verification by the rules of evidence observed in scientific studies. Educators often justify their personality testing as contributing to research, which is important, and the only tenable position in the light of the facts.

In contrast with the clinical and research uses of personality measures, where legitimacy is not primarily a function of the proven adequacy of the measurement instruments employed, surveys such as this one (HES) operate under severe constraints.

The survey scientist must defend the validity and reliability of his instruments as well as the adequacy of his sampling design for the purposes of his survey; both considerations affect the validity of population estimates from sample data.

The choice of a personality measurement instrument for Cycle II must be considered in the context of the preceding discussion. Although the California Personality Test and Cattell's Junior Personality Quiz are, in the opinion of the writer, the most adequately documented of the currently published and objectively scored personality tests for children, neither meets the reliability and validity standards necessary for Survey use and neither is appropriate for the entire age range of 6 through 11 years. Apart from these, no available tests even approach the requirements of this Survey.

In the psychometric sense, the Thematic Apperception Test (TAT) is not a *test.* It is a *projective device* consisting of a series of ambiguous (unstructured) pictures individually presented to the subject (or patient), who is asked to imagine and relate a story. The rationale of the procedure is that people will seek to create structure when a stimulus situation is unstructured and that in doing so they will draw on their own experience, needs, attitudes, and values to provide the details. This process is viewed as a "projection" of inner processes on the unstructured stimulus.

The TAT was developed by Henry A. Murray of Harvard University in 1938 (788). At the same time he presented a report which outlined a motivational system of organismic *needs* and environmental *presses.* This report was highly influential and stimulated much research. Five years later (in 1943), the TAT pictures and a manual for their use were published (799).

From the objective scoring standpoint, it is necessary to recognize that all projective methods share a major problem, since in all of them the testing strategy depends on the process by which subjects add structure to ambiguous stimuli. Although this structuring process does involve projection, in the sense defined above, it also simultaneously involves other factors. Indeed, the structuring process may be as much a function of external, situational factors, to which the subject is responding, as of internal factors. How these various factors combine are only imperfectly understood in the scientific study of perception; they have not, to the writer's knowledge, been investigated in relation to the TAT pictures. In spite of these facts, for the past 60 or more years users of projective techniques have continued to assume that responses to various stimuli represent projection *only.*

Cattell (796) has suggested that "projective" tests (which he thinks should be called "misperception tests"), should employ stimuli of a much lower order of complexity than those of the TAT and the Rorschach inkblots in order to simplify interpretation. Technically this may be an improvement, as Cattell has shown in the misperception tests which he designed for his objective test batteries. In these tests the subject's latitude of response to a specific ambiguity (e.g., estimating the number of communist party members in the United States or the value of a college degree) is extremely limited. A similar conclusion is also implicit in the modifications of the TAT pictures made by McClelland (798) in his studies of motivation measurement in fantasy.

In a complex projective technique such as the TAT, the story produced by a subject may represent his response to the entire picture or only to certain parts of the stimulus picture. In addition, the story itself necessarily requires technical interpretation by the examiner to the extent that it employs idiosyncratic language, symbols, and ideation. Because of the freedom and informality of the method, which is deliberate (in order to avoid prompting or the addition of extraneous variance contributed by the examiner), it is virtually impossible to relate responses to specific internal and external cues or patterns of cues.

The very looseness of the interpretative procedure, in contrast to fixed scoring keys in the case of questionnaires (usually answered "yes," "no," or "?"), led George Kelly (797), in an *Annual Review* article, to observe that while in the case of questionnaires the subject tries to guess what the examiner is thinking, in projective techniques the examiner must guess what the subject is thinking. In either case, there is a good deal of guessing going on.

The TAT has some similarity to the Draw-A-Man Test in that the Draw-A-Man provides an

unstructured stimulus (the instruction to draw a person) and permits wide latitude of response structuring on the part of the subject. It is noteworthy that the Draw-A-Man has produced no acceptable schemes for personality interpretation. However, as pointed out in the discussion of the Draw-A-Man, the most promising results in personality, as well as in cognitive assessment, have been those employing detailed, objective techniques of scoring, such as the point scales.

The selection of five cards of the TAT for the Survey undoubtedly reflects (1) the *appraisal* of existing personality tests mentioned above, combined with (2) the *recognition* of apparent widespread acceptance of the TAT as a projective technique and (3) the *belief* that an appropriate method of objective scoring of responses to them can be developed for the specific use of the Survey as well as for later more general use by professional workers. The basis for this appraisal cannot be documented here, although the writer is prepared to defend it. Reference to the forthcoming *Sixth Mental Measurements Yearbook* (O. Buros, ed., New Brunswick, N.J., The Gryphon Press) might be sufficient for this purpose. The evidence for the recognition of acceptance of the TAT is discussed below, together with an evaluation of the prospects for successful development of an objective scoring procedure.

REVIEW OF THE LITERATURE ON THE TAT

The present review includes abstracts of published research articles, theses, and critical reviews of the TAT literature, as well as 5 general references on the thematic apperception method. These constitute only a small portion of the extensive psychological, anthropological, and sociological research on the TAT and its variants which have appeared in undiminished quantity over the years (e.g., Thompson's Negro edition of the TAT, Symonds' Picture Story Test, Bellak's Children's Apperception Test (CAT), Van Lennep's Four Picture Test, Phillipson's Object Relations Technique, and numerous other techniques which can be traced to the Murray version). Both the TAT procedure and the Murray "need-press" concepts have been used extensively in personality studies and studies of motivation. The items selected for inclusion in this report were judged relevant if they (1) used a measurement approach, (2) were validation or normative studies, (3) had an applicable sample in terms of age, or (4) used an adequate scoring procedure.

Overview

Treatment of the TAT by different writers ranges from uncritical acceptance on the basis of a priori assumptions, illustrated by Henry (749) and Piotrowski (702), through qualified acceptance with a "soft" attitude toward the contradictory evidence, as demonstrated by Mayman (701) and Lindzey (703), to objective evaluation, illustrated by Eron (706), Windle (704), and others. Windle's comment, that there is little agreement among results reported by different investigators, seems to describe accurately this field of research. One area in which some agreement may be found, however, is that of cognitive evaluation (714 and 737-739); this is highly reminiscent of the Draw-A-Man.

The TAT literature abounds in elaborate but largely untested (critically, that is) scoring systems. Most of these are too extensive for brief summarization and go beyond the purposes of this report. However, they have been reviewed in anticipation of a further empirical study of the Survey's Thematic Apperception Test data, and references to 21 additional selected reports are included in the bibliography of section IV.

Most of these, as well as a number of other suggested analytic methods of scoring the TAT, are well summarized in a 1951 publication by Edwin S. Shneidman, Walther Joel, and Kenneth B. Little (800). Although the modes of analysis vary in detail and in terminology, the typical one involves interpretation and frequency counting or evaluation on a rating scale of all or part of the following types of information, usually across all of the stories obtained for a selection of cards. (The full series of cards is often abridged because of practical time limitations, as it is in the Survey.)

Formal (structural) aspects of the stories

Compliance with instructions (including card rejection)

Consistency of stories

Length of stories; vocabulary level

Grammatical forms (nouns, pronouns, verbs, incomplete sentences)

Number and type of situations described

Number and type of characters included

Outcome of stories

Level of response (from description to imaginative interpretation)

Interpretive categories

Feelings, moods, worries, emotional tone

Needs expressed (or implied)

Conflict areas

Presses—physical, emotional, mental, economic, social, religious

Characters—strivings, attitudes, obstacles, barriers, traits, and roles of hero, major characters, and minor characters

Outcomes reflecting success, failure

Thematic content—family dynamics, inner adjustment, sexual adjustment, interpersonal relations, aggression (physical, nonphysical)

Developmental level in Freudian (psychosexual) context

Defense mechanisms utilized

Manner in which environment is assimilated

The number of variables enumerated under these categories is extensive (Murray's need-press system alone exceeds 83), and in most cases the variables require detailed, careful definition and intensive training of scorers. High reliabilities have often been achieved among scorers within a particular laboratory for a given period of tenure of the staff members involved, but these have not generally been maintained with staff changes or when systems have been tried out at other institutions. Often, definitions change over time as new generations of protocols appear, requiring decisions in relation to categories developed on the basis of earlier samples.

In spite of the logical (from some theoretical positions) appeal of these analytic approaches, they do not fit the requirements of psychometric procedures. Such analytic approaches satisfy the needs of various clinicians or investigators in their individual practices and researches, but for survey purposes they are useful primarily because they suggest areas which may be suitable for objective study. With the exception of some formal characteristics (such as length of story and other items that can be counted fairly accurately) which have been related to developmental rather than personality-adjustment concepts, there is so little agreement in the literature on most scoring categories that an investigator seeking to develop an objective scoring procedure might as well start from "scratch."

Research Demonstrating Developmental Factors

Edelstein (737) completed an interesting pilot study demonstrating a system for scoring TAT stories. From her system a total age-adjusted score, correlating well with Stanford-Binet IQ's, could be derived. She used the following six scoring categories—number of words, qualifier/word ratio, number of conditions, number of responses, number of situations involved, and number of characters. Her sample included only 15 boys and 13 girls (ages 9-5 to 12-5), but from a methodological viewpoint her study is promising.

In a conceptually related study, Armstrong (714) administered the CAT (cards 1, 2, 4, 8, and 10) to a sample of 60 children in grades 1 to 3 in the University of Minnesota elementary school. The findings of her study relevant to the present review are as follows: (1) length of story increases with grade, (2) girls' protocols are longer than those of boys, (3) the use of first person pronouns shows a slight but consistent decline with grade progression, (4) girls tend to make more subjective and personalized statements than boys, and (5) girls have a consistently longer reaction time than boys.

Slack (761) gave the TAT to 15 exogenous feebleminded boys and 12 endogenous ones at the Vineland Training School. He correlated a score reflecting *the number of causally and purposefully connected statements* with the Stanford-Binet

and with Thurstone's test of Primary Mental Abilities (PMA). With chronological age held constant, causally or purposefully connected statements correlated with other variables as follows: S-B MA, 0.58; PMA MA, 0.70; PMA Verbal MA, 0.51; PMA Motor MA, 0.72. Length of stories (number of words) correlated as follows with the same variables (CA held constant): S-B MA, 0.31 (ns); PMA MA, 0.34 (ns); PMA Verbal MA, 0.53; PMA Motor MA, 0.48. The age-corrected correlation of *number of purposeful relations* with the PMA Verbal MA was 0.90, and the correlation of *number of causal relations* with the same measures was 0.42. Slack also reported a significant difference between the endogenous and exogenous groups on length of stories.

These studies lend some limited support to the possibility of developing an objective scoring system based on developmental criteria for the five TAT pictures used in the Survey.

Other Relevant Research

The following studies were selected for citation on the basis of their relevance to the Survey problems. Lesser (720) demonstrated how a Guttman-type scale could be developed for measurement of aggressive fantasy. Bijou and Kenny (732) and Murstein (734) investigated ambiguity values of TAT cards. The former found the following ambiguity ranks (out of 21) for the four picture cards used in the Survey (card 16, blank, was not rated):

Card number	*Rank*
1	2
2	3
5	17
8BM	11

The latter reported that cards with medium ambiguity (8BM) were most "productive" of thematic content among college students.

Milam (735) demonstrated the sensitivity of TAT responses to examiner influence. Apparently, the attitudes and behavior of the examiner, as perceived by the subject, account for variance in the TAT responses. This is true of all psychological tests. It is not possible to say whether this is a greater problem on the TAT than on the WISC, for example, but it must be kept in mind as a significant source of uncontrolled variation.

Gurevitz and Klapper (763) found that schizophrenic children characteristically respond to CAT cards with bizarre outcomes, evaluation of stimuli, use of titles, hostility, and verbosity. Holden (766) compared a small sample of cerebral palsied children with normal controls. His results clearly suggest that cerebral palsied respondents tend to describe the cards, while normal controls give more thematic content. The average number of descriptions (out of 10 cards) was 6.0 for the palsied children and 2.8 for the controls. Leitch and Schafer (770) reported a number of response criteria identifying psychotic responses.

From the standpoint of further research on the development of a scoring procedure for the TAT, the following list of specific items has been recorded and evaluated in one or more of the studies reviewed (reference numbers shown in parentheses). In most cases the results were not included in the main discussion either because of sample limitations, subjective methods of scoring, inconclusiveness of results, or unrelatedness to the present problem. Many of them, however, do appear definable and worthy of further study.

- Frequency and duration
  - RT latency (705 and 747)
  - Total reaction time (705 and 747)
  - Number of words (707, 714, 737, 741, 746, 747, and 764)
  - Number of adjectives (737)
  - Number of adverbs (737)
  - Number of nouns (714)
  - Number of pronouns (714)
  - Number of verbs (714)
  - Number of questions (705)
  - Number of ego words (714)
  - Number of situations (737)
  - Number of characters (707 and 737)
    - Male, female
- Nature of action
  - Crying (718)
  - Dancing (737)
  - Disaster (713)
  - Drunkenness (737)

- Paranoid (705)
- Parental anger to child (743)
- Pleasant (705)
- Pleasure (713)
- Sadistic (705)
- Security (713)
- Number of causal relations (761)
- Number of purposeful relations (761)

Outcomes (713, 763, 772, and 775)
- Failure
- Success
- Aggressive (772)
- Clarity of statement (705)
- Bizarre (763)

Self-reference (705)

Number of personalized statements (705 and 714)

Degree of response certainty (705)

Level of interpretation (Eron, 712)
- Symbolic
- Abstract
- Descriptive
- Unreal
- Fairy tale
- Central character not in picture
- Autobiographical
- Continuations
- Alternate themes
- Comments
- Denial of theme
- Rejection
- Peculiar
- Confused
- Includes examiner in story
- No connection between story and picture
- Humorous

PROSPECTS FOR DEVELOPING AN OBJECTIVE SCORING KEY FOR THE SURVEY'S TAT

Although the TAT literature is scientifically "sloppy" in comparison with the material reviewed in relation to the WISC and the Draw-A-Man Test, the following assumptions seemed warranted: (1) a substantial number of items (both formal-structural and thematic-interpretive) can be reliably defined and accurately scored, (2) discriminating developmental criteria can be devised, and (3) an objectively defined scoring system can be developed which will contribute useful information regarding development between ages 6 and 12 years.

It seems unlikely, in light of the literature reviewed, that scoring scales can be constructed which will measure factors such as motivation, affective states, and personality traits. However, this is not serious since there is no indication that these factors have any developmental implications.

The anticipated developmental scales would greatly enrich the information obtained in the Survey by possibly providing developmental norms with regard to behavioral aspects not encompassed by the other tests, such as verbal expression, thematic content of imagination in standard test situations, associations to standard stimuli, role concepts and attitudes in relation to self, peers of same and opposite sex, parental and adult figures, and common cultural values.

While the picture samples are limited, they appear to be well chosen for the purpose. Card 1 has a boy as the central figure; card 2, a girl; card 5, an adult-parental (mother) figure; and card 8BM, a possible stressful situation—involving a father figure—within the experience background of most school-age children. Card 16, the blank card, is completely unstructured. As a set of cards having nearly universal applicability in a United States national sample, the selection appears excellent.

One of the advantages that an investigator working on this problem would have over most of those who have published reports in this area is the large sample obtained under standardized survey conditions. With adequate funds to work with a fairly large sample of perhaps 1,000 or more cases, a good test of these conclusions could be made. Of course, there is no guarantee that the results will be entirely satisfactory, although the prognosis appears good.

However, the Survey is committed to doing something with these data, and no suitable scoring procedure is presently available. In the writer's judgment, the options available were nearly all unsatisfactory, and the one taken may prove to be a wise decision.

BIBLIOGRAPHY

General References to TAT

701. Mayman, M.: Review of the literature on the Thematic Apperception Test, in David Rapaport, *Diagnostic Psychological Testing*. Vol. II, *The Theory, Statistical Evaluation, and Diagnostic Application of a Battery of Tests*. Chicago. Year Book Publishers, 1946. pp. 496-506.

702. Piotrowski, Z. A.: A new evaluation of the Thematic Apperception Test. *Psychoanalyt.Rev.* 37:101-127, 1950.

703. Lindzey, G.: Thematic Apperception Test, interpretive assumptions and related empirical evidence. *Psychology Bull.* 49:1-25, 1952.

704. Windle, C.: Psychological tests in psychopathological prognosis. *Psychology Bull.* 49:451-482, 1952.

705. Hartman, A. A.: An experimental examination of the Thematic Apperception Technique in clinical diagnosis. *Psychological Monographs*. Vol. 63, No. 8 (Whole No. 303). Washington, D.C. American Psychological Association, Inc., 1950.

706. Eron, L. D.: Some problems in the research application of the Thematic Apperception Test. *J.Project.Tech.* 19:125-129, 1955.

707. Lindzey, G., and Silverman, M.: Thematic Apperception Test, techniques of group administration, sex differences, and the role of verbal productivity. *J.Personality* 27:311-323, 1959.

708. Sanford, R. N., and others: Physique, personality and scholarship; a cooperative study of school children, in *Society for Research in Child Development, Monograph,* Vol. 8, No. 2. Washington, D.C. National Research Council, 1943.

TAT: Normative Data

709. Cox, B. F., and Sargent, H. D.: The common responses of normal children to ten pictures of the Thematic Apperception Test series; abstracted, *Am.Psychologist* 3:363, 1948.

710. Bell, J. E.: A comparison of children's fantasies in two equated projective techniques; abstracted, *Am.Psychologist* 3:263, 1948.

711. Whitehouse, E.: Norms for certain aspects of the Thematic Apperception Test on a group of nine and ten year old children; abstracted, *Persona* 1:12-15, 1949.

712. Eron, L. D.: A normative study of the Thematic Apperception Test. *Psychological Monographs*. Vol. 64, No. 9. Washington, D.C. American Psychological Association, Inc., 1950.

713. Cox, B., and Sargent, H. D.: TAT responses of emotionally disturbed and emotionally stable children, clinical judgment versus normative data. *J.Project.Tech.* 14:61-74, 1950.

714. Armstrong, M. A. S.: Children's responses to animal and human figures in thematic pictures. *J.Consult.Psychol.* 18:67-70, 1954.

715. Fisher, G. M., and Shotwell, A. M.: Preference rankings of the TAT cards by adolescent normals, delinquents, and mental retardates. *J.Project.Tech.* 25:41-43, 1961.

716. Brayer, R., Craig, G., and Teichner, W.: Scaling difficulty values of TAT cards. *J.Project.Tech.* 25:272-276, 1961.

TAT: Scoring Schemes

717. Eron, L. D., Terry, D., and Callahan, R.: The use of rating scales for emotional tone of TAT stories. *J.Consult.Psychol.* 14:473-478, 1950.

718. Fine, R.: A scoring scheme for the TAT and other verbal projective techniques. *J.Project.Tech.* 19:306-309, 1955.

719. Friedman, I.:Objectifying the subjective, a methodological approach to the TAT. *J.Project.Tech.* 21:243-247, 1957.

720. Lesser, G. S.: Application of Guttman's scaling method to aggressive fantasy in children. *Educ.Psychol.Measur.* 18:543-551, 1958.

721. Dana, R. H.: Proposal for objective scoring of the TAT. *Percept.Mot.Skills* 9:27-43, 1959.

TAT: Stability, Reliability

722. Porter, F. S.: *A Study of Certain Aspects of the Reliability and Validity of the Thematic Apperception Test.* Unpublished master's thesis, Iowa State University, 1944.

723. Harrison, R., and Rotter, J. B.: A note on the reliability of the Thematic Apperception Test. *J.Abnorm.&Social Psychol.* 40:97-99, 1945.

724. Jeffre, M. F. D.: *A Critical Study of the Thematic Apperception Test Performance of Normal Children.* Unpublished master's thesis, University of Iowa, 1945.

725. Mayman, M., and Kutner, B.: Reliability in analyzing Thematic Apperception Test stories. *J.Abnorm.&Social Psychol.* 42:365-368, 1947.

726. Kagan, J.: The stability of TAT fantasy and stimulus ambiguity. *J.Consult.Psychol.* 23:266-271, 1959.

TAT: Validity Studies

727. Calvin, J. S., and Ward, L. C.: An attempted experimental validation of the Thematic Apperception Test. *J.Clin. Psychol.* 6:377-381, 1950.

728. Saxe, C. H.: A quantitative comparison of psychodiagnostic formulations from the TAT and therapeutic contacts. *J.Consult.Psychol.* 14:116-127, 1950.

729. Davenport, B. F.: The semantic validity of TAT interpretations. *J.Consult.Psychol.* 16:171-175, 1952.

730. Bendig, A. W.: Predictive and postdictive validity of need achievement measures. *J.Ed.Res.* 52:119-120, 1958.

731. Henry, W. E., and Farley, J.: The validity of the Thematic Apperception Test in the study of adolescent personality. *Psychological Monographs*. Vol. 73, No. 17 (Whole No. 487). Washington, D.C. American Psychological Association, Inc., 1959.

TAT: Ambiguity Values of Cards

732. Bijou, S. W., and Kenny, D. T.: The ambiguity values of TAT cards. *J.Consult.Psychol.* 15:203-209, 1951.

733. Davenport, B. F.: *The Ambiguity, Universality, and Reliable Discrimination of TAT Interpretations.* Unpublished doctoral dissertation, University of Southern California, 1951.

734. Murstein, B. I.: The relationship of stimulus ambiguity on the TAT to the productivity of themes. *J.Consult. Psychol.* 22:348, 1958.

TAT: Examiner Influence, Interpreter Influence

735. Milam, J. R.: Examiner influences on Thematic Apperception Test stories. *J.Project.Tech.* 18:221-226, 1954.

736. Young, R. D., Jr.: *The Effect of the Interpreter's Personality on the Interpretation of Thematic Apperception Test's Protocols.* Unpublished doctoral dissertation, University of Texas, 1953.

TAT: Effects of Intelligence, Achievement

737. Edelstein, R. T.: *The Evaluation of Intelligence From TAT Protocols.* Unpublished master's thesis, College of the City of New York, 1956.

738. Kagan, J., Sontag, L. W., Baker, D. T., and Nelson, V. L.: Personality and IQ change. *J.Abnorm.&Social Psychol.* 56:261-266, 1958.

739. Murstein, B. I., and Collier, H. L.: The role of the TAT in the measurement of achievement as a function of expectancy. *J.Project.Tech.* 26:96-101, 1962.

TAT: Personality Variables

740. McDowell, J. V.: *Development Aspects of Phantasy Production on the Thematic Apperception Test.* Unpublished doctoral dissertation, Oklahoma State University, 1952.

741. Cook, R. A.: Identification and ego defensiveness in thematic apperception. *J.Project.Tech.* 17:312-319, 1953.

742. Mussen, P. H., and Naylor, H. K.: Relationships between overt and fantasy aggression. *J.Abnorm.&Social Psychol.* 49:235-240, 1954.

743. Kagan, J.: Socialization of aggression and the perception of parents in fantasy. *Child Development* 29:311-320, 1958.

744. Fitzgerald, B. J.: *The Relationship of Two Projective Measures to a Sociometric Measure of Dependent Behavior.* Unpublished doctoral dissertation, Ohio State University, 1959.

745. Breger, L.: *Conformity and the Expression of Hostility.* Unpublished doctoral dissertation, Ohio State University, 1961.

TAT: Effects of Set, Recent Experience, Stimulus Variables

746. Lubin, B.: Some effects of set and stimulus property on TAT stories. *J.Project.Tech.* 24:11-16, 1960.

747. Newbigging, P. L.: Influence of a stimulus variable on stories told to certain TAT pictures. *Can.J.Psychol.* 9:195-206, 1955.

748. Coleman, W.: The Thematic Apperception Test. I, Effect of recent experience. II, Some quantitative observations. *J.Clin.Psychol.* 3:257-264, 1947.

TAT: Environmental Variations; Culture, Social Class, Race, Ethnic Group, Home Conditions, Sex Role, Sociometric Status, Social Acceptance

749. Henry, W. E.: The Thematic Apperception Technique in the study of culture-personality relations. *Genet.Psychol.Monogr.* 35:3-135, 1947.

750. Mason, B., and Ammons, R. B.: Note on social class and the Thematic Apperception Test. *Percept.Mot.Skills* 6:88, 1956.

751. Fisher, S., and Fisher, R. L.: A projective test analysis of ethnic subculture themes in families. *J.Project.Tech.* 24:366-369, 1960.

752. Mitchell, H. E.: Social class and race as factors affecting the role of the family in Thematic Apperception Test stories; abstracted, *Am.Psychologist* 5:299-300, 1950.

753. Mussen, P. H.: Differences between the TAT responses of Negro and white boys. *J.Consult.Psychol.* 17:373-376, 1953.

754. Mussen, P. H.: Some personality and social factors related to changes in children's attitudes toward Negroes. *J.Abnorm.&Social Psychol.* 45:423-441, 1950.

755. Shields, D. L.: *An Investigation of the Influences of Disparate Home Conditions Upon the Level at Which Children Responded to the Thematic Apperception Test.* Unpublished master's thesis, University of Pittsburgh, 1950.

756. McArthur, C.: Personality differences between middle and upper classes. *J.Abnorm.&Social Psychol.* 50:247-254, 1955.

757. Cox, F. N.: Sociometric status and individual adjustment before and after play therapy. *J.Abnorm.&Social Psychol.* 48:354-356, 1953.

758. Herman, G. N.: *A Comparison of the TAT Stories of Preadolescent School Children Differing in Social Acceptance.* Unpublished master's thesis, University of Toronto, 1952.

759. Milner, E.: Effects of sex role and social status on the early adolescent personality. *Genet.Psychol.Monogr.* 40:231-325, 1949.

760. Butler, O. P.: *Parent Figures in Thematic Apperception Test Records of Children in Disparate Family Situations.* Unpublished doctoral dissertation, University of Pittsburgh, 1948.

TAT: With Feebleminded, Retarded, Handicapped, Brain Injured, Palsied, Disturbed, and Psychotic Children

761. Slack, C. W.: Some intellective functions in the Thematic Apperception Test and their use in differentiating endogenous feeblemindedness from exogenous feeblemindedness. *Train.Sch.Bull.* 47:156-169, 1950.

762. Tolman, N. G., and Johnson, A. P.: Need for achievement as related to brain injury in mentally retarded children. *Am.J.Ment.Deficiency* 62:692-697, 1958.

763. Gurevitz, S., and Klapper, Z. S.: Techniques for and evaluation of the responses of schizophrenic and cerebral palsied children to the Children's Apperception Test (C.A.T.). *Quart. J. Child Behavior* 3:38-65, 1951.

764. Abel, T. M.: Responses of Negro and white morons to the Thematic Apperception Test. *Am.J.Ment.Deficiency* 49:463-468, 1945.

765. Beier, E. G., Gorlow, L., and Stacey, C. L.: The fantasy life of the mental defective. *Am.J.Ment.Deficiency* 55: 582-589, 1951.

766. Holden, R. H.: The Children's Apperception Test with cerebral palsied and normal children. *Child Development* 27:3-8, 1956.

767. Hood, P. N., Shank, K. H., and Williamson, D.: Environmental factors in relation to the speech of cerebral palsied children. *J.Speech & Hearing Disorders* 13:325-331, 1948.

768. Bergman, M., and Fisher, L. A.: The value of the Thematic Apperception Test in mental deficiency. *Psychiat. Quart.Suppl.* 27:22-42, 1953.

769. Ericson, M.: A study of the Thematic Apperception Test as applied to a group of disturbed children; abstracted, *Am.Psychologist*, 2:271, 1947.

770. Leitch, M., and Schafer, S.: A study of the Thematic Apperception Tests of psychotic children. *Am.J.Orthopsychiat.* 17:337-342, 1947.

771. Shank, K. H.: *An Analysis of the Degree of Relationship Between the Thematic Apperception Test and an Original Projective Test in Measuring Symptoms of Personality Dynamics of Speech Handicapped Children.* Unpublished doctoral dissertation, University of Denver, 1954.

772. Christensen, A. H.: *A Quantitative Study of Personality Dynamics in Stuttering and Non-Stuttering Siblings.* Unpublished master's thesis, University of Southern California, 1951.

773. Young, F. M.: Responses of juvenile delinquents to the Thematic Apperception Test. *J.Genet.Psychol.* 88:251-259, 1956.

TAT: With CAT and Michigan Picture Test

774. Symonds, P. M.: *Adolescent Fantasy, an Investigation of the Picture-Story Method of Personality Study.* New York. Columbia University Press, 1949.

775. Light, B. H.: Comparative study of a series of TAT and CAT cards. *J.Clin.Psychol.* 10:179-181, 1954.

776. Andrew, G., Walton, R. E., Hartwell, S. W., and Hutt, M. L.: The Michigan Picture Test, the stimulus value of the cards. *J.Consult.Psychol.* 51:51-54, 1951.

Special Bibliography of TAT Scoring Systems[1]

777. Andrew, G., Hartwell, S. W., Hutt, M. L., and Walton, R. E.: *The Michigan Picture Test.* Chicago. Science Research Associates, Inc., 1953.

778. Arnold, M. B.: A demonstration analysis of the Thematic Apperception Test in a clinical setting. *J.Abnorm.& Social Psychol.* 44:97-111, 1949.

779. Aron, B.: *A Manual for Analysis of the Thematic Apperception Test.* Berkeley, Calif. Willis E. Berg, 1949.

780. Bellak, L.: *A Guide to the Interpretation of the Thematic Apperception Test.* New York. The Psychological Corporation, 1947.

781. Cox, B., and Sargent, H.: TAT responses of emotionally disturbed and emotionally stable children. *J.Project. Tech.* 14:61-74, 1950.

782. Dana, R. H.: Norms for three aspects of TAT behavior. *J.Genet.Psychol.* 57:83-89, 1957.

783. Fine, R.: *Manual for Scoring Scheme for Verbal Projective Techniques (TAT, MAPS, Stories, and the Like).* Washington, D.C. Veterans Administration, 1948.

784. Fry, F. D.: Manual for scoring the TAT. *J.Psychol.* 35: 181-195, 1953.

785. Hartman, A. A.: An experimental examination of the Thematic Apperception Technique in clinical diagnosis. *Psychological Monographs.* Vol. 63, No. 8 (Whole No. 303). Washington. D.C. American Psychological Association, Inc., 1950. pp. 1-48.

786. Henry, W. E.: *The Analysis of Fantasy.* New York. John Wiley and Sons, Inc., 1956.

787. Klebanoff, S.: Personality factors in symptomatic chronic alcoholism as indicated by the Thematic Apperception Test. *J.Consult.Psychol.* 11:111-119, 1947.

788. Murray, H. A.: *Explorations in Personality.* New York. Oxford University Press, 1938.

789. Rappaport, D.: The Thematic Apperception Test, Ch. IV, in *Diagnostic Psychological Testing,* Vol. II, Chicago. Yearbook Publishers, Inc., 1946.

790. Shorr, J. E.: A proposed system for scoring the TAT. *J. Clin.Psychol.* 4:189-195, 1948.

791. Stone, H.: The TAT Aggressive Content Scale. *J.Proj. Tech.* 20:445-452, 1956.

792. Terry, D.: The use of a rating scale of level of response in TAT stories. *J.Abnorm.&Social Psychol.* 47:507-511, 1952.

793. Tomkins, S. S., and Tomkins, E. S.: *The Thematic Apperception Test, the Theory and Technique of Interpretation.* New York. Grune and Stratton, 1948.

794. White, R. K.: *Value Analysis, the Nature and Use of the Method.* New York. Society for the Psychological Study of Social Issues, 1951.

795. Wyatt, F.: The scoring and analysis of the Thematic Apperception Test. *J.Psychol.* 24:319-330, 1947.

Other References Cited in Text

796. Cattell, R. B.: *Personality and Motivation Structure and Measurement.* New York. Harcourt, Brace and World, 1959.

797. Kelly, G. A.: The theory and technique of assessment, in P. R. Farnsworth and Q. McNemar, eds., *Annual Review of Psychology,* Vol. 9. Palo Alto, Calif. Annual Reviews, Inc., 1958.

798. McClelland, D.: *Studies in Motivation.* New York. Appleton-Century-Crofts, Inc., 1955.

799. Murray, H. A.: *Thematic Apperception Test, Pictures and Manual.* Cambridge. Harvard University Press, 1943.

800. Shneidman, E. S., Joel, W., and Little, K. B.: *Thematic Test Analysis.* New York. Grune and Stratton, 1951.

[1]See also 717 to 721.

V. TOTAL PSYCHOLOGICAL TEST BATTERY

The foregoing reviews of the several components of the Survey's psychological test battery have discussed the strengths and weaknesses of each test and the problems involved in estimating population parameters on a national scale from the sample data. In each case a number of specific problems were raised, and suggestions for treatment of data or for further research have been made in the respective sections of the report. However, the most important common problem derives from the examination of the standardization basis of these tests. The norms for the WISC are unquestionably the most satisfactory, with the Draw-A-Man being second; the adequacy of the Wide Range Achievement Test norms has been questioned (see section II). Finally, new norms, related to the scoring system to be developed for the TAT, are yet to be constructed.

In order to achieve the soundest possible basis for population estimates with this battery, it is recommended that new national norms, based on the total Survey sample, be developed for all of the tests before any final population estimates are published. While some preliminary estimates may be warranted, using norms provided by the test publishers, the discussions in the individual sections of the report point up the necessity of the recommended restandardization.

In the event that this work cannot be fully supported, the order of priority indicated by the review would place the reanalysis of the WRAT first, the Draw-A-Man Test second, and the WISC third. It is assumed that this must be done for the TAT when a new scoring procedure is completed and adopted.

The issues in relation to the WRAT are as follows: (1) No adequate sampling plan was followed in standardizing the 1963 revision, and, in fact, the bias of the sample is clearly mentioned in the manual. (2) The test scores used to compile the sample by levels are not equivalent; therefore, only limited confidence can be placed in the resulting norm levels, even though substantial correlation of the WRAT scales with concurrent criteria appears likely.

In the case of the Draw-A-Man Test, it is recognized that (1) the Goodenough norms are outmoded, and that (2) the use of the Harris norms (which is recommended) without analysis of the raw score distributions on the national sample might lead to some errors. The administration of the Draw-A-Man Test in the Survey was different from that recommended by Harris, and it would be prudent to proceed empirically rather than to assume that the Survey drawings are equivalent. In addition, Harris' own norms do not reflect as good a national sample as even the WISC, for which further standardization is unquestionably justified.

One of the major problems with the WISC subtests is that of examining further the optional basis for estimating Full Scale IQ's from the Vocabulary and Block Design scores. Even if restandardization should reveal no need for rescaling the subtest items, the adoption of published conversion tables or direct proration is considered unjustified without further research. This is discussed in more detail in section I.

The information expected from the test battery may be summarized as follows:

1. *WISC Vocabulary*—score. This test individually provides a good estimate of "g," the common "general intelligence" factor in the WISC, and may be accepted as a good measure of the verbal component of the general measure of intelligence.
2. *WISC Block Design*—score. This test is also well saturated in "g" and second only to Vocabulary in reliability. It should be accepted as a strong nonverbal intelligence test and as an estimate of the nonverbal component of the full test.
3. *Draw-A-Man Test*— Goodenough-Harris standard score. The Goodenough-Harris standard score (preferably restandardized on the total Survey sample) can be interpreted as a deviation IQ, in a manner comparable to the WISC IQ's. This score is a reliable and reasonably valid nonlanguage measure of mental maturity.
4. *WRAT Oral Reading*—grade equivalent (Rq).
5. *WRAT Oral Reading*—standard score (Rss).
6. *WRAT Arithmetic*—grade equivalent (Aq).
7. *WRAT Arithmetic*—standard score (Ass).

Both the grade equivalents and the standard scores will be useful for the WRAT Reading and Arithmetic subtests (particularly if they are restandardized on the total Survey sample). The grade equivalents will permit assessment of school retardation, while the standard scores, which have the same characteristics as deviation IQ's, will be more appropriate in pattern analytic combination with the WISC and Draw-A-Man scores.

8. *TAT*— developmental score(s). This may actually be a series of scores. It is entered "symbolically" at this time.

It is possible to think of these data as providing individual profiles or patterns which supplement information represented by the individual scores. For example, some children may rank high or low on all scales, indicating general excellence or retardation in comparison with the general population. There may also be discriminable test patterns associated with such special conditions as reading disability, mental deficiency, scholastic retardation, verbal impairment due to physical or social reasons, behavior disorders, and cultural deprivation. If such patterns exist, it should be possible to identify them by a standard research design based on discrimination of experimentally formed criterion groups. A hierarchical grouping analysis of score profiles, seeking to identify characteristic profiles of groups, would be an alternative approach.

In this procedure, identification of criterion characteristics of the groups would follow rather than precede the main analysis. In either case, criterion data would be obtained from record sources within the Health Examination Survey. In this type of analysis it might also be profitable to explore patterns based on scores representing discrete residuals, with common variance partialled out and represented by an additional variable.

Computer programs for these types of analysis are available, and such studies could be conducted economically on subsamples of the Survey sample.

The inclusion of these psychological tests in the National Health Survey was a very important step which has tremendous practical value to the health, education, and welfare fields and which also has immense scientific value in the life sciences concerned with child development. Despite the technical criticisms, which are inevitable in a problem of the magnitude of this national survey, the tests have been judged to be either a good choice or at least an eminently reasonable compromise with reality within the constraints of the Survey.

The research recommended should be looked on as an unprecedented opportunity to contribute toward adequate mental measurement of children. It is important for those working in this Survey to bear in mind that this is the first general survey of psychological functions of children ever conducted on a sophisticated national sample. The standardization programs for the tests reviewed—and for others referred to—fail to qualify for this distinction. National psychological surveys of adults have been made in both World Wars, and recently a national survey of adolescents was conducted by Project TALENT. However, Cycle II is, to the writer's knowledge, the first one of its kind in the age range of 6 to 12 years.

VI. CROSS-DISCIPLINARY ANALYSES

The complete data of Cycle II may be regarded as composing a matrix of several thousand variables (specific measures or components of measurement procedures) over a sample of nearly 8,000 children. In the processes of data reduction and analysis, many of these variables will remain in the matrix without further manipulation (e.g., height, weight, body temperature, family income level, twin status, number of siblings, and ages of parents). Some will require prescheduled analysis and computation of indexes according to established procedures in the respective fields (e.g., visual acuity, exercise tolerance, and electrocardiogram), while others will require extensive processing on the basis of empirically constructed or revised scoring keys and norms,

as in the case of the psychological tests discussed in this review.

Upon completion of segmental analysis of each testing and examining procedure and reduction of all data to indexes and primary variables, it would be desirable to consider multivariate analysis of the resulting matrix. This type of approach will undoubtedly reveal many significant interrelationships not previously investigated because of lack of appropriate data. It is premature to consider it now, however, before the reduced data schedule is more definitely known.

The primary purpose of the present discussion is to explore possible linkages between the psychological tests in the Survey battery and other variables. This, too, is a formidable task, but some important areas of investigation are opened up by this Survey, and these opportunities for significant research deserve special mention.

DATA AVAILABLE

From various sources within the Survey, data on items such as the following, which have important behavioral implications, will be available:

Parents—age, nativity, education, income level, language spoken, psychiatric history, marital status, handedness, and use of medical care. (The distributions of these variables are of interest. In addition, an SES index of *socioeconomic* level can be derived.)

Siblings—number, twins, ages, education, marital status, work status. (From these data an additional variable, *birth ordinal position,* can be derived.)

Family—size, living status, ethnic classification, race, SES.

Child—*school information:* grade placement; progress rate; absences; characterization as requiring special provision for hard of hearing, visually handicapped, speech therapy, orthopedically handicapped, gifted, slow learning, mentally retarded, emotionally disturbed; description in relation to adjustment, attention, interpersonal relations, discipline, popularity, intellectual ability, academic performance. (These data are worthy of some detailed analysis in order to formulate external rating criteria for independent test validation and to derive further indexes, such as *peer rejection* (based on interpersonal relations and popularity), *general adjustment,* and *general adequacy* (based on a frequency count of negative citation).

Child—*medical history:* prenatal and birth circumstances, food habits, enuresis, thumbsucking, age of walking, talking, early learning rate, attendance at kindergarten, experience of unconsciousness, bad burns (with resulting scars), serious illness, weakness, nightmares, sleeping arrangements, age at puberty (girls). (Frequency distributions of these items, particularly of food habits, which would also provide a basis for judging food idiosyncracies, and sleeping arrangements, which should correlate with SES but may also relate to other variables, should be of great interest. Correlations of many of these items with other data may be extremely important, as, for example, the investigation of *sequelae of early unconsciousness* and the development of a *growth retardation classification,* a *disturbance index,* and a *"weakness" index.)*

Child—*sensory and motor indexes:* visual acuity, color vision, hearing indexes, handedness, grip strength, vital capacity, exercise tolerance.

Child—*body measurements:* height, weight, anthropometry, X-ray, dentition.

Child—*psychophysiological indexes:* blood pressure, temperature, electrocardiogram, phonocardiogram.

Child—*medical findings:* health status, pathology.

Child—*psychological tests:* IQ estimates; verbal ability level; performance ability level; reading, arithmetic, maturity level; adjustment index.

ANALYSES INDICATED

The organization and ordering of the lines of analysis suggested in this section are tentative and are not intended to suggest priorities. In most cases, further study of the literature in the particular areas and consultation with qualified professional persons would be appropriate before committing time and funds to particular studies.

Nevertheless, the richness of this "data bank" is recognized as a source of new scientific knowledge, and it is hoped that it can be adequately exploited.

Growth Indexes

It is expected that mean growth indexes for boys and girls will be computed for as many functions as possible over the six age periods. Analysis of relations among growth trends—separately for boys and for girls—and of growth rate patterns would be of direct interest and would also permit comparison of pattern indexes with psychological test scores. Sex differences in growth patterns and relations of sex-related patterns to test scores are also of great interest.

Other Factors Related to Test Scores

Discriminant pattern analyses might be undertaken systematically in a multivariate design to investigate parental, sibling (including birth order and twin resemblance for the twin sample), family, school, medical, sensory and motor, anthropometric, psychophysiological, and medical correlates of psychological test scores. While this recommendation may appear forbidding in magnitude, the multivariate approach is actually more efficient and economical in total perspective than piecemeal analyses. Among the studies implied in this broad prescription are the following types of investigations:

1. *Reading disability.* Effects of visual and auditory impairment; handedness; SES; growth trends; developmental history; early, recent, and continuing emotional disturbance; illness; birth order, etc.
2. *Mental retardation.* Every item in the above enumeration is potentially related to mental retardation.
3. *School retardation.* Same as above.
4. *Analyses of discrepancies between actual and predicted status in relation to concomitant or associated factors.* These data offer an excellent opportunity to look for significant variance associated with overachievement and underachievement in school grade placement, reading achievement (WRAT and school report), scholastic achievement (school report, WRAT Arithmetic), and peer relations (deviation from central tendency).

While more detailed and specific investigations could be enumerated, it is more constructive to emphasize the advisability of using the multivariate approach, since computer equipment and programs are available for such analyses and since results of greater value can be obtained at a far lower unit cost.

Acknowledgments

The literature review and preparation of abstracts was under the immediate direction of Samuel H. Cox, Research Associate at the Institute of Behavioral Research. Principal persons assisting Mr. Cox were Robert M. Marx, John McCrady, Henry Orloff, and Max S. Taggart II.

The project also was greatly expedited through the efforts of Miss Johnoween Gill, Reference Librarian, Texas Christian University.

Without the loyal and competent help of these individuals this report could not have been completed in only 3 months.

GLOSSARY OF ABBREVIATIONS

BD:	Block Design subtest of the Wechsler Intelligence Scale for Children
CA:	Chronological age
CAT:	Children's Apperception Test
CMAS:	Children's Manifest Anxiety Scale
CRT:	California Reading Test
CTMM:	Chicago Tests of Primary Mental Abilities
E-G-Y:	Kent E-G-Y Test (Scale D, Kent Series of Emergency Scales)
FRPV:	Full-Range Picture Vocabulary Test (by Ammons)
FS:	Full Scale (or Full Score) of the Wechsler Intelligence Scales
g:	General, or "global," intelligence factor
HES:	Health Examination Survey
IQ:	Intelligence quotient
M:	Mean
MA:	Mental age
N.	Number
ns:	Not significant
PPVT:	Peabody Picture Vocabulary Test
PS:	Performance Scale (or Performance Score) of the Wechsler Intelligence tests
R:	Range
r:	Correlation
RT:	Response time
SAT:	Stanford Achievement Test
S-B:	Stanford-Binet Intelligence Scale
SES:	Socioeconomic status
SRA:	Science Research Associates, Inc.
SRA-PMA:	SRA Primary Mental Abilities
SS:	Standard score
TAT:	Thematic Apperception Test
Voc.:	Vocabulary subtest of the Wechsler Intelligence Scales
VS:	Verbal Scale (or Verbal Score) of the Wechsler Intelligence tests
WAIS:	Wechsler Adult Intelligence Scale
WISC:	Wechsler Intelligence Scale for Children
WRAT:	Wide Range Achievement Test

☆ U. S. GOVERNMENT PRINTING OFFICE : 1975 584-528/38

OUTLINE OF REPORT SERIES FOR VITAL AND HEALTH STATISTICS

Originally Public Health Service Publication No. 1000

Series 1. Programs and collection procedures.—Reports which describe the general programs of the National Center for Health Statistics and its offices and divisions, data collection methods used, definitions, and other material necessary for understanding the data.

Series 2. Data evaluation and methods research.—Studies of new statistical methodology including: experimental tests of new survey methods, studies of vital statistics collection methods, new analytical techniques, objective evaluations of reliability of collected data, contributions to statistical theory.

Series 3. Analytical studies.—Reports presenting analytical or interpretive studies based on vital and health statistics, carrying the analysis further than the expository types of reports in the other series.

Series 4. Documents and committee reports.—Final reports of major committees concerned with vital and health statistics, and documents such as recommended model vital registration laws and revised birth and death certificates.

Series 10. Data from the Health Interview Survey.—Statistics on illness, accidental injuries, disability, use of hospital, medical, dental, and other services, and other health-related topics, based on data collected in a continuing national household interview survey.

Series 11. Data from the Health Examination Survey.—Data from direct examination, testing, and measurement of national samples of the population provide the basis for two types of reports: (1) estimates of the medically defined prevalence of specific diseases in the United States and the distributions of the population with respect to physical, physiological, and psychological characteristics; and (2) analysis of relationships among the various measurements without reference to an explicit finite universe of persons.

Series 12. Data from the Institutional Population Surveys.—Statistics relating to the health characteristics of persons in institutions, and on medical, nursing, and personal care received, based on national samples of establishments providing these services and samples of the residents or patients.

Series 13. Data from the Hospital Discharge Survey.—Statistics relating to discharged patients in short-stay hospitals, based on a sample of patient records in a national sample of hospitals.

Series 14. Data on health resources: manpower and facilities.—Statistics on the numbers, geographic distribution, and characteristics of health resources including physicians, dentists, nurses, other health manpower occupations, hospitals, nursing homes, and outpatient and other inpatient facilities.

Series 20. Data on mortality.—Various statistics on mortality other than as included in annual or monthly reports—special analyses by cause of death, age, and other demographic variables, also geographic and time series analyses.

Series 21. Data on natality, marriage, and divorce.—Various statistics on natality, marriage, and divorce other than as included in annual or monthly reports—special analyses by demographic variables, also geographic and time series analyses, studies of fertility.

Series 22. Data from the National Natality and Mortality Surveys.—Statistics on characteristics of births and deaths not available from the vital records, based on sample surveys stemming from these records, including such topics as mortality by socioeconomic class, medical experience in the last year of life, characteristics of pregnancy, etc.

For a list of titles of reports published in these series, write to:

Office of Information
National Center for Health Statistics
Public Health Service, HRA
Rockville, Md. 20852

DHEW Publication No. (HRA) 75-1295
Series 2 - No. 15

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Public Health Service

Health Resources Administration
5600 Fishers Lane
Rockville, Md. 20852

OFFICIAL BUSINESS
Penalty for Private Use, $300

POSTAGE AND FEES PAID
U.S. DEPARTMENT OF H.E.W.

HEW 390

THIRD CLASS
BLK. RATE