AN INVESTIGATION OF THE VALIDITY OF
THE AMERICAN SCHOOL INTELLIGENCE TESTS

A Thesis
Presented to
the Department of Special Education
Indiana State Teachers College

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

by
Roger Alan Hofferth
May 1961
THESIS APPROVAL SHEET

The thesis of Roger A. Hofferth, contribution of the Graduate School, Indiana State Teachers College, Series I, Number 811, under the title, "An Investigation of the Validity of the American School Intelligence Tests," is approved as counting toward the completion of the Master of Science Degree in the amount of six semester hours of graduate credit.

APPROVAL OF THESIS COMMITTEE:

[Signatures and Dates]

APPROVAL FOR GRADUATE SCHOOL

[Signature and Date]
### TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. THE PROBLEM AND DEFINITIONS OF TERMS USED</td>
<td>1</td>
</tr>
<tr>
<td>The Problem</td>
<td>2</td>
</tr>
<tr>
<td>Statement of the problem</td>
<td>2</td>
</tr>
<tr>
<td>Importance of the study</td>
<td>2</td>
</tr>
<tr>
<td>Definitions of Terms Used</td>
<td>5</td>
</tr>
<tr>
<td>Validity</td>
<td>5</td>
</tr>
<tr>
<td>Content validity</td>
<td>6</td>
</tr>
<tr>
<td>Predictive validity</td>
<td>7</td>
</tr>
<tr>
<td>Concurrent validity</td>
<td>7</td>
</tr>
<tr>
<td>Construct validity</td>
<td>7</td>
</tr>
<tr>
<td>Intelligence</td>
<td>7</td>
</tr>
<tr>
<td>Limitations and Delimitations</td>
<td>9</td>
</tr>
<tr>
<td>Delimitations</td>
<td>9</td>
</tr>
<tr>
<td>Limitations</td>
<td>13</td>
</tr>
<tr>
<td>Organization of Remainder of the Thesis</td>
<td>14</td>
</tr>
<tr>
<td>II. THE GROUP STUDIED AND THE MEASUREMENTS USED</td>
<td>15</td>
</tr>
<tr>
<td>The Group Used</td>
<td>15</td>
</tr>
<tr>
<td>The American School Intelligence Tests</td>
<td>16</td>
</tr>
<tr>
<td>Scoring of the ASIT</td>
<td>17</td>
</tr>
<tr>
<td>Standardization of the ASIT</td>
<td>18</td>
</tr>
<tr>
<td>Reliability and Validity of the ASIT</td>
<td>19</td>
</tr>
<tr>
<td>Reliability of the ASIT</td>
<td>19</td>
</tr>
<tr>
<td>Validity of the ASIT</td>
<td>20</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. The Means, Standard Deviations and Standard Errors of the Means for Both the WISC and ASIT at Intermediate Level (Grades 4,5,6)</td>
<td>34</td>
</tr>
<tr>
<td>II. The Means, Standard Deviations and Standard Errors of the Means for Both the WISC and ASIT at Advanced Level (Grades 7,8,9)</td>
<td>34</td>
</tr>
<tr>
<td>III. Coefficients of Correlation Between Both Forms of the ASIT and the Three IQ's of the WISC Intermediate Level--Grades 4,5,6 (64 Cases)</td>
<td>35</td>
</tr>
<tr>
<td>IV. Coefficients of Correlation between Both Forms of the ASIT and the Three IQ's of the WISC. Advanced Level--Grades 7,8,9 (53 cases)</td>
<td>36</td>
</tr>
</tbody>
</table>
CHAPTER I

THE PROBLEM AND DEFINITIONS OF TERMS USED

"To enable the right pupils to receive the right education from the right teachers may be considered the aim of a good educational system." ¹ If such an aim is to be realized, there must be a system of measurement and evaluation set up to determine the specifics of the "right" pupils, the "right" education, and the "right" teachers. The testing of intelligence as a means of determining educational goals, grouping students, and predicting academic success is now commonplace in the public schools of the country. The testing boom of the twenties which followed the success of group intelligence testing in the Army in World War One led to the application of group intelligence tests that far outran their technical improvements. "The fact that the tests were still very crude instruments was often forgotten in the rush of gathering scores and drawing conclusions therefrom."² The time has now passed when intelligence tests and other tests are accepted by teachers, psychologists and other test users.


at face value. A more skeptical attitude now prevails. Before a test is given much weight today, it undergoes much careful preparation by the authors and research departments of test publishers. However, even after such extensive refinement, much more research is done in the field by the consumers of tests and test results to add to the available data on the accuracy of tests.

I. THE PROBLEM

Statement of the problem. It was the purpose of the study to (1) investigate the statistical construct validity of the American School Intelligence Tests (hereinafter referred to as the ASIT) by (2) determining the coefficients of correlation between IQ's on the two forms of the ASIT at two levels with Verbal, Performance, and Full Scale IQ's on the Wechsler Intelligence Scale for Children (WISC) made by the same children, thereby determining if (3) the ASIT and the WISC measure essentially the same thing, namely, intelligence or the kind of thinking necessary to do the kind of work required in the public schools of the country.

Importance of the study. The Technical Recommendations for Psychological Tests and Diagnostic Techniques of the American Psychological Association states that the highest degree of development is needed for tests distributed for
use in practical situations where the user is unlikely to validate the tests for himself. These users have to assume that the test does measure what it presumes to measure purely on the basis of its title and manual. The American School Intelligence Tests fall into such a category. These tests are designed to be administered by the classroom teacher or similar person who has minimal training in test administration and interpretation. Such a person has a very practical use for the results obtained from such tests and probably has neither the time nor the inclination to validate a test himself. It is highly important, then, to make sure that the validity of a measuring instrument placed in his hands has been established beforehand.

Validity refers to the truthfulness of a test. In other words, "Does this test really test what it says it is testing?" More specifically, "does the American School Intelligence Test really test intelligence, or does it test some other quality such as reading ability?" It is not enough to label a test "The Intelligence Test" unless it has been firmly established that the test does indeed measure intelligence. Two tests labelled "Intelligence Tests"

are often assumed to measure the same thing. However, the child who is tested by both tests has been subjected to entirely different sets of tasks. If he is again tested by both tests several years later, he will be faced by more difficult tasks—tasks different from the earlier age level tasks on test A and different from those of test B. Positive correlations between the two scores, however, show that some overlapping does occur, but the relatively low magnitude of such correlations shows that the tests and tasks are by no means identical. 

It is essential then to actually "test the test" to determine the degree to which it actually measures what it purports to measure. It is interesting to note that the earlier intelligence tests were the children of necessity. Alfred Binet, for example, was searching for a means of locating the mentally retarded in the public schools of Paris. It is paradoxical that the Binet-Simon Scale was revised by Terman and used to study the people at the other end of the mental ability continuum, the gifted. The point is, that most of the early developers of intelligence tests had definite, practical purposes in mind for their tests. They were interested primarily in the results obtained and the objective, reliability and validity of their instruments for their own use. Today there are many test publishing

---

companies and book companies in competition for the sale of tests. This is not to say that the test companies are not interested in the accuracy of their tests, but just the opposite. The research divisions of the major test makers along with the authors and interested researchers have combined to make today's tests much more refined in terms of validity, reliability and usefulness than the earlier tests could possibly be.

Whatever the prime motivating force, all tests today are of necessity submitted to the scrutiny of many people and many techniques. The worth-while ones will be strengthened and accepted while the less defensible ones will be either improved or weeded out.

II. DEFINITIONS OF TERMS USED

Validity. "The validity of an evaluation device is the degree to which it measures what is intended to measure."\(^5\) In other words, validity is concerned with determining just what it is the test measures and how well it does so. The term validity must be used specifically, however, since there is no such thing as general validity. A test is valid for some particular purpose and not for others. Cronbach states:

\(^5\) Remmers and Gage, \textit{op. cit.}, p. 122.
A test which helps in making one decision may have no value at all for another—we can't ask the general question "is this a valid test?"—the question is "How valid is this test for the decision I wish to make—or more generally—for what decisions is this test valid?"

It is essential to state the purpose for which a test is being validated before data on validity can be useful. A vocabulary test, for example might be highly valid for measuring intelligence or for predicting future success in an academic curriculum, but might be of no use at all in predicting the success of future airplane pilots.

In addition to the fact that there is no such thing as general validity, another point is of significance to the consumer of validation information. There are at least four general types of validity, each of which is applicable to a particular type of validation research. A concise definition of each is given from the Technical Recommendations of the American Psychological Association below:

**Content validity.**

Content validity is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn. Content validity is especially important in the case of achievement and proficiency measures.

---


7 American Psychological Association, *loc. cit.*
**Predictive validity.**

Predictive validity is evaluated by showing how well predictions made from the test are confirmed by evidence gathered at some subsequent time. The most common means of checking predictive validity is correlating test scores with a subsequent criterion measure.\(^8\)

**Concurrent validity.**

Concurrent validity is evaluated by showing how well test scores correspond to measures of concurrent performance or status. Concurrent validity and predictive validity are quite similar save for the time at which the criterion is obtained. It should be noted that a test having concurrent validity may not have predictive validity.\(^9\)

**Construct validity.**

Construct validity is evaluated by investigating what psychological qualities a test measures, i.e., by demonstrating that certain explanatory constructs account to some degree for performance on the test.\(^10\)

**Intelligence.** In the definition given above of construct validity, (or concept validity as it is sometimes called) it was stated that the psychological quality a test measures is investigated. On an intelligence test, therefore, it is essential to define, at least as far as possible, what it is

\(^8\) Ibid.

\(^9\) Ibid.

\(^10\) Ibid.
that the test is testing, that is, what intelligence is. Freeman sums up the many different definitions into several main groups. The first group might consist of those definitions who place major emphasis on the adaptation or adjustment of the individual to his environment or to the several aspects of his environment. Such definitions stress the adaptability of the individual to new problems and new situations and the ability to react effectively and appropriately in these new situations. Definitions in this category would also stress creativity. The second group of definitions stress the ability to learn. These definitions would have special reference to the kind of thinking stressed in school.

A third group of definitions stress the ability to carry on abstract thinking, including dealing with symbols and concepts that are removed from the realm of the concrete.

It should be emphasized that the above categories of definitions are not mutually exclusive. Most of the definitions put forward by the authorities in the field include ideas from several of Freeman's categories, but the emphasis is usually placed mostly on one.

Since the WISC is used as the criterion measure in the present study, Wechsler's own definition might be presented.

---

Here. He defines intelligence as: "... the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment."^{12}

A final attempt at defining intelligence is presented by George D. Stoddard who goes about the problem in a slightly different way by saying:

Intelligence is the ability to undertake activities that are characterized by (1) difficulty, (2) complexity (3) abstractness, (4) economy, (5) adaptiveness to a goal, (6) social value, (7) the emergence of originals, and to maintain such activities under conditions that demand a concentration of energy and a resistance to emotional forces.\(^{13}\)

In intelligence test making, the test is the working model of the definition or theory, so that the type of test a person makes depends on his prior conception of just what intelligence is. At times, the reverse has been true--the test is made, administered, standardized, etc. before an idea of intelligence can be formulated. Chapter III will discuss such an aspect of the problem.

### III. LIMITATIONS AND DELIMITATIONS

**Delimitations**

Several types of validity were described above. All are different ways of answering the question of how well the

---


test being evaluated does its job of measuring the particular quality it purports to measure. The American School Intelligence Tests purport to measure the kind of intelligence (if it can be said that there are different kinds of intelligence) needed to do work in school. In the present study the problem was approached in one way. The general definition of validity given above could be broken down into two parts. The first might be "the degree to which the test measures anything". Information gathered to determine how well the test tests anything usually comes under the heading of "reliability information". Some information on the reliability of the Wechsler Intelligence Scale for Children and the American School Intelligence Tests is given in the next chapter. The second part of the definition might be concerned with "what it is that the test purports to measure". This second part is probably the most difficult of the two to determine. The difficulty is partly explained by the trouble encountered in the various attempts at defining intelligence. If it were agreed that intelligence is the quality which enabled one to succeed in school, a criterion has been set up which can be fairly well checked. Tests can be given to a selected group and then later school success of the group can be compared with the earlier test results. If those who scored highest on the previous test also did the best in their later school work; and those who did less well on the first test also did
less well in future work, the definition would appear to be justified. The kind of validity described in such a way would be predictive validity. It would show the power of a given test to predict future accomplishment in a given field. Also, the predictive validity has been arrived at empirically, that is, the information has been gained in a practical situation without relying entirely on statistics or entirely on the principles of logic. However, it has still not been concretely proven that the first test given was testing intelligence. It predicted future academic success, so since the high correlation tended to agree with the previously stated principle (that intelligence is the quality that enables one to succeed in school) it could be logically assumed that the quality originally tested was, indeed, intelligence.

In testing the construct (or concept) validity of a test, not only the test itself, but the underlying quality affecting test scores (in the present case, intelligence) is being tested. In the present study the problem was approached in one of several possible ways. The particular approach used will be discussed later in the paper. It should be sufficient to state here that a statistical type of study was used. That is, it was felt that the Wechsler Intelligence Scale for Children measures intelligence and is a good predictor of school success. A high correlation between test scores on the Wechsler and the American School Intelligence
Tests made by the same group of children would tend to show that both tests measure essentially the same thing. Such an approach is statistical and not empirical in nature. That is, there was no attempt to validate the American School Intelligence Tests as predictors of future achievement in school by administering the tests, deriving the scores and then waiting several years to see how well the children who were tested actually did in school. It was felt that if there appeared to be a high correlation between the WISC and the ASIT, then the ASIT would be as good as the WISC in predicting future academic success. The fact that the present study was concerned with one of several types of validity and that one type of validity was arrived at in one of several possible ways shows one of the delimitations imposed upon the present study.

A second major delimitation resides in the sampling used. The ASIT is intended to be used in any school system in the United States and the norms are intended to be representative of the general school population in grades four through nine in the United States. The present study utilized one school (the Laboratory School of Indiana State Teachers College, Terre Haute, Indiana) from one school system (the Vigo County School Corporation) of the many thousands in the county. As explained later in the paper,
not all of the students in the appropriate grades were used. The study, then was delimited to include a small group.

Limitations

The delimitations described above are also limitations on the study. In addition, others might be listed. First of all, the fact that the tests were not all administered scored and recorded by the same person. They were administered over a considerable period of time (some of the WISC scores were as much as three years old). The tests were all designed to be objective, however, which means that the same results should be arrived at no matter who administered or scored them. Some differences can result, nonetheless. It might have been of help to divide the groups tested to see whether those of generally lower intelligence as measured by the WISC (say below 80 IQ) generally scored higher, lower or about the same on the ASIT. A similar study might have been done with those who scored 120 IQ or better on the WISC to see how closely their IQ's correlated with ASIT IQ's. An item analysis might have been done on the ASIT to check the validity of the various items and a factor analysis might have been done to see how well the various types of items (verbal analogies, etc.) correlated with subscores or total scores of the WISC and with the full ASIT scores.
IV. ORGANIZATION OF REMAINDER OF THE THESIS

The remainder of the study will include a discussion of the methods used in determining validity, a description of the group used, a description of the criterion measure used, the technique and results of the study, a summary of the study, and conclusions based on the findings of the study.
CHAPTER II

THE GROUP STUDIED AND THE MEASUREMENTS USED

I. THE GROUP USED

The group used as the population sample in the study included 117 children enrolled in grades 4 through 9 at the Laboratory School at Indiana State Teachers College, Terre Haute, Indiana during the school year 1960-61.

The Laboratory School includes Kindergarten through twelfth grade and also includes a nursery school. In addition, there are two special classes in the school, one for physically handicapped and one for educable mentally retarded children. The school is staffed and operated by Indiana State Teachers College, but is a local school unit of the Vigo County School Corporation and draws its students from the regular attendance area around it. The school is located near the "down-town" section of Terre Haute. A study by Mary Jo Riggle in 1948 concluded that there appears to be a normal IQ distribution in the school.\(^1\) An analysis

\(^1\) Mary Jo Riggle, An Analysis of Mental Ability of Pupils From Grade One Through Grade Nine of Indiana State Teachers College Laboratory School (Unpublished Master's thesis, Indiana State Teachers College, Terre Haute, 1948), 22 pp.
of a more recent summary of achievement testing which included
IQ scores of all grades revealed a similar distribution.2

The sample included children who had been referred to
the Special Education Clinic at Indiana State Teachers
College for individual psychological examinations which in-
cluded administration of the Wechsler Intelligence Scale for
Children. The children were referred to the clinic for
various reasons. All new students to the Laboratory School
aged 8 to 9 or older have a WISC administered to them. Other
children are referred for suspected low mental ability and
for possible admittance to a special summer session for the
gifted. Still other children are referred by teachers for
diagnosis of learning difficulties. The group selected
included 22 students in grade four, 20 in grade five, 22 in
grade six, 27 in grade seven, 9 in grade eight, and 17 in
grade nine.

II. THE AMERICAN SCHOOL INTELLIGENCE TESTS

The American School Intelligence Tests3 (hereinafter
referred to as the ASIT) consists presently of two levels,

2 The Laboratory School, Indiana State Teachers College,
Terre Haute, SRA Achievement Testing Summary, April, 1959.
(Mimeograph Form).

3 Willis E. Pratt, M. R. Trabue, R. B. Porter and
G. A. W. Stouffer, The American School Intelligence Tests
Intermediate (grades four, five and six), and Advanced (grades seven, eight and nine), with two alternate forms, D and E at each level. The tests are based on the Illinois General Intelligence Scale published in 1926. The tests are almost entirely verbal utilizing seven factors (verbal analogies, arithmetic problems, sentence vocabulary, non-verbal analogies, sentence ingenuity, arithmetic ingenuity, and synonyms and antonyms) arranged in separate, timed subtests. The items are of a multiple choice nature with one correct and four incorrect alternatives. The answers are marked directly on the test booklet which utilizes carbon panels to facilitate scoring. The total testing time for each form of the ASIT is approximately 55 minutes. This includes only the actual testing time, approximately another 12 minutes being used for passing out test booklets, giving directions and explanations. The try-outs showed that approximately 90 per cent of the groups tested finished the test within the time limits. The other ten per cent would probably have completed all the items they are capable of comprehending, so that while the test is timed, it is actually more a power than a speed test.

**Scoring of the ASIT**

The total number of correct answers is used to enter the Mental Age (MA) tables provided in the manual. The conventional manner of computing the Intelligence Quotient
(IQ) is then used utilizing the student's chronological age (CA) in the usual formula for the ratio $IQ = \left(\frac{MA}{CA} \times 100\right)$. A table is provided in the manual to simplify the process. The same raw score can then be used to obtain percentile rank and stanine (a standard score with a mean of 5 and a standard deviation of 2). Thus, after obtaining a student's raw score, it is possible to interpret it in four ways, his Mental Age, his IQ, his relative standing in his own grade by means of the percentile table, and also how he stands in his own particular grade in terms of stanines (in which standard ninth of the group his score falls).

**Standardization of the ASIT**

The ASIT was standardized by administration to approximately 20,000 children in schools of the United States which were intended to be representative of the normal population distribution of the country. Tests were administered in 35 states of the country including six hundred sixty-five classrooms in 70 school systems. Both rural and city schools were used and in most cases, other normative data were available. The standardization included administration of both forms of each level of the ASIT to the cases along with the already standardized American School Achievement Tests so that normative data for both series of tests were realized. The manual reports that the pupils used in the standardization
sample were selected on a randomized basis and stratified with respect to geographical location, size of community, school enrollment and grade assignments. The sampling also attempted to take into consideration urban-rural distribution, socio-economic status, race, etc. 4

Reliability and Validity of the ASIT

Reliability of the ASIT. The manual of the ASIT 5 reports a coefficient of equivalence of .846 with a standard error of the correlation of .019 between the two forms at the Intermediate level based on 222 cases who received Form D on one day and Form E on another. The mean of the raw scores was reported as 53.96 for Form D and 58.38 for Form E with standard deviations of 16.196 on Form D, 17.327 on Form E and standard errors of 1.087 on Form D and 1.116 on Form E. On 145 cases at the Advanced level, a coefficient of equivalence of .862 with a standard error of the correlation of .02. The mean of Form D was reported as 59.28 with a standard deviation of 14.3 and a standard error of 1.19. Form E showed a mean of 61.66 with a standard deviation of 14.54 and a standard error of 1.21.


5Ibid., p. 27.
Validity of the ASIT. Since the ASIT is a new test and not yet released for general use, no research has been made available to the writer other than that given in the manual. The manual reports coefficients of correlation of .718 with the SRA Primary Mental Abilities Test, for Form D and .822 for Form E with the California Short-Form Test of Mental Maturity, both of which are similar group tests. The manual also reports a correlation of .669 with the WISC at the Intermediate level. At the Advanced level, correlations of .695 with the SRA Primary Mental Abilities Test, .838 with the American School Achievement Test and .653 with school grades. At the Intermediate level, correlations with the American School Achievement Test and school grades are reported as .816 and .749 respectively. Correlations between the Advanced level ASIT and the American School Achievement Tests and school grades are reported as .838 and .653 respectively.

6 Ibid., p. 29.


III. THE WECHSLER INTELLIGENCE SCALE FOR CHILDREN

Format of the WISC

The WISC\(^9\) is a downward extension and revision of the Wechsler-Bellevue Intelligence Scales\(^10\) for adolescents and adults. The present scale presents a variety of items, some from previously published tests. The scales have been completely restandardized for children from 5 to 15 years of age. The items of the WISC are arranged in subtests keeping the content of each homogeneous in nature. The Verbal scale includes Information, Comprehension, Arithmetic, Similarities, Vocabulary with Digit Span offered as an optional test. The Digit Span was designated as optional by Wechsler because of its relatively low correlation with the other Verbal Subtests; it was used in the present study. The Verbal Intelligence Quotient is computed on the basis of five subtests. The Performance Scale consists of Picture Completion, Picture Arrangement, Block Design, Object Assembly, and Coding or Mazes as an optional subtest (not used in the present study) with the Performance Intelligence Quotient again computed on the basis of 5 subtests.


There are some essential differences between the WISC and other well known tests of intelligence that should be noted here. The Scale probably gives more opportunity for observation by the examiner of a subject's behavior during testing than many other like tests. His manner of attacking problems, his tenacity in working out a solution to a problem and his interaction with the examiner can be noted, but not objectively scored.

A second difference between the WISC and other older scales is its refutation of the popular concept of mental age (MA). As Wechsler states, a mental age has often been wrongly assumed to be an absolute scale and that all persons with the same mental age have identical intelligence levels. "A five-year-old with an MA of 7 has not the same kind of mind as a 10-year-old with an MA of 7". Wechsler derives his IQ's by comparing each subject's performance with scores earned by individuals in his own age group rather than with a composite age group. Wechsler IQ's are computed without the use of MA by using an adaptation of a standard score known as the Deviation IQ. These IQ's are actually standard scores with a mean of 100 and a standard deviation of 15 IQ.

11 Wechsler, op. cit., p. 2.

points which was intended to approximate the 1937 Stanford-Binet\textsuperscript{13} standard deviation. Although the Stanford-Binet standard deviation is not constant at all ages, it fluctuates around a median value of 16. The WISC standard deviation of 15, however, does remain constant at all levels. The use of the term "IQ" to describe such a score can be misleading since it is derived in a way different from that used in age scales. However, the general familiarity of the term IQ justifies its use and also the fact that such scores can be interpreted as IQ's provided their standard deviation is approximately equal to that of other known IQ's.\textsuperscript{14}

\textbf{Scoring of the WISC}

Scaled Scores or standard scores for each of the subtests of the Verbal and Performance Subtests are obtained by entering the appropriate CA tables. These are standard scores with a mean of 10 and a standard deviation of 3. The Verbal subtest standard scores are then combined to total a Verbal score which can be used to find the Verbal IQ by consulting the IQ table provided. After Performance sub-scores and total score are found a separate Performance

\textsuperscript{13}L. M. Terman and Maud Merrill, \textit{The Stanford-Binet Intelligence Scale-Form L-M} (Boston: Houghton Mifflin Co., 1937).

IQ can be found in the same manner. The combination of the total Verbal and Performance scores can then be used to find a Full Scale IQ. Thus a child can be scored in terms of separate subtest standard scores, a Verbal IQ, a Performance IQ, and a Full Scale IQ. In addition, Wechsler, realizing the necessity for MA's for legal definitions, etc., provides a table so that approximate MA standings can be derived. Wechsler had hoped that the subscores could be used in describing various types of mental disorder. However, no objective treatment of the WISC scores has proved able to classify individual patients with a useful degree of accuracy.15

**Standardization of the WISC**

The WISC was standardized on a sample of 100 boys and girls at each age from five through fifteen years. There were 1100 boys and 1100 girls in eleven age groups or a total of 2200 cases, all of whom were white. The selected cases came from four geographical areas of the United States with urban-rural population of the country. The sample attempted approximation of occupational levels of fathers of the children with the occupational distribution in the country.

---

Fifty-five children or 2.5 percent of the population sample were known to be feeble minded. 16

Reliability and Validity of the WISC

Reliability. The WISC manual reports reliability information based on ages $7\frac{1}{2}$, $10\frac{1}{2}$, and $13\frac{1}{2}$. The split-half technique was used with appropriate correction for full length of the test by the Spearman-Brown formula. At age $7\frac{1}{2}$, correlations of reliability of .88, .86, and .92 were realized for Verbal, Performance and Full Scale IQ's. At age $10\frac{1}{2}$, coefficients of .96, .89 and .95 are reported. At age $13\frac{1}{2}$, coefficients of .96, .90 and .94 are shown for the three IQ's respectively. The subtests were also correlated using scaled score units and the coefficients of correlation are not reported here for reasons of space. Coefficients of correlation ranged from .50 on the Digit Span for age $13\frac{1}{2}$ to .91 for vocabulary for age $10\frac{1}{2}$. 17

Validity of the WISC. The WISC manual reports no information on the validity of the test. However, since 1955,


17 Ibid., pp. 13-14.
many studies have been done on the problem. Orr\textsuperscript{18} reports coefficients of correlation of \(0.88 \pm 0.02\) between the WISC Verbal Scale and the Revised Stanford-Binet Intelligence Scale, \(0.61 \pm 0.06\) between the WISC Performance Scale and the Stanford-Binet and \(0.77 \pm 0.04\) between the WISC Full Scale and Stanford-Binet scores based on 50 cases.

Stempel reports a correlation of \(0.68\) between WISC Full Scale IQ's and full scale scores on the Primary Mental Abilities Test by Science Research Associates. She also reports a correlation of \(0.68\) between the WISC Verbal IQ and the PMA Verbal Meaning IQ.\textsuperscript{19}

Grace Altus reports correlations of \(0.71\) between the Verbal IQ of the WISC and the Language score on the California Test of Mental Maturity (short form), \(0.67\) between the Performance IQ of the WISC and the Non-Language IQ of the CTMM and \(0.77\) between the Full Scale WISC IQ's and the total CTMM IQ's.\textsuperscript{20}


\textsuperscript{19}Stempel, loc. cit.

Why the WISC was Selected.

The WISC is rapidly gaining acceptance as an outstanding clinical tool. It appears that, on the basis of the extensive research carried out on the WISC since 1955, it is at least as good a test as the Stanford-Binet Scale in the prediction of school success. Further, it is mechanically an easier test to administer, a point that is well liked by clinicians. The separate Verbal and Performance IQ's enable more diagnostic interpretations than the single Stanford-Binet IQ. Also, the separate subtests and subscores show some possibilities for use in diagnosing mental disorders, brain pathologies, etc., although nothing really concrete has been established through use of the subscores thus far.
CHAPTER III.

TECHNIQUES AND RESULTS OF THE STUDY

I. INTRODUCTION

In arriving at information on the construct validity of a particular test, there are two steps the investigator must take.

First the investigator inquires: From this theory, what predictions would we make regarding the variation of scores from person to person or occasion to occasion? Second, he gathers data to confirm these predictions.\(^1\)

In validating an intelligence test, for example, it is desired to hold as many factors which might influence scores to a minimum. Reasons for the variability of scores are then investigated. On an intelligence test, it is desired that all factors such as previous educational training, cultural background, reading ability and the like are equalized. The only other factor affecting various performances on a test is then designated as "intelligence" or "mental ability". However, on even the best constructed and standardized tests, the coefficient of correlation between the test and the criterion, even when corrected for attenuation, shows some

\(^1\)American Psychological Association, Technical Recommendations for Psychological Tests and Diagnostic Instruments, supplement to The Psychological Bulletin 51:13, March, 1954.
value less than unity.\(^2\) What are some of the reasons for such a finding? The first reason is that in a study such as the present one, the criterion used is not the criterion, but only an indirect measure of it. In other words, the ASIT, for example, is being validated against the WISC which is a criterion of intelligence. The test is actually being validated, then, against a criterion of the criterion. This means that the test is being moved farther from the actual criterion of intelligence. As Goodenough says:

If the new test is frankly called "a test for predicting scores on the test of intelligence," a criterion is thereby chosen about the nature of which there can be no possible misunderstanding.\(^3\)

For example, the group test might have been called "a group test of mental ability". Had this been done, however, the individual test could no longer be regarded as a wholly satisfactory criterion measure. It is merely another approach to the same problem.\(^4\) However, in a study such as the present one, the individual test is set up as the criterion. A low correlation, then, would not be interpreted as meaning that the criterion is an imperfect

\(^2\)Ibid., p. 16.


\(^4\)Ibid., p. 218.
index of intelligence, but rather, the test is regarded as an imperfect index of the criterion.\(^5\) It must be remembered that statistical construct validity is not the validity of the test, but only one type of validity and a test should be validated in all possible ways, pertinent to its claims. A low coefficient of correlation between the test and the criterion measure need not make the test useless. If a test shows a high correlation to future achievement, after an empirical study of predictive validity, it might be found that the first criterion has been ill chosen or that the test being validated has been misnamed. In other words, it is one thing to require that a test be relevant to the criterion and quite another to require that it be equivalent to the criterion.\(^6\) Thus, it might be found that a test is invalid as a measure of intellectual ability but useful as a predictor of school success. "The question forces the test maker to take a clear position. He can develop a superior measure of educated skills, or he can develop a superior measure of unschooled abilities."\(^7\) Either can add to the understanding of pupils.

It has often been said, somewhat facetiously, that intelligence is "what the test measures." Such a definition

\(^5\)American Psychological Association, \textit{op. cit.}, p. 16.

\(^6\)\textit{Ibid.}, p. 15.

can certainly be defended on the basis that it does indeed give an operational definition with which to work. Some of the earliest makers of intelligence actually worked in such a way. People have grown accustomed to thinking of weight as what is measured by a scale designed to measure the pull of gravity on a given mass. Similarly, for the purpose of the study, at least, intelligence was defined, or at least implied to be, what is measured by the Wechsler Intelligence Scale for Children. More specifically, since the ASIT purports to measure the kind of intelligence required to succeed in the schools, it was implied that the WISC measures the particular kind of intelligence needed to succeed in school (if it can be said that there are different kinds of intelligence). A close correlation between the ASIT and WISC, then, would support the hypotheses that (1) the ASIT and the WISC measure the same thing (namely, intelligence) and further, that (2) the ASIT is as good a predictor of school success as the WISC and therefore (3) the ASIT could replace the WISC as a predictor with a saving in time.

II. SELECTION OF CASES

The cases used in the study were not "selected" in the usual sense. Rather the availability of test scores dictated the selection of the pupils used. The files of the Special Education Clinic at Indiana State Teachers College were used
to obtain WISC scores obtained by pupils who also had had
both forms of the ASIT administered to them. The ASIT's
were administered in the appropriate grades in the fall of
1960 while the WISC's were administered at widely scattered
times by the writer and others. Most of the WISC scores
were older than those of the ASIT, but none were obtained
more than three years ago. The final group used in the
study included 117 children, including 22 in grade four, 20
in grade five, and 22 in grade six who all were tested by
the Intermediate form of the ASIT, 27 children in grade
seven, 9 in grade eight, and 17 in grade nine who were tested
by the Advanced form of the ASIT.

III. METHOD OF DETERMINING CORRELATIONS

After grouping the scores (IQ's) obtained on both
forms of the ASIT and the three IQ's of the WISC by each
pupil and then plotting these on a correlation diagram
(scattergram) the coefficients of correlation were computed
by using the Pearson product-moment method. Since the
scattergram was used, the formula utilized was
\[ r = \frac{\bar{x} \cdot \bar{y}}{\frac{S.D. \cdot S.D.}{N}} \]
the coefficient of correlation when deviations are taken from
the two assumed means. Standard deviations were computed by
use of the formula
\[ S.D. = \sqrt{\frac{\sum x^2}{N} - \bar{x}^2} \]
The standard
error of the means were computed by the formula
\[ S.E. = \sqrt{\frac{S.D.}{N-1}} \]
and the standard error of the correlation was computed by the formula \( S.E. r = \frac{1 - r^2}{\sqrt{N - 1}} \). The raw scores for Forms D and E at each level were combined, averaged and the resulting score used to determine MA and IQ in the usual way. The resulting IQ was then termed a "composite IQ".

IV. INTERPRETATIONS OF THE CORRELATIONS

The question of how high a coefficient of correlation should be cannot be answered specifically. Probably the best answer is "as high as possible". Another more satisfactory answer is "about as high as similar coefficients". As was shown in Chapter II, correlations between the WISC and other group tests run between .67 and .77. Cronbach says that it is rare for a validity coefficient to run above .60 although such a correlation shows far from perfect correlation.  

V. CORRELATIONS BETWEEN WISC AND ASIT

Tables I and II, page 34 list the means, standard deviations, and standard errors of the means for both levels of the test expressed in IQ points. Tables III and IV, pages 35 and 36 respectively show the correlations and standard errors of the correlations for both forms at each of the two levels of the ASIT with Verbal, Performance, and Full Scale

\[8\text{Ibid., p. 115.}\]
TABLE I
THE MEANS, STANDARD DEVIATIONS AND STANDARD ERRORS OF THE MEANS FOR BOTH THE WISC AND ASIT AT INTERMEDIATE LEVEL (GRADES 4, 5, 6)

<table>
<thead>
<tr>
<th>TEST</th>
<th>MEAN IQ</th>
<th>STANDARD DEVIATION</th>
<th>STANDARD ERROR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASIT FORM D</td>
<td>89.81</td>
<td>20.4</td>
<td>2.556</td>
</tr>
<tr>
<td>ASIT FORM E</td>
<td>93.87</td>
<td>20.4</td>
<td>2.556</td>
</tr>
<tr>
<td>ASIT COMPOSITE FORM</td>
<td>92.63</td>
<td>20.7</td>
<td>2.594</td>
</tr>
<tr>
<td>WISC VERBAL</td>
<td>99.96</td>
<td>14.0</td>
<td>1.76</td>
</tr>
<tr>
<td>WISC PERFORMANCE</td>
<td>98.8</td>
<td>14.2</td>
<td>1.788</td>
</tr>
<tr>
<td>WISC FULL SCALE</td>
<td>99.81</td>
<td>14.0</td>
<td>1.813</td>
</tr>
</tbody>
</table>

TABLE II
THE MEANS, STANDARD DEVIATIONS AND STANDARD ERRORS OF THE MEANS FOR BOTH THE WISC AND ASIT AT ADVANCED LEVEL (GRADES 7, 8, 9)

<table>
<thead>
<tr>
<th>TEST</th>
<th>MEAN IQ</th>
<th>STANDARD DEVIATION</th>
<th>STANDARD ERROR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASIT FORM D</td>
<td>89.6</td>
<td>19.4</td>
<td>2.71</td>
</tr>
<tr>
<td>ASIT FORM E</td>
<td>91.3</td>
<td>19.1</td>
<td>2.69</td>
</tr>
<tr>
<td>ASIT COMPOSITE FORM</td>
<td>90.92</td>
<td>18.2</td>
<td>2.52</td>
</tr>
<tr>
<td>WISC VERBAL</td>
<td>111.2</td>
<td>16.9</td>
<td>2.366</td>
</tr>
<tr>
<td>WISC PERFORMANCE</td>
<td>103.5</td>
<td>17.6</td>
<td>2.466</td>
</tr>
<tr>
<td>WISC FULL SCALE</td>
<td>104.4</td>
<td>17.9</td>
<td>2.507</td>
</tr>
</tbody>
</table>
IQ's on the WISC. As seen in the tables, the Intermediate forms of the ASIT showed lower correlations, on the whole, than the correlations at the advanced levels.

At the intermediate level, correlations of .65 with Form D, .65 with Form E and .61 with composite form with the WISC Verbal IQ's are shown. Performance IQ's of the WISC correlated .55 with Form D, .48 with Form E, and .46 with the composite form of the ASIT at the same level. Correlations between the Full Scale and the forms of the ASIT were .69 with Form D, .56 with Form E, and .64 with the composite form.

**TABLE III**

COEFFICIENTS OF CORRELATION BETWEEN BOTH FORMS OF THE ASIT AND THE THREE IQ'S OF THE WISC INTERMEDIATE LEVEL-- GRADES 4, 5, 6 (64 CASES)*

<table>
<thead>
<tr>
<th>WISC</th>
<th>Form D</th>
<th>Form E</th>
<th>Composite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal</td>
<td>.650 (.073)</td>
<td>.645 (.073)</td>
<td>.61 (.079)</td>
</tr>
<tr>
<td>Performance</td>
<td>.554 (.086)</td>
<td>.479 (.097)</td>
<td>.457 (.099)</td>
</tr>
<tr>
<td>Full Scale</td>
<td>.691 (.066)</td>
<td>.557 (.086)</td>
<td>.64 (.074)</td>
</tr>
</tbody>
</table>

* Standard Error of the Correlations Shown in Parentheses.
At the advanced level, correlations with the Verbal WISC were .73 for Form D, .70 for Form E, and .71 for the composite form. Correlations with the Performance WISC were .61 with Form D, .67 with Form E, and .66 with the Full Scale WISC. Coefficients of correlation between the WISC Full Scale scores and the ASIT were .66 with Form D, .74 with Form E, and .68 with the composite form.

TABLE IV

COEFFICIENTS OF CORRELATION BETWEEN BOTH FORMS OF THE ASIT AND THE THREE IQ'S OF THE WISC. ADVANCED LEVEL --- GRADES 7, 8, 9 (53 CASES)*

<table>
<thead>
<tr>
<th>WISC</th>
<th>ASIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbal</td>
<td>.728 (.065)</td>
</tr>
<tr>
<td>Performance</td>
<td>.605 (.088)</td>
</tr>
<tr>
<td>Full Scale</td>
<td>.662 (.078)</td>
</tr>
</tbody>
</table>

* Standard Error of the Correlations Shown in Parentheses.
CHAPTER IV

SUMMARY AND CONCLUSIONS

I. SUMMARY

To establish the construct validity of the new American School Intelligence Tests statistically, the test scores of 117 students enrolled in the third through ninth grades at the Laboratory School at Indiana State Teachers College were compared and coefficients of correlation were found between the ASIT test scores and IQ's made on the Verbal, Performance and Full Scale WISC. The population sample included 22 students in grade four, 20 in grade five, 22 in grade six, 27 in grade seven, 9 in grade eight and 17 in grade nine.

There are two forms of the ASIT currently available, Intermediate level--Grades 4, 5 and 6 and the Advanced level, Grades 7, 8 and 9 each with two alternate forms, Form D and Form E. In addition, the writer averaged the raw scores made by each subject on the two forms, used the resulting raw scores to compute MA's and IQ's for a combined or composite form of the two.

The WISC was selected as the criterion measure for the study because of its increasing use for predicting school success and the availability of both Verbal and Performance, as well as Full Scale IQ's.
The group selected as the population sample was composed of children who were attending the Laboratory School at Indiana State Teachers College during the Fall of 1960. The cases were not "selected" in the usual sense, that is, the WISC and the ASIT were not administered to the group expressly for the present study. Rather, the files of the Special Education Clinic at Indiana State Teachers College were utilized to find those pupils who had been administered both forms of the ASIT as well as a WISC within the last three years.

By the use of scattergrams with grouped scores and the Pearson product-moment method, coefficients of correlation between the three WISC IQ's (Verbal, Performance, and Full Scale) and the three IQ's of the ASIT (Form D, Form E, and a composite of the two) were computed. Means, standard deviations and standard errors of the mean and correlations were computed, also and all are reported on tables.

At the Intermediate Level, the ASIT scores correlated with the WISC Verbal IQ's .65 for Form D, .65 with Form E, and .61 with the composite form. Performance IQ's of the WISC correlated .55 with Form D, .48 with Form E, and .46 with the composite form of the ASIT at the same level. Correlations between the Full Scale WISC IQ's and the ASIT forms were .69 with Form D, .56 with Form E, and .64 with the composite form.
At the advanced level, correlations with the Verbal WISC IQ's were .73 for Form D, .70 with Form E, and .71 with the composite form. Correlations with the Performance WISC IQ's were .61 with Form D, .67 with Form E, and .66 with the Full Scale WISC. Coefficients of correlation between the WISC Full Scale scores and the ASIT were .66 with Form D, .74 with Form E, and .68 with the composite form. All correlations were significant at the 1 per cent level.

II. CONCLUSIONS

At the Intermediate level, it should be noted that Form D of the ASIT appears to be a more valid instrument on the basis of correlations with the WISC. Form D correlations with the WISC are higher than any of the Form E or composite forms. However, the low correlation between Form E of the ASIT (a verbal test) and the Performance IQ of the WISC might possibly be more significant. In general, it would probably be correct to state that the administration of both forms of the ASIT would gain nothing in terms of validity although it would probably increase the already high reliability of the test.

Except for the correlation between Form E of the ASIT and the Full Scale WISC, Form D of the ASIT appears to be the more valid instrument. Here, again, in terms of
overall gain in validity and time, it would appear unnecessary to give both forms of the ASIT.

Although it does appear that the ASIT and the WISC (particularly the Verbal WISC) do measure essentially the same thing, and that the ASIT correlates with the WISC about as well as other group tests, the coefficients of correlation are still far from +1.00. Some of the reasons for such a finding might be explored here.

A first source of differences between the ASIT and WISC IQ's might lie in the scoring methods. The WISC uses a deviation IQ which is a standard score with a mean of 100 and a standard deviation of 15 IQ points at all age levels. The ASIT uses a ratio IQ with standard deviations varying from age to age. The smallest standard deviation found by the writer was 18.2 IQ point and the largest 21.75. The ASIT Manual reports Standard deviations as large as 35 IQ points. The IQ scores, therefore, are not strictly comparable. Another reason for lowered correlations between the WISC and ASIT could lie in the methods used for figuring Chronological Age. The WISC drops all days less than 30 and, in the absence of information in the ASIT Manual, the writer rounded all days of CA over 15 to the next higher month. In other words, a child who was 9 years 10 months and 17 days old was given a CA of 9 years, 10 months on the WISC and 9
years 11 months on the ASIT. Such a difference in procedure can account for small variations in IQ.

It is entirely possible that the ASIT might be measuring the quality known as intelligence better than the WISC. At least, it might be measuring the particular type of intelligence required to do school work better than or as well as the WISC in spite of the relatively low coefficient of correlation between them. That is, an empirical study of the predictive validity of both tests, using a follow-up study of the same group of children used in the present study of the same group of children used in the present study would have to be done to determine the actual prognostic value of both instruments.

The ASIT shows coefficients of correlation of .82 to .84 with the American School Achievement Tests and the Intermediate Battery of the ASIT shows a correlation of .75 with present school grades. Such high concurrent validity might be interpreted in two ways. First, the same specific kinds of thinking might be required on both tests. Secondly, it might be decided that the ASIT has been misnamed a test of general intelligence and is actually an achievement test. The first conclusion would have to be verified by further empirical studies. If the second conclusion were found to be true, it would make little difference whether the children in the same class and school were given the ASIT or the
American School Achievement Tests. Either would probably predict about the same. In the case of a child coming from another school, however, an intelligence test would be fairer. A final conclusion would be that while the ASIT shows correlations about as high as other group tests, its primary use will be in screening and should not replace the WISC.

III. RECOMMENDATIONS

There are several recommendations that are made obvious by the findings and conclusions. First, a follow-up study on the predictive validity of the ASIT will be needed. Secondly, it would be desirable for the norms of the ASIT to be given in terms of deviation rather than ratio IQ's. The Technical Recommendations for Psychological Tests of the A.P.A. suggests that standard scores be used unless there is a pressing reason for using other types. The ASIT manual does give stanine values for various scores, but it is probable that the classroom teacher who administers, scores and interprets the tests will use the more familiar IQ rather than the stanines. Thirdly, a similar study to this one should probably be done using larger numbers of cases. A fourth recommendation might be that a factor analysis should be done on the ASIT to determine how well
the various subtests (verbal analogies, etc.) are constructed and also an item analysis might be done to determine the validities of the separate items.
BIBLIOGRAPHY
BIBLIOGRAPHY


The Laboratory School, Indiana State Teachers College, Terre Haute, SRA Achievement Testing Summary, April, 1959. (Mimeograph Form).


Riggle, Mary Jo. An Analysis of Mental Ability of Pupils from Grade One Through Grade Nine of Indiana State Teachers College Laboratory School (unpublished Master's Thesis, Indiana State Teachers College, Terre Haute, 1948).


