5-1-1958

An Analysis and Evaluation of Religious Testing

Stephen A. Watkins

APPROVAL SHEET

This thesis has been approved by the following faculty committee:

First reader: Robert J. Bennett Approved May 14, 1958

Second reader: Nobel V. Sack Approved May 14. 1958

Prof. of Thesis Form: Mildred Wyckoff Approved May 14, 1958

AN ANALYSIS AND EVALUATION OF RELIGIOUS TESTING

by

Stephen A. Watkins

A Thesis

Presented to

the Faculty of the

Western Evangelical Seminary

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Divinity

Portland 22, Oregon

May, 1958

TABLE OF CONTENTS

CHAPTER		PAGE
I.	INTRODUCTION	1
	The Problem	1
	Statement of the Problem	1
	Justification of Study	2
	Delimitation of Study	2
	Definition of Terms	3
	Religious Testing	3
	Psychological Testing and Educational Testing	3
	Correlation	4
	Criterion	4
	Item (Test Item)	4
	Raw Score	4
	Standardized Test	4
	Plan of Procedure	5
II.	HISTORY OF PSYCHOLOGICAL TESTING	6
	Introduction	6
	Biographical Sketches	7
	Alfred Binet	7
	Edward Lee Thorndike	7
	Ernest John Chave	8
	Frank M. McKibben	8
	Testing Until 1900	8
	Rise of Testing (1900-1910)	12
	Development of Testing (1910-1920	13

CHAPTER PAGE

Extension of Standardized Testing (1920-1930) 15

Rise of Evaluation (1930-1940) 16

Extension of Measurement and Evaluation (1940-1950) . 18

History of Religious Testing 19

Summary 22

III. BASIC CONCEPTS INVOLVED IN PSYCHOLOGICAL TESTING 25

The Criteria of a Good Test 25

Validity 25

Reliability 30

Practicality 33

Schedule for Evaluating a Test 35

General Reference Information 36

Validity 36

Reliability 37

Practical Considerations 38

Generalizations Regarding the Problem of Measurement . 40

IV. RELATIONSHIP OF MEASUREMENT TO RELIGIOUS EDUCATION . . . 42

Primary Considerations. 43

Consideration Of Methods 45

Possible Functions of a Testing Program 48

Classroom Functions 48

Guidance Functions 49

Administrative Functions 49

Summary and Conclusions 50

CHAPTER PAGE

V. AN ANALYSIS OF VARIOUS RELIGIOUS TESTS 51

Association Press 51

Laycock Test of Biblical Information 52

Test of Religious Thinking, Form E (Elementary) . . 56

Test of Religious Thinking, Form A (Advanced) . . . 60

C. H. Stoelting Company 64

Test of the Knowledge of Right and Wrong Concerning the Professions 64

Ethical Discrimination Tests 67

Northwestern University Religious Education Tests . . 69

Bible Tests 70

Series B, No. 4, Religious Beliefs 73

My Ideas About Religion 75

The University of Chicago Press 76

A Scale for Measuring Attitude Toward the Church. . 77

Attitude Toward the Bible 80

Attitude Toward God (The Reality of God) Forms A and B 83

Attitude Toward God (Influence on Conduct) Forms C and D 85

Definitions of God 87

VI. EVALUATION AND COMPARISON OF THE TEST ANALYSES 92

A Current Standard Test 92

General Reference Information 92

Validity 93

CHAPTER PAGE

Reliability 93

Practicality 94

Conclusion 95

Tests of Religious Attitude 96

Comparison of Total Points 96

Comparison of Individual Criterion 96

Conclusion 99

Tests of Ethical Discrimination 99

Comparison of Total Scores 99

Comparison of Individual Criterion 101

Conclusion 101

Tests of Biblical Knowledge 101

Comparison of Total Scores 102

Comparison of Individual Criterion 102

Conclusion 104

Summary and Conclusions 104

Comparison of Fields 104

Comparison of Publishers 104

Possible Deductions 105

VII. SUMMARY AND CONCLUSIONS 107

Summary of Important Points 107

History 107

Basic Concepts 108

Relationship of Testing to Religious Education . 109

Analysis and Evaluation 110

CHAPTER PAGE

Conclusions 110

Conclusions as to Reasons for Disappearance . . . 110

Conclusions as to Possible Solutions 111

Suggestions for Further Study 112

BIBLIOGRAPHY 114

APPENDIX A Cole-Von Borgersrode Scale for Rating Standardized Tests 120

APPENDIX B Schedule for Evaluating a Test 123

APPENDIX C Schedule for Evaluating a Test 125

APPENDIX D Otis Score Card for Rating Standardized Tests . 126

APPENDIX E Biblical Test 127

APPENDIX F Test Your "S. S. I. Q." 129

APPENDIX G Method of Inquiry Into Contemporary Religious Experience 131

APPENDIX H Personal Letter From Ross Snyder 134

LIST OF FIGURES

FIGURE		PAGE
1.	Chart comparing tests of religious attitude	97
2.	Chart comparing tests of religious attitude (continued) .	98
3.	Chart comparing tests of ethical discrimination	100
4.	Chart comparing tests of Biblical knowledge	103

CHAPTER I

INTRODUCTION

CHAPTER I

INTRODUCTION

There has developed in the field of secular education the extensive and successful use of testing devices to both measure and improve classroom instruction. These devices, too numerous to name, not only measure achievement, but also aptitude, personality, intelligence, personnel and many other factors. Religious education at one time made an effort to employ these devices but failed.

Therefore, the following questions need to be considered: (1) what were the trends of religious education when religious testing was first used? (2) what general attitude stimulated test development? (3) why was this movement abandoned in religious education while secular education continued to develop and refine the instruments of measurement? (4) is there a practical use for objective measurement in religious education?

THE PROBLEM

Statement of the Problem

The purpose of this research was twofold: (1) to investigate the problem of measurement in the field of religious education, making note of (a) the early development of religious testing devices, (b) their disappearance, (c) to discover reasons for their disappearance with the intent of arriving at a possible solution; and (2) to establish a foundation for future development of modern educational methods in religious education.

Justification of Study.

The premise of this study is that "some kind of measurement or evaluation is inevitable in education."[1] Preliminary research has revealed that at one time religious education proposed and attempted to use refined methods of testing to measure what some call "factors of religion." In secular education these refined methods of testing have remained and are being expanded. They are considered an important part of the total program of education. On the evidence of the lack of available published material and present use it may be assumed that religious education no longer considers these methods important to its program.

This study was made to discover why religious education has discontinued the use of refined testing methods and by discovering the reasons, to lay a foundation for re-establishing the principles of evaluation in religious education.

Delimitation of Study.

The area of measurement in education is far too broad to be fully considered in one thesis. Measurement in its entire scope could easily touch on every aspect of Christian education. Because of the broadness of measurement, this study will be limited to a brief survey of the testing movement in religious education and an analysis of available religious education tests.

---

[1]C. C. Ross, Measurement in Today's Schools (New York: Prentice-Hall, Inc., 1947), p. 45.

DEFINITION OF TERMS.

There are a few terms with a broad meaning which are used interchangeably. The words measurement, evaluation and occasionally testing are used to designate general meanings in relation to instruments, methods and movement. The interchanging use of these terms is significant only in that each has a shade of difference in meaning and all of these meanings need to be included. Measurement, as used in this study, implies the use of some tool such as a test or scale. Evaluation is a more inclusive concept and includes factors other than definite tools. It may be used to describe a comparison of the realized with the ideal. Test is even more limited than measurement. It implies only a type of instrument. All of these terms are used in reference to methods and movement.

Religious Testing.

The term religious testing refers to methods of measurement and evaluation used in religious education.

Psychological Testing and Educational Testing.

Distinction is not made in this thesis between psychological testing and educational testing. The two are so closely related in the development of the movement as a whole that it is impractical to separate them. Educational testing is more limited in scope than psychological testing. The latter is used in counseling, guidance, business personnel and many other areas while educational testing is limited to education.

Correlation.

The term correlation will be used frequently. It is defined as "Relationship or 'going-togetherness' between two scores or measures."[1]

Criterion.

Criterion may be defined as "A standard by which a test may be judged or evaluated."[2]

Item (Test Item).

An item is a single question or exercise in a test.

Raw Score.

The term raw score is mentioned a few times in Chapter Five. It is defined as:

> The first quantitative result obtained in scoring a test. Usually the number of right answers, number right minus some fraction of number wrong, time required for performance, number of errors, or similar direct, unconverted, uninterpreted measure.[3]

Standardized Test.

A standardized test may be defined as:

---

[1]Roger T. Lennon, "A Glossary of 100 Measurement Terms," Test Service Notebook (New York: World Book Company, n.d.), p. 2.

[2]Ibid.

[3]Ibid., p. 5.

> A systematic sample of performance obtained under prescribed conditions, scored according to definite rules, and capable of evaluation by reference to normative information. Some writers restrict the term to tests having the above properties, whose items have been experimentally evaluated and/or for which evidences of validity and reliability are provided.[1]

PLAN OF PROCEDURE.

The logical place to begin is with the history of testing, both secular and religious. An abundance of material is available on the history of secular testing. Virtually no material is available to give a history of religious testing. However, much value may be gained to this study by a thorough understanding of secular history and a brief discussion of religious testing history. Following this the basic concepts of testing will be discussed. Criteria of a good test demanded by modern secular education will be studied and from this a schedule for evaluating tests will be proposed. The available religious education tests which are now in existence will then be analyzed and evaluated with the intent of reaching some conclusions concerning the testing movement in religious education.

---

[1] Ibid.

CHAPTER II

HISTORY OF PSYCHOLOGICAL TESTING

CHAPTER II

HISTORY OF PSYCHOLOGICAL TESTING

INTRODUCTION

It is impossible to understand testing as it is today without first becoming acquainted with the history of its rise and development. The methods used in modern education at the present time have been developed through experimentation and trial and error. Religious education had a history similar to that of secular education. The methods which have had lasting value have been developed through the difficult road of experimentation. The one method retained by secular education that has not been retained by religious education is that of scientific objective measurement and evaluation. Some reasons for this fact may be suggested from a study of the history of secular testing. It was the writer's purpose to discover the general trends of measurement in the field of education and to attempt to relate these trends to the problem as stated in the introductive chapter of this thesis. The procedure will begin with a brief biographical sketch of four men who were leaders in the field of testing--two from the field of secular education and two from the field of religious education. Measurement will then be discussed from its earliest suggestion on through to the present time. The significant trends in psychological testing and measurement will then be summarized.

BIOGRAPHICAL SKETCHES

The first significant pioneer in the field of psychological testing was Alfred Binet (1857-1911) who was a French experimental psychologist. He was connected with the laboratory of psychology and physiology at the Sorbonne and held a position of director for a number of years. He began and published a journal of psychology which expressed the French movement in psychology. Alfred Binet is known generally for his research on human intelligence and specifically for his scales and tests to measure intelligence. He wrote five books in this area.[1]

Edward Lee Thorndike (1874-) has been one of the most important figures in the development of secular educational testing and measurement. He received continual promotion at Teachers college, Columbia University until he became professor in 1904. Later he became director of the division of psychology of the institute of educational research at Teachers college. During World War I Thorndike was chairman of the committee on classification of personnel in the Army. While he served in this capacity, he was instrumental in establishing an efficient system for the classification and distribution of troops. A recent compilation of his writings show a total of over three hundred titles, more than thirty of which are well known books, many in the area of testing. It has been said of him "no other person has touched the measurement movement at so

---

[1]Walter Yust, ed., Encyclopedia Britannica (Chicago: 1947), III, pp. 581, 2.

many points or has contributed so much to it."[1]

Ernest John Chave (1886-) has been prominent in the field of religious education. He has spent the greatest part of his career at the University of Chicago Divinity School as professor and has held a number of important positions on education boards and in organizations.[2] His chief interests have been in progressive educational philosophy and methods. He has published a number of books important in the area of testing including the following: Measurement of Attitudes, Measure Religion, and Personality Development in Children.[3]

Frank Melbourne McKibben (1889-), who was head of the Department of Religious Education at Northwestern University and Garrett Biblical Institute for many years, distributed a number of religious tests. He holds the following degrees: A.B., S.T.B., M.A., and Ph.D. He is the author of a number of books in the field of religious education.[4]

TESTING UNTIL 1900

The earliest mention of any form of test may be found in the Bible:

> And the Gileadites took the passages of

---

[1]Ibid., Vol. XXII, p. 155.

[2]J. C. Schwarz, ed., Who's Who In The Clergy (New York: no publisher given, 1936), p. 216.

[3]Lefferts A. Loetscher, ed., Twentieth Century Encyclopedia of Religious Knowledge (Grand Rapids: Baker Book House, 1955), p. 232.

[4]Schwarz, ed., op. cit., p. 778.

> Jordan before the Ephraimites: and it was so, that when those Ephraimites which were escaped said, Let me go over; that the men of Gilead said unto him, Art thou an Ephraimite? If he said, Nay; then said they unto him, Say now Shibboleth: and he said Sibboleth: for he could not frame to pronounce it right. They then took him, and slew him at the passages of Jordan: for there fell at that time of the Ephraimites forty and two thousand.[1]

China's remarkable stability, according to the sociologist, can be attributed to five factors, one of which is their highly organized examination system. There is a great difference of opinion as to the beginning of this system. One author states that it began as early as 2200 B.C.[2] and another claims it did not have its beginning until 225 B.C.[3]

The system has been described as being "thoroughly democratic, ruthless, invariable, and orthodox."[4] The candidates were confined to isolated cells for hours at a time and compelled to write lengthy papers or treatises on assigned topics.[5]

The oral examination was used in the universities during the medieval times. The University of Bologna by 1219 A.D. and the University of Paris before the close of the thirteenth century

---

[1]Judges 12:5,6, King James Translation.

[2]Harry A. Greene and others, Measurement and Evaluation In The Secondary School (New York: Longmans, Green and Co., 1943), p. 37.

[3]C. C. Ross, Measurement in Today's Schools (New York: Prentice-Hall Inc., 1947), p. 27.

[4]Ibid.

[5]Greene and others, loc. cit.

required degree candidates to defend their theses orally. The first written educational examination probably made its first appearance at Cambridge in England in 1702.[1]

With the increase of students in the 1800's the Boston school committee was forced to change its method of yearly inventory which had included an oral examination of all its pupils. Their first solution was to quiz only the highest grade but this also became impossible as the number of students grew. In 1845 a sub-committee appointed to survey the grammar departments decided to use written examinations. This was the beginning of an awakening to the need for carefully worked out written examinations that were as fair as possible. This incident made a real impression on Horace Mann, who was prominent in education at that time, and he published his comment; thereby putting this before the public.[2]

Credit for devising and using the first objective measures of achievement is given to Rev. George Fisher--an English schoolmaster. His "scale books" were in use in the Greenwich Hospital School as early as 1864. They scaled performance by units of one-fourth from one, representing the highest, to five, representing the lowest degrees of efficiency. It is interesting to note that his work produced no lasting results because "he lived too far in advance of the thought and educational practice of his day."[3] This may hint toward the solution of the problem of this thesis.

---

[1] Ibid., p. 38.

[2] Ibid., p. 39.

[3] Ibid., p. 41.

In America it is noted that Dr. J. M. Rice discovered an idea for comparative tests in 1894 which made him the "real inventor of comparative tests."[1] He administered a list of spelling words in many school systems and analyzed the results. His conclusions were very revealing and quite a shock to the Department of Superintendence of the National Education Association. He was highly criticized and consequently it was not until ten years later that significant attention was brought to the objective method in education testing.[2]

However, apart from educational testing there was development in objective scientific testing and measurement before 1900 in the area of psychology. Galton, with the publication of Hereditary Genius in 1869, brought the scientific study of individual differences into focus.[3] After this the first name to appear in the area of intelligence tests was that of Wilhelm Wundt at Leipzig in 1879. His interest, however, was confined to reaction times and did not include the problem of individual differences. Nevertheless, he did influence the course of psychology considerably and especially the work of other German psychologists who introduced many forms of separate tests which were borrowed later.[4]

Following this limited approach, Alfred Binet entered the scene. Binet was, in a sense, daring and imaginative for he was not afraid of

---

[1] Ibid.

[2] Ibid.

[3] Ibid.

[4] Ross, op. cit., pp. 30, 31.

making errors as he searched for methods to measure "intelligence" even though he never seemed quite sure of what he meant by the term.[1] In 1895 Binet and Henri described ten types of tests which they thought were likely to discriminate between levels of mental ability.[2] Intelligence tests were still vague and general until after the turn of the century.

RISE OF TESTING (1900-1910)

Measurement and evaluation methods were becoming more accepted by this time and the shocking results of J. M. Rice's study began to be recognized. Alfred Binet was beginning to make an impression and in 1905 he introduced the first scale for the measurement of intelligence. This first scale, though crude, still has served as the pattern for all subsequent tests and scales the world over. The 1908 revision was a definite improvement and introduced the "mental age" concept.[3]

Although J. M. Rice's analysis of teaching spelling by comparative tests was introduced earlier, the first actual test for measuring achievement was the Stone Arithmetic Test which was published in 1908 and the first scale was the Thorndike Handwriting Scale announced in 1909 and published the following year.[4]

---

[1] Ibid., p. 34.

[2] Greene and others, op. cit., p. 43.

[3] Ross, op. cit., p. 36.

[4] Ibid., p. 44.

DEVELOPMENT OF TESTING (1910-1920)

Much of the brush had been cleared in the field of testing by this time and it was becoming the current "band wagon" in the area of psychology as well as gaining momentum in education. Beginning about 1910, several studies in rapid succession showed the unreliability of school marks and examinations. Variations which were found in grading and testing gave significance to this research. Marks for German showed 17.1 per cent A's and 8.4 per cent F's while marks in English showed 6.5 per cent A's and 15.5 per cent F's. (Taken from the University of Chicago High School) Does this mean that English is harder than foreign languages? Another example of the subjectivity of measurement and evaluation in education at that time is found in another study. An English composition was given to one hundred English teachers to mark, assigning it a percentage value and also indicating the school grade in which they would expect that quality of work to be done. The percentage values varied from sixty to ninety-eight and estimated grade location from the fifth grade to a junior in college. The composition had been the best found in a survey at Gary, Indiana and was written by a high school senior whose special interest was journalism. Many other startling surveys and studies were conducted during this period of time which promoted the development of testing.[1]

Intelligence testing took great strides during this period due to certain circumstances which will be discussed in the following paragraph. Lewis Terman (1877-) revised the Binet Scale and

---

[1]Ibid., pp. 44-49.

adapted it for use with American children--normal as well as subnormal. This revision appeared in 1916 along with a most complete manual titled The Measurement of Intelligence. This test, now called the "Stanford-Binet", was still limited to use with separate individuals.[1]

In 1917, when the United States was faced with the necessity of training men for position as commissioned officers and regulars, a number of psychologists volunteered their services. Thorndike was elected chairman and they proceeded to devise a method of classifying thousands of men. The old and expensive method of testing each individual separately was discarded and the Army Alpha was created. It was limited to those who could read and understand the English language. So a second test, which was a non-language test for use with illiterates and men who could not read and understand English, was devised and called the Army Beta.[2] The use of these two tests demonstrated three things: (1) the value of mental tests for revealing individual differences in mental ability among people of normal intelligence, (2) the fact that mental testing need not be a costly, individual procedure, and (3) the value of the tests in the practical classification of men. This experience, then, aided greatly the growth of testing and immediately after the war group tests were adapted to the elementary and high school levels.[3]

---

[1]Ibid., pp. 37,38.

[2]Greene and others, op. cit., pp. 43,44.

[3]Theodore Torgerson and Georgia Adams, Measurement and Evaluation (New York: The Dryden Press, 1954), pp. 28,29.

In 1918 Thorndike published one of the most influential articles that has ever appeared on the subject of educational measurements. In it he challenged men to put action to their criticism and improve their methods. According to R. B. Buckingham it was in 1919 that "test-making passed from an amateur to a professional basis." Quantity production had been achieved but tests were not of the highest quality. There was a growing conviction that emphasis should be placed upon quality of work rather than quantity alone.[1]

EXTENSION OF STANDARDIZED TESTING (1920-1930)

With the progress made during World War I in standardizing tests, testing experienced a real improvement in quality. Although personality and character tests had their introduction before this period, they came into wider use in 1921. Voelker devised some actual test situations for measuring character. The Woodworth Personal Data Sheet created in 1917 to measure the ability of soldiers to adjust to trying conditions of army life was adapted by Mathews in 1923 for school use.[2]

An important step in educational testing was the organization of achievement tests into batteries. This took place just before or during the early 1920's. By the administration of a test battery which included subtests on skills in reading, arithmetic, language

---

[1]Ross, op. cit., pp. 49-50.

[2]Ibid., p. 53.

and other subjects, measures could be obtained of children's comparative achievement in these different areas. In this way the achievement of class and school groups could be interpreted by comparison with the average achievement of children in the same age group or grade levels.[1]

During this period well known tests such as the Otis Intelligence Test and the Stanford Achievement Test batteries appeared. According to a recent authority more than one thousand standardized tests appeared before 1930.[2] Another notable development of this period was that of statistical techniques of test analysis.

RISE OF EVALUATION (1930-1940)

During this period two main trends are noted: (1) a more critical attitude toward tests, and (2) the development of new methods of measurement and evaluation. In the late 1920's and the early 1930's, testing was being pushed by enthusiasts who had "seen the light." Tests of intelligence and achievement were administered widely and somewhat indiscriminately. Their results were accepted quickly and uncritically and became a basis for unjustified judgments and actions in reference to individuals. As one authority stated, "Many sins were committed in the name of measurement by uncritical test users."[3]

---

[1]Torgerson and Adams, op. cit., p. 32.

[2]J. Wayne Wrightstone and others, Evaluation in Modern Education (New York: American Book Company, 1956), p. 7.

[3]Robert Thorndike and Elizabeth Hagen, Measurement and Evaluation in Psychology and Education (New York: John Wiley & Sons, Inc., 1955), p. 6.

This same authority continues:

> After a while the pendulum began to swing back. More and more sharply voiced criticisms of tests and of the uses made of tests began to be heard. Heredity-environment discussions became acrimonious. The use of test scores as a basis for classroom grouping became the subject of bitter attack. Criticism was directed at specific tests in terms of their limited scope and their emphasis upon restricted and traditional objectives. It was also directed at the whole underlying philosophy of quantification and the use of numbers of express psychological qualities.
>
> The critical attack had the healthy effect of forcing the test enthusiasts themselves to become more critical of their assumptions and procedures and to broaden their approach to the whole problem of psychological and educational appraisal.[1]

The field of testing expanded with the critical attack and new approach to include many new tests in the areas of personality, interests, attitude and sociometric techniques. Tests such as the Rorschack and others using projective techniques appeared. Anecdotal records were introduced as a technique of evaluation. The use of this type of test and technique showed the effort being made to measure the less tangible objectives of the modern educational program.[2]

In Hildreth's A Bibliography of Mental Tests and Rating Scales, titles of all tests are listed. In the 1933 edition 3,500 titles were listed; in the 1939 edition 4,279 titles were listed; and in the 1945 supplement to the 1939 edition, there were 5,294 titles of tests and rating scales listed.[3]

---

[1] Ibid.

[2] Wrightstone and others, op. cit., pp. 6,7.

[3] Torgerson and Adams, op. cit., p. 30.

EXTENSION OF MEASUREMENT AND EVALUATION (1940-1950)

Several authorities in a publication dated 1943 made this statement concerning the outlook on testing at that time:

> Although educational and mental measurement are still unquestionably in the developmental stages, their merits and appropriate uses are increasingly coming to be recognized. On the other hand, many of their shortcomings are thoroughly realized. The modern emphasis on the guidance function of the teacher and the increased familiarity of teacher and evaluation techniques have resulted more and more in a transfer of measurement functions from the specialists to the teacher and in cooperative attacks of test specialists and subject matter specialists on common problems in this field.[1]

Another authority indicates that:

> The criteria of validity and reliability... were increasingly applied in the selection of tests for school use. The concept of measurement was extended to include appraisal of a variety of outcomes in relationship to the goals of education and the potentialities of the individual. Measurement had grown from a static concept to a dynamic force.[2]

Since World War II three main trends may be noted especially in the elementary school: (1) the more frequent use of standardized tests, (2) a continuing program of test improvement, making testing more useful than it was a decade ago, and (3) the application by classroom teachers of many methods of studying the adjustment and development of children.[3]

---

[1] Greene and others, op. cit., pp. 43,44.

[2] Torgerson and Adams, op. cit., p. 33.

[3] Ibid., p. 44.

HISTORY OF RELIGIOUS TESTING

The bulk of religious tests seems to have been published between the years 1927 through 1933 with few tests published after that time. To verify this the writer has examined nineteen available religious tests to discover the year in which they were published. If a graph were drawn the peak would have been reached around 1930. Indications of this are also found in a number of publications. Goodwin B. Watson (1899-) states that in 1921 tests related to religious education "could have been counted on a speaker's fingers."[1] Watson, at the time of writing his book (about 1926) listed thirty tests related to religious education that were available at that time and seventeen tests which were not yet available[2] and he seemed very optimistic about future mass production of religious tests.[3]

Valuable to this study would be a brief summary of the educational philosophy behind religious testing at this time. Watson was careful not to express the philosophy of that time as his own. He speaks very generally in implying a progressive idea in religious education. He says:

> . . . . the interests of religious education are expanding. At one time, perhaps, Bible knowledge tests would have been sufficient. Today there is a widespread conviction that Biblical material is taught, not for its own sake, but in order to influence attitudes and

---

[1]Goodwin B. Watson, Experimentation and Measurement in Religious Education (New York: Association Press, 1927), p. 67.

[2]Ibid., pp. 70-105.

[3]Ibid., p. 67.

> behavior. Hence, it is demanded that tests be applied to the real objective, the resulting ideals and conduct, rather than merely to the intermediate objective of Bible information acquired. Many would expand their concern beyond everyday morality into the wider concerns of social welfare. They would not care to pronounce upon an individual's religious education until his attitudes toward war, birth control, and extraterritoriality have been made clear. Still others would contend that no particular information or conviction should be regarded as the goal for religious education. They would measure results in terms of the process of living which has been set up.[1]

Notice his use of the pronoun "they" in regard to the goals of religious education. Although he does not relate himself to this progressive idea, he states that tests should be created to measure these goals: thus indorsing that which he refused to speak of as a personal conviction.

Dr. Chave, who was referred to earlier in this chapter, should be noted here for his contribution to this field. He was an earnest advocate of objective scientific measurement and he influenced this area greatly during this period of test popularity. As was noted in the Twentieth Century Encyclopedia of Religious Knowledge, Dr. Chave's chief interests have been in progressive educational philosophy and methods.[2]

The movement at that time (early 1930's) was stimulated by the desire to make use of all the tools used by secular education. The ideas of progressive educational philosophy had been borrowed from

---

[1] Ibid., pp. 67, 68.

[2] Loetscher, ed., op. cit., p. 232.

secular education and then the use of tests as a tool of education was also apparently copied from secular education. The question arises as to whether these methods of measurement were adopted without careful consideration and qualification of religious goals apart from the secular?

Religious tests began to disappear in the early 1940's because there was no demand for them. The leading men in the religious education department seemed to swing away from objective tests. From the University of Chicago, the one time stronghold of progressive education and pioneer work in religious tests, there came a complete reversal of position. Mr. Ross Snyder, chairman of the Religious Education Department of the University of Chicago Divinity School offered this statement:

> . . . . not much has been done for quite a while in the field of tests and measurements in religious education. Partly, everybody has been so busy trying to find new foundations and design new shapes of program. . . . .
>
> Another factor is our discovery that maybe what we can find out by tests - even attitude tests; leaves us with pretty surface information. . . . In order to find depth, paper and pencil tests are misleading. We have come closer to the methods of a therapeutic interview; we have to find ways whereby a "startling encounter" with a situation or a person's faith can take place, and then see what this awakens in a person. . . .
>
> We have done nothing along the lines of measurement since Dr. Chave retired. We have been doing considerable open-ended and depth interviewing, using projective tests. . .[1]

Recent articles in magazines dealing with religious education

---

[1]Statements by Ross Snyder, personal letter, January, 1958, Used by permission.

and chapters in recent books, however, have been devoted to discussion of tools of measurement. In 1943, Gaines S. Dobbins of the Southern Baptist Theological Seminary in Louisville, Kentucky, included a chapter in his book which was titled "Let Us Test Our Teaching."[1] In this chapter he advocated the use of scientific objective measures as well as some subjective methods.

C. B. Eavey in 1953 discussed evaluation methods in two chapters of his book.[2] Ralph Heim of Gettysburg Seminary discussed measuring in the Sunday Church school in his book which was published in 1950.[3] Findley B. Edge suggested various types of tests and areas in which this method might be used in a chapter of his book published in 1956.[4] Although in some areas there is a revival of interest in accurate evaluation methods, to the writer's knowledge there are no tests being published at the present time for use in religious education.

SUMMARY

Ross summarizes the recent trends of the testing movement in two major areas: (1) test construction and (2) use of tests. He states that the emphasis in test construction has changed from

---

[1]Gaines S. Dobbins, The Improvement of Teaching in the Sunday School (Nashville: Southern Baptist Convention, 1943), pp. 135-154.

[2]C. B. Eavey, The Art of Effective Teaching (Grand Rapids: Zondervan Publishing House, 1953), Chapters IX and X.

[3]Ralph D. Heim, Leading a Sunday Church School (Philadelphia: The Muhlenberg Press, 1950), pp. 310-327.

[4]Findley B. Edge, Teaching for Results (Nashville: Broadman Press, 1956), pp. 168-177.

"quantity to quality." He says:

> . . . .test makers as a group no longer unblushingly make the enthusiastic claims for their products that were common even a decade ago. Instead there has grown up a more critical and becomingly modest attitude, which is probably the most characteristic feature of the present trend.[1]

He also states that another recent trend is to extend the field of measurement into new areas to develop new and more specific types of tests.

Ross suggests four possible stages in the use of tests through the years in which testing has been developing. In the first stage there was a lack of interest and a large amount of indifference and suspicion by all except those who invented the tests. The second stage was that of curiosity and the third stage was that of confidence and in some instances overconfidence. The fourth stage may be labeled that of critical caution, not in curtailment in the use of tests, but toward a more critical use of tests with more caution in the interpretation of test scores.[2]

The testing movement in the field of religious education has followed the same general pattern as that found in the field of secular education. Leaders in religious education were quick to catch the enthusiasm for testing and began to produce tests in quantity. Opposition came and testing became almost extinct with an apparent feeling of bitterness on the part of some leaders. Not until the

---

[1] Ross, op. cit., p. 62.

[2] Ibid., p. 63.

present day has there been any reconsideration of the need for objective evaluation in religious education. From the trends which have been discussed in detail in secular education and the trends implied in religious education, the writer will later suggest some conclusions which may have a bearing on the solution to the problem.

CHAPTER III

BASIC CONCEPTS INVOLVED IN PSYCHOLOGICAL TESTING

CHAPTER III

BASIC CONCEPTS INVOLVED IN PSYCHOLOGICAL TESTING

It is the purpose of this chapter to establish a background for the evaluation of tests. A thorough understanding of the criteria of a good test, including a working knowledge of the different kinds of validity, of reliability and the factors that make a test practical, must first be acquired. These concepts, which will be defined later, will then be organized into a schedule for evaluating tests and this schedule will be used to evaluate the religious tests which are available to the writer.

THE CRITERIA OF A GOOD TEST

Validity.

The first and most important question to be asked in regard to any testing procedure is: How valid is it? A definition of the term validity follows:

> . . . . validity is that characteristic which indicates the degree to which the instrument measures or provides a diagnosis of the psychological characteristics that it purports to measure.[1]

That is to say, does the test measure what it is supposed to measure, all of what it is supposed to measure and nothing but what it is supposed to measure?

---

[1] J. Wayne Wrightstone and others, Evaluation in Modern Education (New York: American Book Company, 1956), p. 42.

The following criticisms illustrate the types of weaknesses which may be found in tests: (1) Many tests permit minor or irrelevant factors to influence the score, (2) Cultural factors make it difficult to obtain valid tests of intelligence, (3) Personal habits of responding in borderline judgements dilute some achievement and personality tests, (4) Traditional tests in school subjects have been criticized because they do not show the student's ability in all aspects of the course, (5) The meaning of any test score is also lowered by chance errors of measurement.[1]

In minimizing these weaknesses and determining how valid a test is, there are four things that must be taken into consideration. They are: (1) predictive validity, (2) concurrent validity, (3) content validity, and (4) construct validity.

Predictive validity is the ability of the test to predict future behavior or success. One authority states:

> . . . . the effectiveness of our test procedure will be judged by the accuracy with which test scores predict a suitable measure of later success. This later measure is called a criterion measure.[2]

The term "criterion measure" is important to this discussion. An example of this is the percentage of cadets eliminated from pilot training at different aptitude levels. Of those receiving an aptitude rating of 1 (the lowest), 82.4 per cent were eliminated from pilot training. Of those receiving a rating of 9 (the highest),

---

[1]Lee J. Cronbach, Essential of Psychological Testing (New York: Harper & Brothers, Publishers, 1949), pp. 49, 50.

[2]Robert Thorndike and Elizabeth Hagen, Measurement and Evaluation in Psychology and Education (New York: John Wiley & Sons, Inc., 1955), p. 116.

only 5.5 per cent were eliminated.[1]

The predictive validity can be estimated by determining the correlation between test scores and a suitable criterion measure of success on the job. In the example given on the preceeding page, the correlation coefficient was .49.

According to Thorndike and Hagen it is very difficult to find a suitable criterion measure.[2] However, they state that there are four qualities that are to be desired in a criterion measure. These are, in the order of their importance: (1) relevance, (2) freedom from bias, (3) reliability, and (4) availability.[3] It must first be decided how closely related the test score is to ultimate success such as in a job. This must be done necessarily by professional judgment because of the lack of emperical evidence.

The next question is: How free from bias is this criterion measure? Is it affected by economic status, working conditions or quality of equipment?

Reliability of criterion measure is this: a measure of success on the job must be stable or reproducible if it is to be predicted by any type of device.

Finally, how available is the criterion measure? How long is it going to take to get a criterion score for each individual and how much will it cost? All of these must be considered in predictive validity.

---

[1] Ibid.

[2] Ibid., p. 117.

[3] Ibid., p. 118.

Concurrent validity simply "indicates the correspondence . . . between a measure and the more or less immediate behavior or performance of identifiable groups."[1] In contrast to predictive validity, which is concerned about future prediction, concurrent validity is concerned about the analysis of present behavior as the diagnosis of personality difficulties.

Content validity is concerned with the "accuracy with which the content of the test represents the content of the course of instruction."[2] One history teacher may consider a test invalid because it overemphasizes military events. Another may consider it invalid because the test stresses memorizing of facts. If a scale is designed to measure attitude toward the Bible, a question or statement about the Methodist church would be invalid. The problem, therefore, is to find out how well the test corresponds to the course content or to the goal which has been set for it. According to an authority, the content with which to compare the test may be:

> (1) the content of a particular local text or course of study, (2) the common content of a number of texts or courses of study, (3) the judgment of experts as to what should be emphasized in a course of study, (4) the activities the individual carries out or the errors he makes in the general activities of life, or (5) the knowledge and skills that must be displayed in a particular job.[3]

Closely related to the establishing of content validity is

---

[1] Wrightstone and others, op. cit., p. 44.

[2] Theodore Torgerson and Georgia Adams, Measurement and Evaluation (New York: The Dryden Press, 1954), p. 48.

[3] Thorndike and Hagen, op. cit., p. 112.

construct or concept validity which is concerned with the effectiveness of expression. However, the term "effectiveness of expression" is broad, abstract, and indefinite. The meaning as used here expresses these ideas: Test items must be specific, concrete, and precise and they must consist of definite limited tasks. Concepts such as "good citizenship," "fairmindedness" and "scientific thinking" are too broad and often indefinite. These terms must be analyzed into their "behavioral components."[1] An outline for the analysis of these components is suggested by Thorndike and Hagen.[2] An analysis of this area differs, however, from content validity in that construct validity is concerned with the functions or processes that are applied to content while content validity is concerned with the subject matter acted upon. The same authorities that judge one, however, would also judge the other.

A recent authority summarizes the criteria of validity as follows:

> . . .survey of validity indicates the central importance of a meaningful criterion, and clearly indicates the complex and difficult nature of establishing validity. Satisfactory criterion measures are difficult to achieve. Criteria for judging proficiency in a job, a course of study, or in personal-social adjustment require an immense investment of time and professional skill, despite which the results are often limited in scope and of low reliability. The limitations which these difficulties present lead to the conclusion that obtaining satisfactory criterion data is perhaps the most difficult

---

[1] Ibid., p. 113.

[2] Ibid., p. 112.

and costly aspect of measurement and evaluation.[1]

Reliability.

The next consideration of a good test is the question of reliability. Is the test reliable? A technical definition of reliability is: ". . . an estimate of the degree of consistency or constancy among repeated measurements of individuals with the same instrument."[2] The problem here is not what the test measures; as it is in validity. Rather, how accurately the test measures what it purports to measure. The question is asked: What is the precision of the resulting score? How accurately will it be reproduced if the individual is measured again? Reliability is next in importance to validity. Ross states that "although high reliability is no guarantee that the test is good, low reliability does indicate that it is poor."[3] He indicates that the ideal test tells the truth consistently.

When reliability is discussed, the terms "coefficient of reliability," and "standard error of measurement" are used. These terms may be defined as "the amount by which an obtained score differs from a hypothetical true score."[4]

---

[1]Wrightstone and others, op. cit., p. 46.

[2]Ibid., pp. 46, 47.

[3]C. C. Ross, Measurement in Today's Schools (New York: Prentice-Hall Inc., 1947), p. 83.

[4]Roger T. Lennon, "A Glossary of 100 Measurement Terms," Test Service Notebook, Number 13, (New York: World Book Company, n.d.), p.5.

There are several causes for variation in psychological measurement: (1) Actual difference among individuals in the psychological characteristic being measured, (2) Differences in ability to take a specific test as in the ability to comprehend direction, effects of practice in taking previous tests, and facility in dealing with test exercises, (3) Chance factors such as fluctuations in performance, memory, reasoning, right guesses, etc., (4) Differences of personal temporary nature such as fatigue, motivation, emotional tension, etc. which affect the performance of the individual, (5) Differences connected with external conditions such as heat, light, ventilation, noise, broken pencil, and interference.[1] These factors, some of which can be controlled and others which cannot, should always be taken into consideration.

Several factors involved in the reliability coefficient must also be considered in relation to the above discussion of variation.

1. The reliability coefficient depends on the length of the test.
2. The reliability coefficient depends on the spread of scores in the group studied.
3. A test may give reliable measures at one level of ability, and unreliable measure at another level.[2]

In regard to the length of a test, the more questions asked of the same general type, the more accurate the estimate of ability will be. If a math test contained only one addition problem, a very poor sample of the individual's ability would be obtained.

---

[1] Wrightstone and others, op. cit., p. 48.

[2] Cronbach, op. cit., p. 60.

Short tests can be made more reliable by lengthening them. It must be taken into consideration though, that if the test is too long, an individual may become bored and this reduces the reliability of the test.

Part-scores based on a few test items are of limited value. Interpreting subscores separately may prove to be very unreliable. Cronbach indicates, in relation to this, that when there is a wide spread of scores, the chances for reliability are greater.[1]

A test which is reliable on one level of ability may prove to be unreliable on another. A pitch-discrimination test was given to Navy recruits twice. On the first test those receiving a score of 85 varied on the second test from 72-95. Those receiving a score near the chance level (55) on the first test varied in their scores when retested from 40 to 87.

The above three factors should be taken into consideration when evaluating the reliability of a test. At this time, a brief statement of the methods of determining coefficients may be helpful. There are three major procedures: (1) The administration of two equivalent tests and correlation of the resulting scores, (2) Repeating the administration of the same test and correlation of resulting scores, and (3) subdividing a test into two or more equivalent fractions. The consideration of these methods is not essential to test evaluation. It should simply be noted whether or not the tests being evaluated have gone through the correct procedures in order to establish a reliability coefficient.

---

[1] Ibid., pp. 61, 62.

A test may have a reliability coefficient from 0.00 to 1.00. The question then arises: What is the minimum reliability that is acceptable? Thorndike and Hagen state that there is no general answer to this question. In a test measuring something objective such as mathematics, a reliability coefficient of 0.85 would not be unreasonable while a test to judge leadership ability would do well to have a coefficient of 0.60.

Thorndike and Hagen conclude their discussion of minimum reliability with this statement:

> Thus, a test with relatively low reliability will permit us to make useful studies of and draw accurate conclusions about groups, but relatively high reliability is required if we are to have precise information about individuals.[1]

In conclusion, a test that is valid but which has a low reliability coefficient is a poor test. In judging reliability one must be aware of causes of variation--both external and internal. A test, to be useful, must have an established reliability coefficient demonstrating a tried consistency. However, even though a test is valid and highly reliable, there is one more criterion which must be considered for its use in education.

Practicality.

An educative instrument may be ideal in every aspect and still be virtually useless to a school program if the instrument is impractical.

---

[1]Thorndike and Hagen, op. cit., p. 140.

This is an especially serious consideration to the Christian educational program. Factors which must be considered in the practicality of a testing program are such things as cost, ease of administration, ease of scoring, ease of interpretation and the time involved.

Validity and reliability of different tests being equal, cost may well be the deciding factor in the usefulness of a test. A church school with its limited budget will not accept a new method if its use means greater expenditures. It is a self-evident fact that no matter how useful the publisher feels his test is, church schools many times will reject a method solely because of its expense. In order to be practical, an instrument of measurement must be available at minimum cost.

Ease of administration is an important factor in practicality. With the untrained personnel of the church school, tests which are difficult to administer are of little use. A test difficult to administer would, in many cases, be rejected by students of the Sunday school who are not forced to attend. Also, these circumstances being present--untrained personnel and volunteer attendance--the validity and reliability of the test would drop if administration was difficult.

Scoring of tests, another important factor, must not be too difficult. Some tests, such as the individual Binet examination for intelligence and the Rorschach projective technique for personality appraisal, require expertly trained scorers. Many tests are so complicated in the primary scoring that the person scoring them would have to be acquainted with statistical methods before being able to score the test.

Two things should be considered in examining the scoring procedure for a test: (1) how complicated the scoring is, and (2) the amount of time it takes to score the test.

Ease in interpretation of the test results is another quality that must be evaluated. In order that the interpretation be valid, specially trained personnel have to interpret many tests. Many tests of attitude, aptitude, interest, personality, etc. require such skill. Tests under consideration for use should be studied carefully to determine how easily the results can be interpreted by the available personnel.

With the briefness of the Sunday school class and the short period of time allotted to other units of Christian education, a test cannot afford to be too long. This presents a problem, for at the same time a test that is too short becomes unreliable. However, there are many tests which are divided into subtests and can be given in two different periods. In addition to this, the longer tests have an important effect upon the cooperation, interest, and effort of the individual who is examined.[1]

SCHEDULE FOR EVALUATING A TEST

In order to systematically and objectively evaluate the available religious tests, a standard form or procedure will be suggested here. Only three such forms have been found during the writer's research. Two of these, the "Otis Score Card for Rating Standardized Tests" and the "Cole-Von Borgersrode Scale for Rating Standardized

---

[1]Wrightstone and others, op. cit., pp. 54-56.

Tests," will be placed in the appendix. The third, Rinsland's "A Form for Briefing and Evaluating Standardized Tests," is too complicated to include in this study. Adaptations will be made from these three for use in analyzing religious tests. Credit should also be given to Thorndike and Hagen for their suggested schedule[1] which the writer found especially helpful. This and other suggested outlines will also be included in the appendix.

General Reference Information.

Included in this area is information which does not necessarily affect the usefulness of the test but which identifies it. Included will be the name of the test, the author's name and position (if available), the publisher, date of publication, the cost and any other information helpful in this area such as the statement of purpose.

Validity.

Three sources of evidence of validity will be more or less considered. It should be noted that, of these sources, one or more will usually be absent. These evidences are merely suggested to cover every possible phase of the test being examined.

The first source is from the plan for the test. Does the manual discuss the procedures for determining the scope of the test, and for the particular content to be covered? How closely do the test objectives correspond to objectives desired in particular areas of

---

[1] Thorndike and Hagen, op. cit., pp. 147-149.

religious education? A number of the tests will say very little concerning the procedures for determining validity.

The second source is found in the test blank itself. Do the test items appear appropriate for the objectives that are to be evaluated? Are the items well constructed? It is very important to note whether or not they are free from ambiguity. Do they have attractive wrong-answer choices? This will be the largest source for judging validity although it is not the most important.

Another important source of evidence is from the statistical studies of the test in use. Has the test been correlated with concurrent measures? With what later criterion measures has the test been correlated? How does the evidence concerning statistical validity compare with that for other tests?

Reliability.

Reliability is concerned with the consistency of the test. It is virtually impossible to judge reliability by examining the test. The facts given in the manual are the only source of information.

The first question asked should be: How adequately are data reported? How large and what is the nature of the groups on which the data are reported? Do they indicate the type of reliability coefficient (coefficient of internal consistency, coefficient of equivalence, or coefficient of stability) which has been computed? The coefficient of internal consistency indicates how accurately or

consistently the test measures the individual's performance at a particular moment. This is computed by the split-half method. Coefficient of equivalence is concerned with the same thing and the fluctuations from day to day of the individual and the test. This is computed by using parallel tests. Coefficient of stability is concerned with the score over a period of time. This is computed by the test and retest method. It should be observed whether or not these factors and methods are mentioned in the test manual.

Question number two concerning reliability is: What are the facts on reliability? All data should be listed and if possible, compared with other existing tests of the same type. Items to look for are: age or grade, size of group, mean, standard deviation, and so on.

Practical Considerations.

To the educational system, secular or religious, this area of consideration is important and in many cases may decide the usefulness of the testing method. Included will be factors in administration, in scoring, in interpretation and finally factors in continued use, concluding with a discussion of the format.

Many questions need consideration in evaluating the method of administration. How adequate is the manual? How complicated are the procedures in relation to the student? How complicated is the procedure which must be followed by the examiner? Do the procedures become too involved? How much time is required to administer this

test? Will it fit into the average period of a Sunday school class?

The amount of time required to score the test is important to consider. The answer form and the type of key used for the test will influence the time needed for scoring. Are special skills such as skilled or trained judgement and qualitative interpretation required in scoring the test?

Interpretation is perhaps the most important factor in the practical aspect of a test. Considering that most of the religious tests were primarily designed to be used in areas of religious education such as church Sunday schools, most of the users will have been untrained in test procedures. These questions then confront us: Are the types of norms suggested in the manual appropriate? Are they complete? How readily may raw scores be converted into derived scores? How complete and helpful are the aids to interpretation which are provided in the manual? Are any suggestions made for a remedial program? With these considerations it should be noted whether the purpose of the test is diagnostic or survey. Although not nearly all of these questions will be answered in one analysis, they are suggested to meet any phase present in a test.

Factors in continued use include the following questions: Are there comparable forms? How many? How well is comparability established? Does the cost permit continued use?

Last and one of the least important but significant enough to be included is the consideration of the format of the test. The arrangement of the printed matter should be noted as well as the

legibility of the type and the quality of the paper. Test blanks should be free from distractions.

The questions under each heading are only suggested as a guide to an analysis. In the evaluation of individual tests, many of the questions will not be considered because of the limitations of the test. However, as much as is practical, general consideration of each phase will be incorporated into the discussion. An outline of the schedule discussed above will be found in the appendix. A suggested chart for comparison of different tests by means of number value as adapted from the "Otis Score Card for Rating Standardized Tests" will also be found in the appendix.

GENERALIZATIONS REGARDING THE PROBLEM OF MEASUREMENT.

Some generalizations on tests and measurement as quoted from Ross will prove valuable to this study. They are as follows:

1. Some kind of measurement or evaluation is inevitable in education.
2. All measurement is subject to error.
3. These errors of measurement are due in part to the imperfection in the measuring instruments available.
4. The limitations of the methods used are a still more important source of error in measurement.
5. Teachers and school administrators must not only understand and appreciate the functions of measurement in education, but they must realize more fully the limitations of present measuring instruments.[1]

---

[1]Ross, op. cit., pp. 95-98.

A realization of the truth of these statements is important to a balanced and objective outlook on testing. Tests, if they are misused, can be dangerous. Used carefully and wisely, however, with full realization of their limitations, tests can prove to be a valuable instrument.

CHAPTER IV

RELATIONSHIP OF MEASUREMENT TO RELIGIOUS EDUCATION

CHAPTER IV

RELATIONSHIP OF MEASUREMENT TO RELIGIOUS EDUCATION

It is the purpose of this chapter to establish the relationship of measurement and evaluation methods in general to religious education. The writer has observed that whenever the terms measurement, test, or evaluation are mentioned within the circles of religious education, people frown or quickly state that these should never be adapted into Christian religious education. The general opinion seems to be that any method of measurement or evaluation is completely unrelated to religious education. A few current leaders in Christian education have dared to speak on this subject. Before further study can be justified in this line, the evidence as presented from these religious leaders aforementioned and evidence from other sources such as secular writers, logic, and observation of present methods must be presented here.

Certain primary considerations must first be taken as a foundation for the relationship of measurement and evaluation to Christian religious education. The inevitability of measurement, along with a realistic view of the limitations of any method or instrument of evaluation, comes in the category of primary consideration. This is the primary relationship but it is not complete without a brief survey of the methods now used in religious education. Certain leaders in the area of religious education have found evidence from observation that present methods need re-thinking. Three areas will be benefited if

the present methods are rethought, improved, and utilized. These will be discussed in the concluding pages of this chapter.

PRIMARY CONSIDERATIONS.

The terms "measurement" and "evaluation" as used in this section will imply a broad and general meaning. Evaluation may be defined as "the measurement and appraisal of a comprehensive range of objectives . . . . . through the use of a variety of techniques."[1] The "variety of techniques" may include judgments made from impressions or an informal discussion comparing two individuals. With this definition of terms, four primary considerations will be made.

The first consideration is that some kind of measurement or evaluation is inevitable in education. Ross states that:

> This generalization is amply supported by the history of every recognized science, and of education itself, regardless of whether it is to be classified as a full-fledged science or not.[2]

It is evident that even in religious education evaluation is unavoidable. Ask a Sunday school teacher how her class is doing and she will begin to evaluate the class on the basis of her own criteria. Dobbins states that teaching will inevitably be tested. He says:

> Whether we desire it, or even are aware of it, we and our teaching are constantly being put to the test. Sunday by

---

[1]J. Wayne Wrightstone and others, Evaluation in Modern Education (New York: American Book Company, 1956), p. 7.

[2]C. C. Ross, Revised by Julian C. Stanley, Measurement in Today's Schools (New York: Prentice-Hall, Inc., 1954), p. 131.

> Sunday, month by month, year by year, time and change measure relentlessly our self-improvement, the scope of our knowledge of the Bible and of people, the thoroughness of our preparation and the skill of our practice.[1]

Our second consideration is that all measurement and evaluation is subject to error. Ross states that this is true even in the exact sciences and quotes F. W. Westaway, who in referring to physics and chemistry says: "We may, in fact, look upon the existence of error in all measurements as the normal state of things."[2] Though these errors can be reduced by refining the methods of measurement, they can never be completely eliminated.

The third consideration is that the errors of measurement are due to the imperfection of the measuring instruments and the limitations of the methods used. It is reasonable to assume that if the instrument used is personal judgment and the method used is observation, that the degree of error will be very large. Personal judgment is subject to variation caused by such factors as mood, personal ideas, prejudices, and many other factors. Observation is subject to variation depending on such factors as visibility, environmental factors, change, the object being observed and many other factors. Ross states that "these errors can be reduced but never wholly eliminated."[3]

The fourth consideration which logically follows is this: If

---

[1]Gaines S. Dobbins, The Improvement of Teaching in the Sunday School (Nashville: Southern Baptist Convention, 1943), p. 135.

[2]Ross (Revised), op. cit., p. 132.

[3]Ibid.

the leaders of religious education will realize the inevitability of measurement in education and that all measurement is subject to error depending on the method and the instrument; then present methods should be analyzed; the best methods (new methods if necessary) chosen; and they should be carefully and completely refined to meet the qualifications for a good instrument. This fact needs no discussion--if the first three observations are correct and valid, this step must be taken. In order to improve the methods it is necessary to understand what methods are currently being used in religious education.

CONSIDERATION OF METHODS.

In the chapter on history it was noted that the most common method of testing in early American education was by oral recitation, which was subject to the individual teacher's own judgment. The standard of judgment varied with the teacher. This was brought vividly to the attention of educational leaders in the late 1800's with such shocking results that the fact was rejected for a time. It was shown through a survey how inconsistent the teacher's judgment can be. This was the beginning of the scientific objective testing movement in secular education.

The history of the testing movement in religious education is not as clear and thorough as in secular education. No one has ever made a complete survey of the evaluation methods used in religious education. Evidence of present methods used is limited to the personal observation of this writer and of a few Christian education leaders.

Heim observes that:

> The usual way of answering such questions is in terms of enrollment, attendance, pupil interest, offering, and the like. It has been assumed that a Sunday Church School should produce results in somewhat vague terms of Christian profession, biblical knowledge, church membership, and wholesome character. And it has been assumed further that schools which, for example, are keeping up their attendance will produce those results.[1]

It would appear that if this statement is true, methods of evaluation in religious education are about a century behind those in secular education. Heim has observed this as an existing fact. To this fact we add Dobbin's words:

> In view of the seriousness and high importance of our task as teachers of religion, should we not undertake to replace haphazardness with accuracy, guesswork with system, uncertainty with certainty, at every possible point? Would it not therefore be wise for us to rethink this whole matter of testing and measuring the results of our teaching?[2]

It should be noted here that not all churches are haphazard in the measurement of their goals. However, the writer of this thesis has observed little or no effort toward objective measurement in Christian education.

Secular education, in the early part of this century, faced the fact that evaluation was inevitable and that it was woefully short of valid measurement. As a result, many devices are employed today for the purpose of evaluation. The concept of measurement and evaluation is far broader than methods such as paper and pencil tests.

---

[1]Ralph D. Heim, Leading a Sunday Church School (Philadelphia: The Muhlenberg Press, 1950), p. 310.

[2]Dobbins, op. cit., p. 137.

Some of the methods used by secular education now include: tests, interviews, case studies, case conferences, group discussions, anecdotal records, observation, files of sample materials, questionaires, rating scales, check lists, inventories, logs, diaries and sociograms. Definite programs employing these methods have been set up in most school systems. The teachers are trained in the use of these methods and schools have established a standard of constant improvement.

It is understood by the writer that because the secular schools do this, religious education should not necessarily be expected to follow the same pattern. When a measurement program is mentioned, the statement is usually made that there is no time for such a program in the religious educational program. It is not the purpose of this thesis to set up a program of measurement in religious education. If the reader wishes some help along this line it is suggested that he refer to the book by Eavey, The Art of Effective Teaching.[1] However, it may be generally observed that there is an unwillingness to analyze one's own program of teaching. When asked whether the pupils are getting anything the answer may be, "time will tell." The responsibility of evaluating the effectiveness of present teaching is put off into a vague and indefinite future. The question may be asked: Is religious education afraid to evaluate teaching and learning in terms of the immediate and the objective?

---

[1]C. B. Eavey, The Art of Effective Teaching (Grand Rapids: Zondervan Publishing House, 1953), especially pp. 251-253.

POSSIBLE FUNCTIONS OF A TESTING PROGRAM.

The negative reaction to testing in religious education may be the result of a limited knowledge and view regarding testing. The question of function will aid in showing the relationship of measurement and evaluation to religious education. What areas may be most benefited by such a program? Thorndike and Hagen list possible functions for a secular school testing program.[1] Possible adaptations for religious education will be suggested from this list. Three areas will benefit: (1) classroom, (2) guidance, and (3) administrative.

Classroom Functions.

In the Sunday school classroom, procedure guiding the planning of activities for specific individual pupils will be aided. A measurement instrument such as a Biblical knowledge test will point out remedial students and the particular area in which they are lacking. With this information, compensation could be made in the teaching and possibly special assignment could be made.

Another function is determining reasonable achievement levels for each pupil and evaluating discrepancies between potentiality and achievement. This function will be limited in Sunday school use, however, to determining achievement levels according to norms set by standardization.

---

[1]Robert Thorndike and Elizabeth Hagen, Measurement and Evaluation in Psychology and Education (New York: John Wiley & Sons, Inc., 1955), p. 6.

Guidance Functions.

Guidance functions will facilitate classroom procedure. Tests will help students realize their need by building a realistic picture of themselves. In the upper grades, tests will aid the pupil in making immediate choices. Tests will aid high school juniors and seniors in determining educational and vocational goals. Finally, tests may help the teacher and the parent to understand problem cases within the Sunday school class.

Administrative Functions.

One of the most important areas aided by the proper use of testing is administration. Achievement tests in Biblical knowledge and possibly attitude tests could aid in evaluating curricula and curricular emphasis. It may evaluate any new curricula experiments. Testing may help evaluate teachers both as to their training and aptitude. Finally, testing results, if used carefully and ethically may provide helpful information to the different agencies of the church as well as outside agencies.

More could be added to these functions as time and use provide new knowledge. These and many others are already employed by secular education. If later studies should find the educational method of measurement and evaluation useful to religious education, these areas discussed above will then receive the help they need.

SUMMARY AND CONCLUSION.

It is reasonable to assume that since testing is inevitable in education it should be refined and used. Added to this assumption is the observed fact, supported by a number of authors, which places present religious evaluation methods on a low level of validity and reliability. However, even though the present level is low, sufficient incentive is provided by the suggestion of many functions proposed and adapted for religious education to rethink and improve in this area. With the conclusion that measurement and evaluation are definitely related to religious education, the writer will proceed to analyze available religious tests.

CHAPTER V

AN ANALYSIS OF VARIOUS RELIGIOUS TESTS

CHAPTER V

AN ANALYSIS OF VARIOUS RELIGIOUS TESTS

It was the purpose of this chapter to analyze each available religious education test and to attempt to discover the reason for the rejection of these tests. It should be noted again that almost all religious tests are no longer being published and have not been in print for a number of years. This limits the availability of said tests but the writer feels that the tests that were available for analysis give sufficient representation of religious tests as a whole to arrive at a fairly valid conclusion. The tests were divided into four groups according to the publishers.

ASSOCIATION PRESS.

The Association Press is the official press of the Y. M. C. A. At one time they maintained a "Test and Research Division" which produced a number of testing devices. They discontinued this division and allowed their tests to go out of print as the stocks became exhausted during the depression. At the present time they are publishing one test entitled Roger's Test of Personality Adjustment which the writer of this thesis has not been able to obtain as yet. Three tests (now out of print) are on hand and these will be analyzed. Two of them are different forms of the Test of Religious Thinking and the other is the Laycock Test of Biblical Information.

Laycock Test of Biblical Information[1]

GENERAL REFERENCE INFORMATION. The author of this test is Samuel Ralph Laycock. The only information given concerning the author is that he was from the University of Saskatchewan, Saskatoon, Canada. This test was published by the Association Press of New York in the late 1920's. The exact date is not given. There is no information given concerning the cost. The purpose of the test is to measure Biblical information of pupils eleven years of age and over.

VALIDITY. The test is divided into seven subtests which are spread over four pages. Each test varies a little in method. The first test is multiple choice in which the pupil chooses the correct answer and underlines it. The second test is also multiple choice with an "x" to be placed next to the correct answer. The third test is true and false. The fourth and sixth tests are again multiple choice using the "x" beside the correct answer. The fifth and seventh tests are multiple choice with the correct answer to be underlined. The plan appears to lend itself well to the validity of the test. Choice of the areas of knowledge appears to be standard.

The information required of the pupil seems to be reasonable. Some of the knowledge expected includes knowledge of patriarchs, leaders of the Old Testament period as well as the New Testament period, understanding of well-known passages such as the Ten Commandments, the Beatitudes, the thirteenth chapter of I Corinthians,

---

[1]S. R. Laycock, Laycock Test of Biblical Information (New York: Association Press, n.d.).

the 23rd Psalm, the 19th Psalm, outstanding dissertations by Jesus, the Lord's Prayer, and acquaintance with the important books of the Bible such as Proverbs, Ezekiel, Psalms, Amos, I Corinthians, and the Gospels. The test items range in difficulty from items designed to test pupils of the low average bracket in Biblical knowledge to those pupils with superior Biblical knowledge. There is very little in the test which could be labeled ambiguous. The test items appear to be fair. Wrong answers appear as attractive as right answers which diminishes the possibility of right guesses.

Very little is said about the statistical validity of this test. In its preliminary standardization, the test was given to 1,115 pupils from grades seven through nine. The norms which were established will be discussed under factors of interpretation.

RELIABILITY. Very little is given concerning the reliability of this test. The split-halves method was used to establish the reliability coefficient. It was used on 102 pupils of grades seven, eight and nine and yielded a correlation of .80. This figure in itself appears very excellent; however, there are some discrepancies which appear upon closer examination. Only one method was used of the three methods possible in determining the reliability coefficient and this was used on only 102 pupils. Another interesting observation which remains unexplained is that the grade norm becomes smaller as the pupils grow older. There is no comparison with other tests. The total discussion of reliability is limited to two short sentences.

PRACTICALITY.

Factors in Administration. Is this test practical to administer? The manual contains very explicit instructions for administration. In the directions, items such as the size of the class, writing equipment, atmosphere (attitudes toward testing, etc.), discipline, manner of the administrator and strictness of adherence are discussed. The test procedure is written out to be followed verbatim. Some inconsistency appears between the "General Directions" and the "Test Procedure." Instructions are given to avoid anything that would call attention to the test as a test or to excite the pupils but the term "test" is used frequently in the verbatim instructions in reference to this particular device. The general directions say to be pleasant and sympathetic while the composition of the procedure is cold and mechanical and at times appears hard and unsympathetic.

The instructions are not complex and are easily understood by the pupils. The sample given before each subtest is adequate and simple. The instructions to be given by the examiner are simple and the timing is simple.

Total time allowed on the test amounts to ten minutes so this test could easily be given within the period of a Sunday school class.

Factors in Scoring. Time required to score the test is approximately ten to twelve minutes for each test. After the test has been corrected, very little time is required to compute the pupil's score. About two hours would be required to correct and score all the tests for a class of ten. No special training is required. All of the

answers are objective.

Factors in Interpretation. By using 1,115 cases, three types of norms were established: (1) grade norms, (2) age norms and (3) sex norms. The user of this test would compare his scores with the norms established in these three areas. The validity of the norms may be open to some question. Although standardization was determined from a good representative cross section, not enough cases were used to give valuable norm scores. As was discussed earlier under reliability, some questions arise concerning the norms for this test. However, a "Return Sheet" is provided so that the scores of many users could be used to give a better picture of the norms.

Nothing is given to help interpretation outside of listing these norms. There is no suggestion for a remedial program. Actually, interpretation may prove difficult in this test.

Format. The arrangement of the printed matter on the test sheet is superior to that of the manual. The test is neatly arranged with the title and other items set off by use of different size type. The quality of paper used for the test sheet is also superior to that used for the manual.

CONCLUSION. This test rates high on two of the most important aspects of criteria for a good test. From the evidence examined, the test rates high on validity and practicality. The test, if carefully used, could prove to be very useful in a church situation. A few terms used in the manual would need to be defined and explained if an untrained worker were to use this form.

Test of Religious Thinking, Form E (Elementary).[1]

GENERAL REFERENCE INFORMATION. General information is limited. The author is not given and there is no mention of cost. It was published by the Association Press in 1928. The purpose of the test, as quoted from the manual, is to "discover opinions, judgments, and attitudes regarding God, Jesus, prayer, the church and Kingdom, other religions, life purposes and to measure agreement with a liberal point of view."[2]

VALIDITY. The scope of the test is very broad--perhaps too broad. It endeavors to discover opinions, judgments and attitudes in seven different areas. The test contains a total of four pages (including the title page) and is divided into six parts which touch upon the seven areas mentioned above. These six parts are: (1) Ideas of God and Religious Education, (2) Ideas of Jesus, (3) Ideas of Prayer, (4) Ideas of the Church and Kingdom, (5) Ideas of Other Religions, and (6) Ideas of Life Purpose.

The manual implies that the most favorable position is one which agrees with a liberal point of view but this viewpoint is almost obscure in the test items themselves. They might just as well be used to measure agreement with a conservative point of view. The pupil has adequate opportunity to express his view thus giving a fairly valid measure of his religious thinking. The section

---

[1]No author given, Test of Religious Thinking, Form E (New York: Association Press, 1928).

[2]Ibid., "Manual of Directions," p. 1.

"Ideas of Life Purpose" contains the least opportunity to express a conservative and spiritual position. It should also be noted that the vocabulary used in the test items can be understood quite readily by those for whom the test was devised--ten to fourteen year olds.

Practically no evidence of this test's validity can be obtained by statistics. It is a revision of a former test and was distributed before statistical validity could be established. It is assumed that the validity of this test was improved over the former test.

RELIABILITY. Very little is said in the manual concerning the reliability of this test. It does state that work was being done on the reliability. The former test from which this test was made was given to about one hundred elementary school pupils in protestant Sunday schools. The split-halves method was used to obtain a reliability coefficient which was .76. The producers of this test assume that this new revision is an improvement. It is questionable whether or not a correlation of .76 is accurate if only one hundred pupils were used. No norms were established because of the lack of research prior to the publishing of this test and so nothing can be said concerning such things as standard deviation.

PRACTICALITY. A limited amount of information is given in the manual on such items as interpretation, cost, comparable forms for the same age group. Consideration will be made of the contents of both the manual and test.

Factors in Administration. The instructions given in the manual for administration of this test are extremely brief. All that is

said concerning procedure is "Follow procedure suggested in 'General Directions for All Tests.'"[1] There is no section in the manual bearing that title and no reference is made as to where these "General Directions" may be found. If the administrator is accustomed to giving tests, he will doubtlessly know enough about test procedure to read the directions on the test sheet out loud. Out of four pages in the manual, only a small portion of one page contains material that is of any help in administering the test.

The instructions to the pupils themselves are not complex but there are no sample items given in their directions.

The manual states that ordinarily thirty minutes is enough time to allow for completion of this test. This would crowd the average class in Sunday school and perhaps make the test impractical to give.

Factors in Scoring. The scoring key is objective with a number value given to each answer. No special skill is required of the one scoring the test. However, the values attached to each answer may be subjective and open to question. This is a question of interpretation. The key is crowded onto one page which makes the scoring process slower than it would be otherwise. The form for the answers and the scoring key could be revised to increase the practicality and speed of scoring.

Factors of Interpretation. Very little help in interpretation of the test results is given in the manual. Two suggestions are made:

---

[1] Ibid., p. 1.

(1) to evaluate a group or class showing how many favored each proposed answer, (2) to measure agreement with a liberal point of view by scoring the tests instead of tabulating them as would be done if the first suggestion were followed. Absolutely no norms are given. These were being obtained through a return sheet which was to be filled out by users of the test. The only help given was total possible scores on each part and the added total score of the whole test.

It should be noted that a user of this test wishing to measure agreement with a conservative point of view would need to completely revise the answer values for the test items although the test items themselves are generally valid. The help given in the manual for interpretation is contained in three short paragraphs and nothing is said to distinguish the three things--opinions, judgments, attitudes--which the test was supposed to reveal.

Format. The general quality of the test sheet is superior to that of the manual. The arrangement of the test sheet is good. The first page is devoted to title, score, a few directions and general information for the pupil. This arrangement avoids any distraction from the test items. The following three pages contain the test items which are neatly and clearly arranged. Different sized type is used to indicate the title and subsidiary items.

The manual used typewritten type and is consequently limited to upper and lower case. Its arrangement is average. The quality of paper is inferior.

CONCLUSION. An examination of this test indicates an inferior instrument which is low in validity, very low in reliability, and very low in practicality. Its scope is limited to one point of view which makes it useless to conservative churches without a great deal of revision--especially of the answer values.

Test of Religious Thinking, Form A (Advanced).[1]

GENERAL REFERENCE INFORMATION. As in Form E of this test which was just discussed, this test lacks mention of an author and cost. It was published about 1928 (date of revision) by the Association Press. The statement of purpose is:

> To discover opinions, judgments and attitudes regarding God, Jesus, prayer, the Kingdom of God, the church, Sunday observance, religious education, immortality, religions other than Christianity, and life purposes. To measure the agreement of these opinions with liberal protestant groups.[2]

The purpose, as stated, seems quite broad. The title indicates that this is a test of religious thinking which then must include all these parts. The test is designed for adults and intelligent high school students.

VALIDITY. Although the scope is discussed, the methods which were used to decide the scope are not mentioned. The authors assume a liberal position which lessens greatly the validity of this test

---

[1]No author given, Test of Religious Thinking, Form A (New York: Association Press, 1928).

[2]Ibid., "Manual of Directions," p. 1.

for those who adopt a conservative position. In an attempt to cover religious thinking adequately, the producers of this test cover the nine areas already listed in the statement of purpose. This gives a very excellent cross-section of the pupil's thinking.

Since this test is designed to test adults and intelligent high school students, the vocabulary is fitting and most of the concepts can be understood. There are a few items which could be interpreted as being ambiguous. A "yes" or "no" answer is required of the following question: "Do you think God is interested mainly in having people obey the Bible?"[1] Although this is a question which is intended only to get an opinion from the student, a feeling of hesitancy or uncertainty about the real meaning may be felt. There is insufficient choice in a number of statements concerning ideas of other religions. Knowledge, rather than opinion, was required in a number of areas. For example: "Did Jesus write the Twenty-third Psalm ('The Lord Is My Shepherd')?"[2]

In spite of the fact that this test is designed to measure agreement with a liberal point of view, the items could be used by those with a conservative point of view. As is true with Form E of this same test, the answer values would need revision before the test would be valid from a conservative viewpoint.

This test was still in the process of standardization when it was published and so there are no statistical studies available on it.

---

[1] Ibid., question number 6, p. 2.

[2] Ibid., question number 17, p. 2.

RELIABILITY. Since this is a revision of a former edition and no research had been done on this revision, there are no data reported that are of real value. The data reported from the former edition show a reliability coefficient of .84. The method used to obtain this coefficient was the split-halves method which was applied to tests given to one hundred high school pupils.

PRACTICALITY. Evidence in this test of practicality will be taken from the limited source of the manual. Little is said about administration and interpretation.

Factors of Administration. Exactly the same problems face the user of this form as faced the user of Form E. Two very small paragraphs give very limited and general instructions to the administrator. If untrained personnel were to give this test there would be some confusion as to the correct procedure which might cause some distraction to those taking the test.

The instructions to the pupil appear to be simple. However, there are no sample exercises given at the beginning of each new section.

Time required to administer the test is about forty-five minutes. Although it is wise to have a test which is longer for the sake of validity, this test is too long to be used in the ordinary Sunday school class period. The use of this test would be limited to particular sessions of longer length.

Factors in Scoring. The answer form is very inconvenient in that it requires a maximum of eye movement which slows down the

scoring process. There are six pages to be scored and three pages in the manual containing the scoring key. The arrangement and size of type of the key differs from that of the test sheet. The exact amount of time required to score this test would depend on the individual ability of the person scoring it but in any case, scoring would require too much time.

The key is objective and does not require subjective or qualitative interpretation.

Factors in Interpretation. Very little help is given in the manual on interpretation. Two suggestions are made concerning the interpretation. The first is to discover to what extent a point of view is held. To do this the leader notes the number who answer "yes" to a particular opinion and the number who answer "no" and convert this into percentage. The second suggestion is to measure agreement with a point of view. In this method the scoring key is used and each paper is given a grade which may range from about seventy-five to a total possible score of three hundred. The higher the score, the greater the agreement. Agreement with a liberal point of view is measured if the scoring key in the manual is used. If agreement with a conservative view is desirable, the key would have to be revised by the method discussed by Watson.[1]

The only norms which are given are the ones established on the former edition. The use of these norms is misleading and they should have been omitted.

---

[1]Goodwin B. Watson, Experimentation and Measurement in Religious Education (New York: Association Press, 1927), p. 67.

Format. The discussion on format for the test just preceeding this test will apply to this test also.

CONCLUSION. This test is invalid from the conservative point of view without revision of the scoring key. The test items in themselves need little change. If the test could be shortened without damage to the validity and the scoring key revised in arrangement, it would become more practical. The acceptability of this test in its present form is very poor.

C. H. STOELTING CO.

The C. H. Stoelting Company is one of the leading publishers of tests. They have published tests which were adapted and used in religious education. This company gave lack of demand as the reason for the disappearance of religious tests. Two tests will be examined and evaluated for this study but as they are not directly concerned with religious education, the analysis will be less extensive.

A Test of the Knowledge of Right and Wrong Concerning the Professions.[1]

GENERAL INFORMATION. The author of this test is Matthew Hale Wilson. It was published by the C. H. Stoelting Co. in 1933. The purpose of this test is to measure moral insight. The philosophy of the test is stated as follows:

> The purpose to do that which is right or
> the purpose to do that which is wrong

---

[1]Matthew H. Wilson, A Test of the Knowledge of Right and Wrong Concerning the Professions (Chicago: C. H. Stoelting Co., 1933).

> reveals a fundamental difference in people. A test which measures either of these important attitudes or their combination in an individual reveals the essential character of that individual. In the judgment which he passes upon the work of others the one taking this test reveals himself.[1]

VALIDITY. The manual thoroughly discusses the procedure for deciding the scope of the test. Material was gathered from the code of ethics for different professions which are commonly known. The author assumes that moral insight is measured on the basis of interest in moral problems. The test covers professions such as banker, editor, physician, teacher, clergyman and lawyer. The purpose of this test could correspond well with certain goals of religious education. The test items appear very valid in their composition and very little ambiguity can be detected.

This test was correlated with various intelligence tests in 628 cases. The correlation was .35. In seventy-one cases the test correlated .09 with Wilson's Test of Religious Experience. In seventy-one cases this test correlated .39 with Watson's Social Relations Test. To establish a predictive validity, seventy students were selected for observation. Thirty-five had received a high score and thirty-five a low score. At the end of a designated time the two groups were compared. Of those receiving a high score, twenty-one were in residence and one was dropped for discipline. Of those receiving a low score, twelve were in residence and three had been dropped for discipline. This indicates a high correlation.

---

[1] Ibid., "Manual of Directions," p. 4.

RELIABILITY. The reliability coefficient of this test is .88. The split-halves method was used to establish this coefficient. A total of 526 cases from three different colleges were used. This test seems to be desirable in so far as reliability is concerned.

PRACTICALITY.

Factors in Administration. The instructions are simple but are too difficult to find and too brief. The arrangement of the manual obscures the instructions in test data and interpretation. The instructions to the student are brief and simple. No practice exercise is given. The only statement concerning the length of time required to administer the test is that it should be finished in an ordinary classroom period. The test sheet consists of five pages of 112 items. Since the test is designed for college students and older, it could, perhaps, be completed within a Sunday school class period.

Factors in Scoring. The answer form is simple since the items in the test are answered by circling the letter T (true) or the letter F (false). The scoring key is objective and does not require the one using it to possess any special training for subjective judgment. Each statement is either mainly right (true) or mainly wrong (false). A minimum amount of time would be required to score the tests.

Factors in Interpretation. Virtually nothing is said concerning interpretation. No norms are given. All that is indicated is that those receiving a high grade are more likely to succeed scholastically.

Format. The manual rates very low in this area. It is a typewritten mimeographed paper containing six pages stapled together with one staple in the upper left hand corner. The print in much of the manual is not clear. The quality of the paper is average.

The test sheet is printed with legible type and neat arrangement. There are no distractions among the test items.

CONCLUSION. A great deal of effort was spent on this test to standardize it. The research and scope are thoroughly discussed. The test rates high in both validity and reliability but the manual is woefully lacking in aids of interpretation. The test almost loses its value because of this lack.

Ethical Discrimination Test.[1]

GENERAL REFERENCE INFORMATION. The author of this test is S. C. Kohs. It was published by the C. H. Stoelting Co. No publication date is given. There is no statement as to what the test is intended to measure.

VALIDITY. The test is composed of six exercises covering the areas of social relations, moral judgment, proverbs, definitions of moral terms, offense evaluation, and moral problems.

The test items are objective in construction. However, the choice of the correct answer could easily be debatable. There is no answer value given to near right choices. In the case of moral

---

[1]S. C. Kohs, Ethical Discrimination Test (Chicago: C. H. Stoelting Co., n.d.).

judgment, one thing may be almost as bad as another but there is no second choice. Only one correct answer for each item is given in the scoring key.

The test has not been standardized and what statistics are quoted are tentative. Therefore, nothing can be said concerning statistical validity.

RELIABILITY. No data are reported and nothing is said about the reliability of this test.

PRACTICALITY.

Factors in Administration. The manual is very adequate so far as the administration of the test is concerned and the directions are complete and thorough. Twelve items are listed under the general directions and then each exercise is handled individually. Altogether, six and one-half pages out of a total of eight pages are used for instructions in administration. The directions for the student are brief and simple. A sample exercise is given in each division. The instructions for the examiner are also simple and to the point. This test requires about thirty minutes to administer.

Factors in Scoring. The answers are objective and do not require the examiner to make subjective judgments. The form for the answers and the arrangement of the key facilitate scoring. With a little revision the speed of scoring could be increased. As it is, the correct answers are listed on two pages of the manual but if a scoring card were provided to match the test items in arrangement, a great deal of time could be saved.

Factors in Interpretation. The norms given in the manual are limited and tentative. Barely one hundred cases ranging in age from eleven to adult college students and employment managers were used. Actually, such limited material may be of more harm than help. No suggestions outside of this brief discussion of norms are made to aid interpretation.

Format. The arrangement of the printed matter in both the manual and tests appears to be very excellent. Different sized type is used to set off titles from subordinate items. In the test sheet the directions are clearly separated from the test items which causes a minimum of distraction. The quality of paper used in both cases is excellent.

CONCLUSIONS. Much of the value of this test is lost because it was premature. Lack of standardization, norms, and aids to interpretation greatly limit the usefulness of this test. Nothing is said of the processes of validation or methods used to establish reliability. Due to these facts, this test would seem to be of little value.

NORTHWESTERN UNIVERSITY RELIGIOUS EDUCATION TESTS.

Northwestern University, under the direction of Frank M. McKibben, distributed a number of religious education tests. According to the tests and promotion on hand they published tests in three fields: Bible, religious beliefs, and citizenship. All of these tests are available. The tests for each general area will be

considered as a whole rather than evaluating each test separately. The test on citizenship will not be considered as it is irrelevant to this study.

Bible Tests.[1]

Six separate test sheets are included in this area with each covering a particular area of knowledge or understanding. They are divided into two groups: Series A is concerned with knowledge of the Bible and is called "Information Tests"; Series B is concerned with the understanding of Bible passages and is called "Comprehension Tests." Three areas are covered: (1) the life and teaching of Jesus, (2) Old Testament times and teachings, and (3) the Acts and Epistles. These six tests are covered by one manual.

GENERAL REFERENCE INFORMATION. The tests were published by Northwestern University. No author is given. Series A was published in 1927 and Series B was published in 1929. The manual does not define the purpose of these tests. However, on the first page of the test in the instructions to the student, the following statement is made: "The purpose is to see how well you understand these passages."[2]

VALIDITY. Evidence of validity from the test plan itself is very limited. The procedure followed in making this test is not

---

[1]No author given, Northwestern University Religious Education Tests, Bible Tests, Series A and B (Evanston, Illinois: Northwestern University, 1933).

[2]Ibid., Series B, p. 1.

mentioned and the scope of the content is not discussed. The test objective as mentioned above could correspond with the objectives desired in local situations. It is one of the primary objectives of a Sunday school class to relay information and create an understanding of the Bible. A measuring instrument in this area would be helpful.

The content of the test appears to be valid. However, it is stated that these tests are suitable for grades five through twelve and yet many of the items are too difficult for that age group. Series A, Bible information tests, are objective and seem to be free from ambiguity. Series B, Comprehension tests, do not seem to be influenced by a liberal theological view but remain close to the standard interpretation of Bible passages. In either case it is difficult to guess the right answers.

Nothing whatsoever is said about statistical validity or correlation with any other test.

RELIABILITY. There is no information given on reliability.

PRACTICALITY.

Factors in Administration. The manual is one letter size sheet of paper folded in half to make four pages. One page and a half is given to directions for administration. Another page contains the scoring key and the final page contains norms. The manual is most adequate in directions for administration in spite of its lack of size.

The directions and procedures are simple. Directions to the administrator cover both the details in preliminary preparation and

the actual giving of the test. The instructions to the pupil are direct and simple. A sample exercise is included on each test sheet. Each test sheet contains just one type of item which simplifies the procedure.

Since the Bible area is divided into six separate tests, a great deal of time from the class period is not required in giving the test. The average time required to do the test is not given.

Factors in Scoring. The form for the answers is designed to ease scoring. An "X" is placed beside the correct answer and all the answers are arranged in a direct line down the page. The scoring key, though it does not match the answer forms, is arranged simply. To convert the raw scores into per cents, the number of correct answers is multiplied by a certain number which depends upon the test. This is a relatively simple process. One thing which may hinder the validity of the answers is the fact that in Series A the answers for tests one, two, and three follow the same pattern. For example, in question one of tests one, two and three, the second item is the correct one.

Factors in Interpretation. The only aid to interpretation is the listing of norms. The norms are arranged into grade levels four through twelve. Norms are given for each of the six tests. In a few instances, so few cases were used that the norms are of little significance. The number of cases used ranged from 19 to 1,522. Most of the norms are based on between five hundred and six hundred cases which should be adequate to make them significant. Individual inter-

pretation in local situations would be a comparison with the established norms. In contrast to the Laycock Test of Biblical Information, the norms generally rise with the increase in age.

Factors in Continued Use. The tests, since they are divided as they are, could lend themselves to continued use. The cost at the time of publication was 3 cents per test. This test could be used in conjunction with the Laycock Test of Biblical Information which was evaluated at the beginning of this chapter.

Format. Quality of paper, legibility of type and arrangement of printed matter are excellent. The printed matter forms a logical arrangement. The tests are free from unnecessary items such as general information, instructions and needless titles.

CONCLUSIONS. The most important disadvantage of these tests is the inadequacy of the manual. The user of these tests will have no idea of their validity or reliability. However, tests of this nature, providing the content is satisfactorily valid, are usually helpful.

Series B. No. 4. Religious Beliefs.[1]

GENERAL REFERENCE INFORMATION. The material available on this test is very limited. No manual accompanied the test and no reference is made to either a manual or a scoring key on the order list put out by the publishers. However, it is the opinion of the writer that a manual must be in existence somewhere. A test of this nature would be

---

[1]No author given, Series B. No. 4. Religious Beliefs (Evanston, Illinois: Northwestern University, 1927).

almost entirely useless without a manual. A brief analysis will be made on the basis of the material available.

VALIDITY. The test sheet itself is the only material available for observation and evaluation. A number of the items could easily have a double meaning. It may also be observed that the items indicate the author's point of view although this is not too evident. A number of problems are presented that would be completely new and unreal to the people in some circles. Much of the wording and many of the terms used are unfamiliar to those of a conservative theological position.

RELIABILITY. Absolutely no information is available.

PRACTICALITY.

Factors in Administration. Directions on the test sheet are brief and simple. To the writer they appear to be clear. The approximate time required to administer this test would be about fifteen minutes.

Format. Generally speaking, the format would be rated fairly high. The print is very legible and the arrangement appears to be logical. Test items appear on the last half of the title page which might cause some distraction.

CONCLUSION. Only incomplete information is available due to the lack of the manual so no definite conclusion can be made concern-

ing final validity and reliability. However, from the evidence which is available, the writer would not be willing to use this test.

My Ideas About Religion.[1]

As in the preceeding test, the absence of a manual limits the information which is available for analysis. No significant conclusion will be reached.

GENERAL REFERENCE INFORMATION. No author is mentioned. It was published by Northwestern University with a copyright in 1933. The cost was 3 cents per test. The purpose is not stated but it is assumed that it is to measure the student's attitude toward religion.

VALIDITY. No definite plan can be observed from the test sheet. There are a total of seventy-five items with questions on almost every doctrine of the Bible. Some of these are: heaven, hell, creation, church, Kingdom of God, prayer, Jesus, God, sin, works, baptism, inspiration, and many others.

Evidence from examining the test itself is of little help here without the comments of the manual.

RELIABILITY. There is no information available on reliability.

PRACTICALITY. Only two factors will be discussed and they are administration and format.

---

[1]No author given, My Ideas About Religion (Evanston, Illinois: Northwestern University, 1933).

Factors in Administration. The instructions on the test sheet are very brief but they are clear. There is no time limit and the student is advised not to hurry. The size of the test indicates that it could be administered within the time limits of an ordinary Sunday school period.

Format. The general quality of this test is high. The print is very legible and attractive. Although the test items begin on the title page there is enough division between the two to eliminate any confusion. However, the presence of the title and instructions on the same page with a number of the test items may cause some distraction.

CONCLUSION. It should be noted here again that no significant conclusion can be reached.

THE UNIVERSITY OF CHICAGO PRESS.

A number of tests were published by the University of Chicago Press while Dr. Ernest J. Chave was at the head of the Department of Christian Education. These tests as a whole bear the latest dates of any of the tests evaluated. A number of the tests published by this press are available through a collection compiled into a paper bound book and published in 1939.[1] However, nothing has been done in this field by the University of Chicago since the retirement of Dr. Chave.

---

[1]Ernest J. Chave, Measure Religion (Chicago: The University of Chicago, 1939).

A Scale for Measuring Attitude Toward the Church.[1]

GENERAL REFERENCE INFORMATION. The authors of this test are L. L. Thurstone and E. J. Chave, both of the University of Chicago. The name Thurstone has been connected with tests and measurements almost since their beginning. Dr. Chave also did a great deal of pioneering in this area. His interest was in progressive education. This test was published by the University of Chicago Press in 1929. The purpose is to describe people's attitudes toward the church. The authors felt that no attitude should be right or wrong, favorable or unfavorable to the church. They assume that every one has a right to his own opinion. This seems to be expressive of the progressive educational philosophy.

VALIDITY. The list of opinions in this scale were selected from a much larger number of opinions which were subjected to a series of psychophysical experiments. The opinions were scaled so that they represented an evenly graduated series covering the whole range of opinions from plus to minus.

Further evidence of validity is found in the content of the test. Nothing is said concerning the age range of the test. Judging from the terminology used in the test and the type of statements which are made, the test would be invalid for anyone under adult age. The statement "I regard the church as a parasite on society"[2] uses a

---

[1] L. L. Thurstone and E. J. Chave, A Scale for Measuring Attitude Toward the Church (Chicago: The University of Chicago Press, 1929).

[2] Ibid., p. 3.

term too difficult for children and a concept with which only a thinking adult would be acquainted. The statements represent a good range of thought but seem to be peculiar to the time of publication. However, this is true of only a few statements.

There are no data on statistical validity.

RELIABILITY. Nothing is said in the manual concerning reliability.

PRACTICALITY.

Factors in Administration. The manual is thorough in the areas which it discusses but is incomplete in several important topics which will be discussed later. The manual does not give simple step by step instructions. The manner of giving the test is discussed in several paragraphs. If the test administrator was not trained in testing, he might have some difficulty in understanding the instructions.

The instructions to the students are brief and perhaps a little incomplete. The instructions are composed of three sentences and appear between the title and the student information blanks. No sample exercises are given.

According to the manual, the test usually requires about fifteen minutes to administer. It is emphasized that there is no time limit and that speed does not count. The time element is about the only favorable feature in the administration of this test.

Factors in Scoring. The scoring key is very inadequate. Each of the forty-five items is assigned a number value according to its

nature. The key, rather than taking the statements in their order, puts them in their value order. This means that the scorer will sometimes have to look through the whole key before finding the item for which he is looking. This procedure is extremely complicated and slow.

The scorer is not required to have any technical training outside of a knowledge of math. He should be endowed with patience and an alert mind.

Factors in Interpretation. There is no discussion of norms whatsoever. There is no evidence that this test is even standardized. The suggestions for interpretation given in the manual are, as a whole, too complicated to be useful to a local situation. They suggest six categories ranging from "strongly favorable to the church" to "strongly antagonistic."[1]

Factors in Continued Use. Two other tests of attitude are published by the University of Chicago Press which could be used in conjunction with this test. No mention is made of cost.

Format. The format is attractive to the eye. Several different sizes of type are used. The type is very legible and the quality of the paper good. The instructions are placed on the test in such a way as to be a distraction and this is a disadvantage.

CONCLUSION. This test lacks some of the most important things. The absence of norms, and the lack of data on validity and reliability, make this test almost useless. The complexity of the instructions for

---

[1]Ibid., Test Manual, p. 3.

administration and the methods of scoring make the test extremely impractical.

Attitude Toward the Bible.[1]

GENERAL REFERENCE INFORMATION. This test was prepared by E. J. Chave and edited by L. L. Thurstone. It was published by the University of Chicago in 1933. There are two forms (A and B) of this test, one to be given a period of time after the first is given. The purpose is not actually stated but it may be assumed that it is designed to measure the individual's attitude toward the Bible.

VALIDITY. It should be noted that nothing is said concerning the scope of the test or its validity. Evidence of these will be drawn from the test itself.

The goals of the test could be useful and could correspond to the goals of a local situation. However, they are not as readily adaptable as other goals may be.

There are several things within the items themselves which hinder the validity of the test. First, terms used in almost every item limit the test to college age and above. Nothing is said concerning age in this test. If it were to be used with high school age students, many of the terms such as "tremendous," "supernatural," "fanaticism," "unscrupulous," and others would need to be defined. Another observation on validity is that the statements themselves re-

---

[1]E. J. Chave, Edited by L. L. Thurstone, Attitude Toward the Bible, Forms A and B (Chicago: University of Chicago Press, 1933).

flect the theological position of the authors of the test. The test is too short to be valid. Some of the statements are ambiguous. All of these things greatly lower the validity of this test.

RELIABILITY. No statement is made in the manual concerning reliability.

PRACTICALITY.

Factors in Administration. Instructions in the manual for the administrator are very general. The manner in which the test should be introduced to the students is left to the imagination of the user. If he is untrained, this would be especially difficult with this test.

Instructions given to the student are much clearer than those with the test analyzed just previous to this one. The instructions are outlined and give the procedure step by step. The only thing that might be confusing is the form for the answer. A check is used to show agreement and a cross to show disagreement. The use of plus and minus may be more standard.

The manual states that this test requires about ten minutes to administer. This could easily fit into the religious education program.

Factors in Scoring. Each statement has a number value attached to it. Only the ones with which the student agrees are counted. The median score, which is the person's score, is then found and it is the basis for interpretation. The form of the scoring key is convenient and facilitates the speed of scoring. Since the test is short and the key adequate, only a minimum amount of time is required for scoring.

Factors in Interpretation. The greatest difficulty is in this area. First it should be noted that the number values attached to some of the statements make an interpretation invalid from a conservative point of view. A relatively high value is attached to the statement, "The Bible helps me although I have no illusions as to its supernatural origin."[1] The value attached is 7.9 which, according to the manual, indicates a strong belief and devotion to the Bible. This is a reflection on a liberal view of inspiration.

The authors of this test attach no value to negative answers and allow no opportunity for any neutral position. If the student happened to agree with only one statement, the number value of that one statement would be his total score. This one factor limits the interpretation greatly and makes the test less reliable.

The use of median scale value to reach the student's score may also present a very perverted interpretation.

Finally, there are no norms of any type. No information on standardization is reported. Nothing is said concerning what should be done with those who are extremely prejudiced against the Bible. What little discussion is given concerning the interpretation of this test could be dangerous.

Format. The format as a whole is excellent. The test is nearly free from any distractions such as instructions, unnecessary type and illegible type.

---

[1] Ibid., Test Sheet, Form A, p. 2.

CONCLUSION. The weakest areas of this test are validity and interpretation. These two are about the most important factors to be considered in the use of a test and their absence makes this a very poor test.

Attitude Toward God (The Reality of God) Forms A and B.[1]

GENERAL REFERENCE INFORMATION. The authors are E. J. Chave and L. L. Thurstone. The test was published by the University of Chicago Press in 1931. No age or grade level is given and the purpose is not stated.

VALIDITY. Although each form of this test is short, containing only twenty statements each, a good range of opinions is listed. The student will probably not be influenced by the way the statement is made. There seems to be freedom from any ambiguity.

However, the validity cannot be judged too highly. There is a total lack of any evidence of validity in the manual. This test was not correlated with any other test. The procedure for planning the scope is not discussed. Nothing is said about standardization. The interpretation which will be discussed more fully later, is in danger of being invalid. Evidence from the test itself is not sufficient without statistical evidence of validity.

RELIABILITY. Although reliability is not as important as

---

[1]E. J. Chave and L. L. Thurstone, Attitude Toward God (The Reality of God), Forms A and B (Chicago: The University of Chicago Press, 1931).

validity, it still needs to receive definite attention in the manual. Nothing is said about reliability in the manual for this test.

PRACTICALITY.

Factors in Administration. Instructions given in the manual are extremely general and subjective. If a novice administered this test, he would be confused. The instructions are not complex but they are confusing. The instructions for the student are much simpler and much less confusing. The answers are illustrated and the illustration is repeated just before the test items on the next page. There may be a slight danger of over simplification in the instructions for the student.

The manual states that ten to fifteen minutes are all that is required to administer this test. It emphasizes that this is not a speed test.

Factors in Scoring. The mechanical aspect of scoring is fairly convenient. The form of the key is simple and practical. The person's score is derived by taking the median scale value. The method of scoring in this test is identical to the one just analyzed.

Factors in Interpretation. The usefulness of this test is handicapped because of this factor. First, it should be noted that no norms are given. Secondly, the method for determining a person's score is inadequate. Thirdly, no help is given in interpretation outside of listing seven attitudes. Fourthly, no remedial course is suggested or recommended. There is very little to be said in favor of the interpretation of this test.

Format. The first page contains the title, student information, and test instructions. The test is then turned over for one page of test items. This arrangement gives freedom from distraction. The type varies in size, thus setting off different items. It is also very legible and its general appearance is impressive.

CONCLUSION. The manual is unsatisfactory as it is lacking in data on validity, reliability and norms for interpretation. No assurance is given that this test is reliable. Evidence which is presented indicates that the validity is low. The test is actually of little help in the purpose for which it was designed.

Attitude Toward God (Influence on Conduct) Form C and D.[1]

This test follows almost exactly the same general pattern that the foregoing tests published by the same press have followed. Therefore, little elaboration will be made on this test or the next test to be examined.

GENERAL REFERENCE INFORMATION. The test was prepared by E. J. Chave and L. L. Thurstone. It was published by the University of Chicago Press in 1931. No mention is made concerning the purpose of the test or the age groups for which it was designed.

VALIDITY. There are no data concerning the validity of this test. Nothing is discussed concerning the scope of the test and how

---

[1]E. J. Chave and L. L. Thurstone, Attitude Toward God (Influence on Conduct) Forms C and D (Chicago: University of Chicago Press, 1931).

it was determined. Very little can be discovered from the pattern of the test because of its shortness. It contains twenty-two items, all of which are listed on one page. The test statements show a good range of opinions with a number of the statements classed as neutral and others varying in degree for or against God. With this very limited information it is necessary to judge validity low.

RELIABILITY. Since the only source of information for a discussion of reliability is from the data given in the manual, there can be nothing said concerning this area.

PRACTICALITY.

Factors in Administration. As with the other tests, instructions in the manual for administration are very general. The instructions on the test sheet to the student are identical to the tests previously discussed and are fairly clear. The manual states that ten to fifteen minutes are required to administer this test.

Factors in Scoring. The form of the key and the method of scoring is identical to the tests just discussed. A maximum amount of convenience is obtained with just a short time required to score each test. The method for obtaining a person's score is debatable since the median score is the final score.

Factors in Interpretation. Absolutely no norms are given for any age level. This statement should be sufficient in evaluating the aids to interpretation. Anything said in the test that is not supported by norms is of little value. However, there are several very general

suggestions given in interpretation.

Format. The format is identical to the other tests from the University of Chicago Press.

CONCLUSION. The conclusions reached concerning this test are the same as those reached regarding the preceeding test. The only difference is that this test has a slight change in emphasis. It may be concluded that this test is of little value.

Definitions of God.[1]

GENERAL REFERENCE INFORMATION. This test was prepared by E. J. Chave and L. L. Thurstone. It was published by the University of Chicago Press in 1931. A statement of purpose is as follows: "The purpose of this list is to discover the meanings of the term God to those who use it, and some of the factors that probably have helped to make the present attitude."[2] If this may be called an objective test, then it should be labeled a highly subjective objective test. Evidence of this will be seen later in the analysis. No specific age is mentioned in relation to this test. On the title page where student information is required, a place is made available for the student to underline his age group. This ranges from under twelve to the age group 50-99.

VALIDITY. The nature of this test makes it difficult to

---

[1]E. J. Chave and L. L. Thurstone, Definitions of God (Chicago: The University of Chicago Press, 1931).

[2]Ibid., Test Instruction, p. 1.

determine whether it is valid or invalid. There are no data reported which, in this case, may not be a disadvantage. The test items are well worded and perhaps in some cases too well worded. The terms used would not be understood by many people and especially those under college age. Some of these terms include: fundamentalist, conservative, hypothesis, integrating, interblended, and many other such words. Some of the concepts presented in the statements are unfamiliar to the average person and would be most difficult to understand.

Another area which may be open for criticism is the information required on the first page. The instructions are to underline phrases which best describe the student's religious attitudes and practices. Following are ten items, each containing several phrases expressing different attitudes and practices on certain subjects. Some of the information asked for here may be regarded as highly personal and the student may either hesitate to mark the item or may mark one not expressing his own feeling but that of the leader. The criticism of terminology as mentioned above may also apply here.

In conclusion it may be stated that the only real criticisms are in regard to the construction of the statements and the personal information required.

RELIABILITY. No data are reported on reliability. The nature of the test makes it very difficult to establish any coefficient.

PRACTICALITY.

Factors in Administration. In considering the manual, hesitancy may be expressed in making any judgment. Since the test is so subject-

ive, the instructions are quite subjective also. It cannot be stated exactly that these instructions are complex or confusing and if the administrator studies them carefully they may be adequate but study is required. The instructions to the students become a little too complex for the average person taking this test in a limited class period.

Nothing is said concerning the time factor but judging from the nature of the test (subjective) it would take all of an average Sunday school class period (approximately thirty minutes) and maybe longer to complete this test. Therefore, it may be considered that the time element is a bit impractical.

Factors in Scoring. Scoring this test is highly subjective as the scorer evaluates the student's position on the basis of the information given on the first page and the response to the statements on the second page. A scale based on the scorer's judgement is made to show the position of the student. The average person could not use this test because of the special skill required in scoring and interpreting the results. The time which would be required to score each test would be another disadvantage.

Factors in Interpretation. Interpretation would be made as the test is scored (if the term scored may be used). The response to each individual item is evaluated separately. A place is allowed at the bottom of the page for further comment by the student. The manual makes this statement concerning the interpretation of the test: "From the information given on the first page and the state-

ments checked on the second page the investigator will have to formulate his estimate of the general position of the subject."[1] If the user of this test is not trained to know what to look for, this test may be misleading and a little dangerous. Since this test is subjective in nature and the interpretation is dependent upon the scorer's judgment, the outcome will vary greatly from user to user and this actually decreases the value of the test.

Factors of Continued Use. In relation to this factor the writer would like to quote the following paragraph:

> The greatest value of the list will probably be to furnish data that may be used in correlational studies of different types. Changes in concepts may easily be discovered by having this form checked at different times, after sufficient interval has elapsed to warrant a possible change, or after a definite experience of some kind that might be expected to modify existing attitudes. In using the two scales "Attitude toward God" (The Reality of God") and "Influence on Conduct" it will be distinctly helpful to interpret the results if one has the information from this form on "Definitions of God." Wherever a person's general philosophy of life might be considered in the study of any social attitudes the index derived from this "Definition of God" or from the "Attitude toward God" scales should be of distinct value. The God concept is often the integration of the person's philosophy of life, and is at least a measure of his larger values.[2]

CONCLUSION. This test is invalid and impractical for use in an average church situation. Evaluating it on the basis of the

---

[1] Ibid.

[2] Ibid., p. 2.

criteria for a good test shows it to be inadequate for use in most instances.

CONCLUSION.

The contents of this chapter are summarized and concluded in Chapter VI, "Evaluation and Comparison of the Test Analyses."

CHAPTER VI

EVALUATION AND COMPARISON OF THE TEST ANALYSES

CHAPTER VI

EVALUATION AND COMPARISON OF THE TEST ANALYSES

It was the purpose of this chapter to compare and evaluate the religious tests which were analyzed in the previous chapter. The tests of similar character were grouped together for the sake of evaluation and compared. The evaluation included a comparison of the tests as a whole and then a comparison of the individual components, noting any pattern which appeared.

The average reader will not be able to appreciate the evaluation of these religious tests which were published thirty years ago unless some standard is suggested as a basis for comparison. A strict adherence to the criteria of a good test is demanded today by those who produce and use psychological and educational tests. Therefore, the writer has chosen a measuring device published in 1957 which he will discuss briefly in order to suggest what reasonably should be expected of a standard test or measuring device. Thereby the reader will see more vividly any weaknesses present in the religious tests. This test will be treated only briefly.

A CURRENT STANDARD TEST.

General Reference Information.

The title of this device to be evaluated is Life Experience Inventory.[1] The authors are Gilbert L. Betts and Russell N. Cassel,

---

[1]Gilbert L. Betts and Russell N. Cassel, Life Experience Inventory, (Cincinnati: Published by the authors and distributed by C. A. Gregory Company, 1957).

both of whom had Ed. D. degrees. The manual devotes an entire page to the discussion of these two men, fully giving their qualifications. (It was noted that this did not occur in a single religious test.) This device was published in 1957 by the authors and distributed by C. A. Gregory Company of Cincinnati, Ohio. This test is:

> concerned with assessing certain cogent areas in the life history or experience of an individual, and of providing an objective and quantitative score indicative of this evaluation.[1]

Following this the introduction then thoroughly discusses the purpose and the philosophical background of this instrument.

Validity.

The manual discusses very thoroughly the procedure by which the scope of the test was decided. The validity of purpose is established by discussing the role of life experience in human behavior. The history of the inventory and the development of validation are discussed. The following items are discussed and their validity established: face validity, content validity, status validity, prediction validity, and construct validity. Complete statistics are also given on validity. These different types of validity were never mentioned in the religious tests.

Reliability.

The methods used to establish the reliability of this instrument are discussed. Reliability coefficients are given from eight

---

[1] Ibid., Test Manual, p. 3.

different population samples under each of the three parts of the test. The total average coefficient is .76. The completeness of the data and information stimulates confidence in the reliability of this measuring device. In only four of the thirteen religious tests analyzed was any reliability coefficient given and in these cases, very little was said as to the procedure that was followed in establishing the coefficient.

Practicality.

FACTORS IN ADMINISTRATION. Almost every question of administration is answered in the manual. The test is self-administering and no technical skill is required of the administrator. Anyone who is able to do successful school work at the fifth grade level would have no difficulty in filling out the inventory. The instructions to the student are simple and brief. They are quickly and easily understood. An hour or even less is required to finish this test. This is not too long for public school purposes.

FACTORS IN SCORING. This test can be scored by anyone who is able to count. Absolutely no special skills are required. Each section of the test can be quickly scored in the minimum amount of time.

FACTORS IN INTERPRETATION. Norms are provided for both sexes and for typical and delinquent individuals. They are based on cases numbering from 160 to 1,710. A chart is presented for use in predicting delinquency proneness in individuals. The prediction of delinquency proneness is also given step by step. Interpretation,

aside from the norms and charts, is thoroughly discussed and adequate aids are given.

FACTORS IN CONTINUED USE. Two other tests have been validated for use with scores on this inventory. This provides a broader description of the individual being tested.

FORMAT. It is sufficient to say that the format on this current test is superior to any of those previously analyzed.

Conclusion.

The manual is complete and thorough in every phase of testing. The qualifications of the authors are discussed thoroughly; complete statistics are given in validity, reliability and in interpretation. It has required fifteen large pages to contain this important and valuable information in the manual. This manual, by its completeness, makes the test itself more valuable and usable. This test is not an exception to the rule. It only illustrates a standard which has made secular testing successful and valuable.

In scoring this test, this writer would compute the total points to be 93.5 out of a possible 100. The scoring card method (which will be used in the remainder of this chapter) does not imply that a test scoring 100 is a perfect instrument which is free from all error. It is only a standard for judgment on the items which a test should contain. If a test is to be acceptable, in the judgment of this writer, it should have a score of 90 to 100.

TESTS OF RELIGIOUS ATTITUDE.

The first area to be considered is that of religious attitude. The majority of the published religious tests available were in this area. Each of the nine tests in this area has been evaluated for its individual components and recorded on the score sheets found in Figures 1 and 2. For the sake of discussion, a comparison of the total points will be considered, and then a comparison of individual criterion will be made.

Comparison of Total Points.

In comparing the total number of points for each test, it was noted that no test received more than 50 points. The test rated lowest was Definition of God, published by the University of Chicago, which received a total of 27 points. Religious Thinking, Form E, published by the Association Press received the highest score. Two of the tests, because of incomplete information, could receive no total. The remaining five tests ranged in score from 47 through 49. The average total score was 45.

Comparison of Individual Criterion.

It is significant to note in which individual criteria these tests were rated lowest. The test manuals received an average rating of four out of a possible eight points. No manuals were available with two tests. Validity, for all the tests, received an average of nine points out of a possible twenty. Reliability was especially low since there were no statistics given in the manuals. An average of

ITEMS	Standard number of points	University of C. "Attitude Toward Church"	University of C. "Attitude Toward Bible"	U. of Chicago "Attitude Toward God" (Reality)	U of Chicago "Attitude Toward God" (Conduct)	U. of Chicago "Definitions of God"	
1. Manual	8	4	4	4	4	4	
2. Validity	20	10	5	8	8	5	
3. Reliability	11	3	1	1	1	1	
4. Ease of Administration (21)							
a. Special preparation	7	6	4	3	3	1	
b. Adequate Directions	7	1	4	5	5	3	
c. Time	7	6	6	6	6	4	
5. Ease of Scoring (15)							
a. Objectivity	8	4	4	4	4	1	
b. Convenient Form of Key or Method	4	1	3	3	3	0	
c. Time Required	3	1	3	3	3	1	
6. Ease of Interpretation (20)							
a. Types of Norms	10	0	0	0	0	0	
b. Directions or aids	10	7	6	6	6	3	
7. Format	5	4	4	4	4	4	
TOTAL	100	47	44	47	47	27	

Figure 1, Chart Comparing Religious Attitude Tests

ITEMS	Standard Number of Points		Association Press "Religious Thinking" Form E	Same as last test, Form A	Northwestern U. "My Ideas About God"	Northwestern U. Religious Beliefs	Average of Attitude Tests
1. Manual	8		4	4	?	?	4
2. Validity	20		14	12	?	10	9
3. Reliability	11		6	8	?	?	3
4. Ease of Administration (21)							
a. Special Preparation	7		5	5	5?	5?	4
b. Adequate Directions	7		1	1	4?	5?	3.2
c. Time	7		4	2	6	6	5
							12.2
5. Ease of Scoring (15)							
a. Objectivity	8		6	6	?	?	4
b. Convenient form of Key or Method	4		2	2	?	?	2
c. Time Required	3		1	1	?	?	2
							8
6. Ease of Interpretation (20)							
a. Types of Norms	10		0	0	?	?	0
b. Directions or aids	10		5	5	?	?	5.6
7. Format	5		2	2	4	4	3.5
TOTAL	100		50	48	?	?	45

Figure 2, Chart Comparing Religious Attitude Tests (continued)

three points was received out of a possible eleven. Ease in administration and scoring received about half of the total possible points. One of the most important factors in practicality is interpretation. All of the tests rated especially low in this area; receiving an average of 5.6 total points out of a possible 20. The greatest weakness in interpretation was the lack of any norms. All of the tests except for two which had mimeographed manuals rated high in format.

Conclusion.

The average total score of 45 for tests of religious attitude is a fair judgment which is comparable to the score of 93.5 given to the current test which was discussed at the beginning of this chapter. These instruments are very inadequate and may even be somewhat dangerous. Judgments based on the information furnished from these tests and from the manual's interpretation of scores could be misleading. This does not mean that no attempt should be made to measure religious attitude, but it does show that the work done in this area is far short of what should be done.

TESTS OF ETHICAL DISCRIMINATION.

Comparison of Total Scores.

Examination of Figure 3 indicates that Wilson's Ethical Discrimination Test received a total of 61 points while Koh's test received a total of 59, giving an average total points of 60. Although the outcome of these two tests is about the same, their qualities vary.

ITEMS	Standard Number of points		"Wilson's Ethical Discrimination"	"Koh's Ethical Discrimination"	Average		
1. Manual	8		3	2	2.5		
2. Validity	20		18	10	14		
3. Reliability	11		9	6	7.5		
4. Ease in Administration (21)							
a. Special Preparation	7		6	6	6		
b. Adequate Directions	7		2	6	4		
c. Time	7		4	5	4.5		
					14.5		
5. Ease in Scoring (15)							
a. Objectivity	8		8	8	8		
b. Convenient Form of Key or Method	4		3	2	2.5		
c. Time Required	3		2	2	2		
					12.5		
6. Ease in Interpretation (20)							
a. Types of Norms	10		0	5	2.5		
b. Directions or aids	10		3	3	3		
					5.5		
7. Format	5		3	4	3.5		
TOTAL	100		61	59	60		

Figure 3, Chart Comparing Ethical Discrimination Tests

Comparison of Individual Criterion.

These two tests differ from each other in the quality of individual criterion. Wilson's test was high in validity, reliability, and ease of scoring. Koh's test was high in ease of administration and format. Wilson's test was especially low in the quality of the manual, ease of administration, and ease of interpretation. Koh's test was especially low in validity, reliability and ease of interpretation. It was noted that both tests were low in ease in interpretation. Wilson's test lacked any norms and although norms were given in Koh's test, they were limited and tentative pending further investigation.

Conclusion.

No pattern can be drawn from a comparison of these tests. Although they generally are not weak in the same areas, their weaknesses fall at crucial points. Noting the chart, it was observed that even their strong areas fall too short of the standard. These instruments are supposed to produce information which can be used as a basis for action. However, action taken from the basis of the information given by these tests, could be very misleading and disillusioning to the users of the tests.

TESTS OF BIBLICAL KNOWLEDGE.

Two tests were available in the field of Biblical knowledge--Laycock Tests of Biblical Information and Northwestern University Bible Tests. It should be recalled that the Northwestern University

Bible Tests was a battery containing six tests, all of which were considered under one manual. The score card evaluation of these tests is found in Figure 4.

Comparison of Total Scores.

Perhaps it can be said that this area has produced the most valid tests in Christian Education. The Laycock test received 72 total points and the Northwestern University Tests received 66 total points for an average of 69. However, according to the standard mentioned earlier, this is 21 points below the minimum total points for an acceptable test.

Comparison of Individual Criterion.

Examination of the individual criterion demonstrates the weak-areas of these tests. The weaknesses, as noted in Figure 4, are found in the areas of the manual, reliability, and ease of interpretation. The reasons for a low rating in the manuals differ in each case. The manual for the Laycock test is inferior because of the quality of type and paper. The manual for the Northwestern University tests is inadequate because of its lack of information.

However, it should be noted that these tests are not necessarily as unreliable as the chart shows. They had to be graded low because of either incomplete information or total lack of information.

The norms were furnished in both manuals as aids to interpretation but almost no instructions were present for interpreting the norms. However, in a test of this nature, a complete listing of norms may be adequate for use in interpreting the test results.

ITEMS	Standard number of points		"Laycock Test of Biblical Information"	Northwestern U. Bible tests	Average scores		
1. Manual	8		5	3	4		
2. Validity	20		18	15	16.5		
3. Reliability	11		5	0	2.5		
4. Ease in Administration (21)							
a. Special Preparation	7		7	7	7		
b. Adequate Directions	7		5	6	5.5		
c. Time	7		7	7	7		
					19.5		
5. Ease of Scoring (15)							
a. Objectivity	8		8	8	8		
b. Convenient Form of Key or Method	4		3	3	3		
c. Time Required	3		2	2	2		
					13		
6. Ease of Interpretation (20)							
a. Types of Norms	10		7	9	8		
b. Directions or aids	10		2	2	2		
					10		
7. Format	5		3	4	3.5		
TOTAL	100		72	66	69		

Figure 4, Chart Comparing Biblical Knowledge Tests

Conclusion.

These tests of Biblical knowledge received the highest score of any field. This writer was impressed by their usefulness and general practicality. It may be said that if the manual had furnished more complete information, all of the components would have received a higher rating. It may also be noted that use of test results in the area of Bible knowledge may not have such a discriminating affect as would be found with attitude and ethical tests. However, for the sake of comparison, these tests are still generally inadequate.

SUMMARY AND CONCLUSIONS.

Comparison of Fields.

Upon observing the tests that were available, it appeared that the greatest interest and activity was in the area of religious attitude. Yet the tests in this particular area were the most deficient of any of the tests evaluated. Tests of religious attitude had an average total of 45 points. Ethical discrimination tests, with a total of 60, were evaluated much higher. Biblical knowledge tests received the highest total with an average of 69. The average total score for all the tests combined was 58.

Comparison of Publishers.

Examination of average total points from each publisher is significant. The poorest tests, as far as criteria of a good test are concerned, were published by the University of Chicago Press.

The average score for their tests combined was 42.4. Association Press tests were next with a combined average of 57 total points which was an improvement over the tests from the University of Chicago Press. Northwestern University's tests did not contain sufficient information to compute any average. The tests from the C. H. Stoelting Company did not receive the highest individual total scores but they did have the highest average of total points when they were combined. Their average was 60. This company was and is in the test making business while the other presses only published tests as a sideline.

Possible Deductions.

Three possible deductions may be made from the analysis and evaluation of religious tests.

The first is that a critical examination of these tests shows that results from using them could and probably did have a negative affect on religious education. There is no information available concerning the actual use of these tests in churches but this deduction is reasonable on the basis of the analysis and evaluation.

Second, more interest was evidenced in quantity than in quality, and it was noted several times that statistics and norms were admittedly incomplete and tentative. The primary concern, as implied in some of the manuals, was to get the tests out into circulation and then to gather the needed data. No restrictions were placed on their use. Qualifications of the examiner were never questioned and a great deal was assumed on the part of the authors of these tests.

The institutions represented by the publishers were not as well

equipped to publish tests during the 1920's and 1930's as they are at the present time. The approach to testing is different today and a great deal more care is taken in both the producing and the distributing of tests.

The last deduction will be made in the form of a suggestion. It is possible that religious education never recovered from the careless production and distribution of tests.

CHAPTER VII

SUMMARY AND CONCLUSIONS

CHAPTER VII

SUMMARY AND CONCLUSIONS

SUMMARY OF IMPORTANT POINTS.

History.

Significant facts were mentioned in the area of the history of testing. It appeared that there has been no time in history when tests were not in existence. Men, since time began, have devised means of measurement. It was noted that at a time approximately one hundred years ago, a few men used and advocated objective tests but it was also noted that these men lived before their time and their methods were not accepted and adopted for use. Later, J. M. Rice administered a set of spelling words in a large number of schools and from the results established several facts which were in opposition to common educational belief. This shocked the educational leaders and he was sharply criticised. After 1910, other men conducted innumerable surveys which showed how unreliable school marks were and how unreliable and subjective the teacher's judgments were. These surveys awakened leaders in the field of education to the need for standardized objective testing.

The time at which these things took place was very significant. Conditions were ripe for the development of secular tests. The unrest caused by the surveys, the sudden need for mass classification of personnel during World War I, and the movement for modern progressive methods in education all contributed to this development of testing.

Testing became popular and many different agencies began devising and producing tests; flooding the field with countless numbers of instruments. As a result of the great quantity of tests which were produced and placed in circulation, there were many ill effects and with the flood of tests also came a flood of criticism. This had a healthy effect on the testing movement and caused a deepening in the quality of tests being published. The limitations of tests were then being recognized.

Meanwhile, about the time that secular testing was being criticised, religious testing became popular with certain religious education leaders who began to push the use of measurements in that area. As a result of the popular progressive idea, emphasis was placed on attitude and interest instead of Biblical information. Religious education seemed to be doggedly following secular education as if it were a duty. Then, for reasons which have never been recorded, religious tests disappeared and religious institutions lost interest in a testing program. Not until the last few years has any voice been lifted in favor of the use of objective tests in religious education.

Basic Concepts.

The purpose in discussing basic concepts of testing was not to find evidence to solve the mystery of the disappearance of the religious tests but to establish a background of information in order to adequately understand and evaluate the tests which were published and their weaknesses. The following facts were noted: In studying the material and making an effort to condense it into a few short

pages, the writer of this thesis was impressed with the quantity and quality of criteria of a good test. Many types of validity were mentioned, as was brought out in Chapter III. The meaning, importance and complexity of procedure for establishing reliability was also mentioned. These criteria cannot be established by a brief research. It requires literally years of work to produce an instrument of measurement that is adequate. After all this work has been done, the testing instrument is still weak in some points and subject to different outside influences.

Any number of physical and mental factors may cause a test to be unreliable. A realization of this fact deepens the respect for the work involved in producing a good test and also brings the understanding that any instrument is subject to error.

Another aspect of testing especially important to religious institutions is that of practicality. Until the Christian education leaders in a church can see that these methods are practical and valuable, they will probably completely reject any testing device as unnecessary.

Relationship of Testing to Religious Education.

The most important fact in the relationship of testing to religious education is that measurement and evaluation are inevitable in education--secular or religious. Another important fact was that methods of measurement now used in religious education are grossly inadequate. However, if an objective view was to be taken toward evaluation, nearly every area of Christian religious education would

benefit.

Analysis and Evaluation.

It is sufficient to say that the available published religious tests were inadequate.

CONCLUSIONS.

Conclusions as to Reasons for Disappearance.

In the first place, several important events preceeded the secular testing movement. The demand that student's work be evaluated before promotion and the surveys showing the unreliability of evaluation methods gave impetus to the testing movement in secular education. No evidence of any similar incentive for testing is found in the religious education field. It appears that the leaders decided it was a good idea and so adopted testing without preparation and without understanding the basic principles of testing. Therefore, it was concluded that religious education, as a whole, was not ready for methods of testing.

In the second place, the validity of religious tests was not considered and their reliability was not established. Norms were tentative, pending further study and investigation. The users of the tests were expected to know more than was reasonable as they were not adequately trained. The authors of the tests assumed too much. Therefore, it was concluded that most of the tests produced were published and put into circulation prematurely before adequate research had been made on them.

Instruments of testing in attitudes, interests, ethics and such should be used only by trained administrators. A qualified person should be available to oversee the use of these instruments. Other restrictions would need to be investigated and adopted.

Finally, those concerned with religious education must realize that evaluation is much broader than "tests" as many people think of them. The writer, through contact with just a limited amount of the unlimited material on measurement, has come to realize that evaluation is a part of every day life. This realization has created a respect for the area of measurement which had not been his experience before this research.

SUGGESTIONS FOR FUTURE STUDY.

The writer would like to close this thesis with several suggestions for future study. It is felt by him that this work was necessary for a foundation toward the utilization of the methods of testing. Therefore three suggestions are made.

First, a thorough investigation into the present methods of evaluation need to be made. A survey similar in nature to those made during the period of 1910 to 1916 could be made showing the unreliability of present methods (if they are unreliable). Questionaires and interviews could indicate how Sunday school teachers now evaluate their students and the work done by their students and the work done by the teachers themselves.

Second, a program could be instigated and established to train teachers in the primary essentials of evaluation. The teachers would first need to learn to accept evaluation. Secondly, they would need

to learn how to evaluate their work and their students' by using the many different methods available. A report of experiments in this area would be helpful to the field of religious measurement and evaluation.

Third, tools could be prepared for evaluation and tests produced for religious education. These should be standardized and refined to meet the qualifications of a good test. This area would require the greatest amount of work.

BIBLIOGRAPHY

Buros, Oscar Krisen (ed). The Mental Measurements Yearbook. New Brunswick, N. J.: Rutgers University Press, 1938.

Cattel, Raymond B., The Measurement of Personality. Yonkers: World Book Co., 1947.

Chave, Ernest J., Measure Religion. Chicago: The University of Chicago, 1939.

Davis, Frederick B., Utilizing Human Talent. Washington: American Council of Education, 1947.

Dobbins, Gaines S., The Improvement of Teaching in the Sunday School. Nashville: Southern Baptist Convention, 1943.

Eavey, C. B., The Art of Effective Teaching. Grand Rapids: Zondervan Publishing House, 1953.

Edge, Findley B., Teaching for Results. Nashville: Broadman Press, 1956.

Flanagan, John C., Factor Analysis in the Study of Personality. Stanford University: Stanford University Press, 1935.

Flanagan, John C., Measuring Interests. Advisory Service Bull. No. 4. New York: Coop. Test Service, 1940.

Greene, Edward B., Measurements of Human Behavior. New York: Odyssey Press, 1941.

Greene, Harry A., Jorgensen, Albert N., and Gerberich, J. Raymond, Measurement and Evaluation In The Secondary School, New York, Longmans, Green and Co., 1943.

Greene, Harry A. and Jorgensen, Albert N., The Use and Interpretation of Elementary School Tests, New York, Longmans, Green and Co., 1935.

Hawkes, Herbert E., and others (eds.). The Construction and Use of Achievement Examinations. Boston: Houghton Mifflin, 1936.

Heim, Ralph D., Leading a Sunday Church School. Philadelphia: The Muhlenberg Press, 1950.

Hollingworth, H. L., Judging Human Character. New York: Appleton-Century, 1922.

Hunt, Thelma, Measurement in Psychology. New York: Prentice-Hall, Inc., 1936.

Jennings, Helen, and others, Sociometry in Group Relations. Washington: Ameri. Council on Education., 1948.

Lee, J. Murray, A Guide to Measurement in Secondary Schools. New York: D. Appleton-Century Co., Inc., 1936.

Lincoln, Edward A., and Workman, Linwood L., Testing and the Use of Test Results. New York: The Macmillan Co., 1935.

Monroe, W. S. (ed.). Encyclopedia of Educational Research. New York: Macmillan, 1941.

Mursell, James T., Psychological Testing. New York: Longmans, Green, 1947.

Nelson, M. J., Test and Measurements in Elementary Education. New York: The Cordon Co., 1939.

Orleans, Jacob S., Measurement in Education, New York: Thomas Nelson and Sons, 1937.

Paterson, Donald G., and others., Student Guidance Techniques. New York: McGraw-Hill, 1938.

Ross, C. C., Measurement in Today's Schools. New York: Prentice-Hall, Inc., 1947.

Ross, C. C., Revised by Stanley, Julian, Measurement in Today's Schools. New York: Prentice-Hall, Inc., 1954.

Schwarz, J. C. Ed., Who's Who In the Clergy. New York: No Publisher given, 1936.

Smith, Eugene R., Tyler, Ralph W., and others., Appraising and Recording Student Progress. New York: Harper, 1942.

Smith, Henry L., and Wright, Wendell W., Tests and Measurements. New York: Selver, Burdett and Co., 1928.

Thorndike, Robert L. and Hagen, Elizabeth, Measurement and Evaluation in Psychology and Education, New York: John Wiley & Sons, Inc., 1955.

Thurstone, L. L., and Chave, E. J., The Measurement of Attitude. Chicago: University of Chicago Press, 1929.

Thurstone, L. L., Multiple-factor Analysis. Chicago: University of Chicago Press, 1947.

Torgerson, Theodore L. and Adams, Georgia S., Measurement and Evaluation. New York: The Dryden Press, 1954.

Tyler, Ralph W., Constructing Achievement Tests. Columbus: Ohio State University, 1934.

Watson, Goodwin B., Experimentation and Measurement in Religious Education. New York: Association Press, 1927.

Wood, Hugh B., Evaluation of Pupil Growth and Development. Eugene: University of Oregon Coop. Store, 1940.

Wrightstone, J. Wayne, Justman, Joseph, and Robbins, Irving, Evaluation in Modern Education, New York: American Book Company, 1956.

PERIODICALS

Anderson, Harold H., and Brewer, Joseph E., "Studies of Teachers' Classroom Personalities, II." Appl. Psychol. Monogra. 1946, No. 8.

Beebe, H. Keith, "Teaching by Testing," Religious Education. March-April, 1951. pp. 96-99.

Bloom, Benjamin S., "Test Reliability for What?" Journal of Educational Psychology, 1942, 33, 517-526.

Cockrum, Logan V., "Personality Traits and Interests of Theological Students." Religious Education, January-February, 1952, pp. 28-32.

Cronbach, Lee J., "Test 'Reliability': Its Meaning and Determination." Psychometrika, 1947, 12, 1-16.

Crosby, R. C., and Winsor, A. L. "The Validity of Students' Estimates of Their Interests." J. Appl. Psychol., 1941, 25,408,414.

Darley, J., "Changes in Measured Attitudes and Adjustments." Journal Social Psychology. 1938, 9, 189-199.

Ellis, Albert., "The Validity of Personality Questionaires." Psychol. Bull., 1946, 43, 385-440.

Fischer, Eunice, "Test Your 'S. S. I. Q.'" Christian Life, Vol. 12, No. 3, (July, 1950) P. 54.

Haggard, Ernest A., "Observations on the Measurement of Moral Character." Religious Education. May-June, 1955, pp. 156-161.

Lewis, Don., "The Learning Function." Amer. Psychologist, 1946, 1, 260.

Rinsland, H. D., "A Form for Briefing and Evaluating Standardized Tests" Journal of Educational Research, 42: 371-375, January 1949.

Swineford, Frances, "Analysis of a Personality Trait." Journal of Educational Psychology, 1941, 32, 438-444.

Thayer, C. R., "Psychological Tests in Pastoral Counseling," Church Management, April, 1953, pp. 18-23, 79.

TESTS

Chave, E. J. and Thurstone, L. L. (ed.). Attitude Toward the Bible, Forms A and B. Chicago: The University of Chicago Press, 1933.

Chave, E. J. and Thurstone, L. L. Attitude Toward God (Influence on Conduct) Forms C and D. Chicago: The University of Chicago Press, 1931.

Chave, E. J. and Thurstone, L. L. Attitude Toward God (The Reality of God), Forms A and B. Chicago: The University of Chicago Press, 1931.

Chave, E. J. and Thurstone, L. L. Definitions of God. Chicago: The University of Chicago Press, 1931.

Kohs, S. C. Ethical Discrimination Test. Chicago: C. H. Stoelting Co., n.d.

Laycock, S. K. Laycock Test of Biblical Information. New York: Association Press, n.d.

No Author. My Ideas About Religion. Evanston, Illinois: Northwestern University, 1933.

No Author. Northwestern University Religious Education Tests, Bible Tests, Series A and B. Evanston, Illinois: Northwestern University, 1933.

No Author. Series B. No. 4 Religious Beliefs. Evanston, Illinois: Northwestern University, 1927.

No Author. Test of Religious Thinking, Form A. New York: Association Press, 1928.

No Author. Test of Religious Thinking, Form E. New York: Association Press, 1928.

Thurstone, L. L. and Chave, E. J. A Scale for Measuring Attitude Toward the Church. Chicago: The University of Chicago Press, 1929.

Wilson, Matthew H. A Test of the Knowledge of Right and Wrong Concerning the Professions. Chicago: C. H. Stoelting Co., 1933.

APPENDIX

APPENDIX A

COLE-VON BORGERSRODE SCALE FOR RATING STANDARDIZED TESTS

I. Preliminary Information
   1. Exact name of test?
   2. Name and position of Author?
   3. Name of publisher and nearest address?
   4. Cost?
   5. Date of copyright?
   6. Purpose of test?

II. Validity (25)
   A. Curricular (15)
      1. Exact field or range of education functions which test measures?
      2. Ages and grades for which intended?
      3. Criteria with which material was correlated?
      4. Do questions parallel good teaching procedures?
      5. How wide is sampling of important topics?
      6. What is the social utility of questions?
      7. Is test claimed to be diagnostic? (If so, proof, and see VI, 5,c, below)

   B. Statistical (10)
      1. Correlated against what outside criteria?
      2. Size of coefficient of correlation?
      3. Size and representativeness of sampling?
      4. Proof of validity of items? (such as statements as to experimental tryout of items individually to determine that no large percentage is failed or passed by all pupils and that the items show a consistent increase of percentages of successes with successive age or grade levels).

III. Reliability (25)
   A. Most important items
      1. Correlated with what?
      2. Size and representativeness of sampling?
      3. Reliability coefficient?
      4. The means of the distributions?
      5. The standard deviations of the distributions?
      6. If some other measure than the above three is given to prove reliability, what is it?
      7. Inter-correlations?

   B. Less important but desirable
      1. Order of giving various forms of test?
      2. Is test reliable enough statistically for individual measurement, or can it be used only for groups?
      3. Evenness of scaling? (see II, B, 4)

4. Are pupils accustomed to this type of test?

IV. Ease of Administration (15)
1. Manual of Directions (3)
   a. How complete and simple is the manual?
   b. Does manual control test conditions well?
   c. Typographic make-up?
2. Simplicity of Administration (8)
   a. Amount of explanation needed for pupils by examiner?
   b. Are directions to pupils clear, detailed, comprehensive?
   c. Is arrangement of test convenient for pupils?
   d. Are samples and "fore-exercises" given when needed?
3. Alternate forms (3)
   a. Number?
   b. Evidence of reliability?
   c. Evidence of equivalency?
4. Time needed for giving (1)

V. Ease of Scoring (10)
1. Degree of objectivity--purely objective or some judgment on part of examiner?
2. Are adequate directions given--clear, equal to all emergencies?
3. Is scoring key adjusted to size of test?
4. Time needed to score one test?
5. Simplicity of procedure?
   a. Number of processes needed to get final score?

VI. Ease of Interpretation (20)
1. Norms (6)
   a. Kind--age, grade, percentile, etc.?
   b. Derivation--size and representativeness of sampling?
   c. Tentative, arbitrary, or experimental?
   d. For separate parts?
   e. How expressed?
2. Is class record provided?
3. Are there provisions for graphing results?
4. Is interpretation of raw scores easy or hard?
5. Application of results (10)
   a. Are directions or suggestions given for applications of results to benefit teaching or administration?
   b. Are tests survey or diagnostic?
   c. If diagnostic---
      (1) Proof of diagnostic value?
      (2) What principle or principles underlie construction?
      (3) How many different skills, abilities, or aspects of the subject are analyzed or measured?
      (4) Does the analysis of total subjects into unit abilities follow teaching practices or needs?
      (5) Is the diagnosis individual or class-proof?
      (6) Does the test demand tabulations of individual pupils' errors to secure diagnosis?
      (7) Is a remedial program provided or suggested?

VII. Miscellaneous (5)

1. Typography and make-up?
   a. Arrangement of printed matter?
   b. Legibility of type?
   c. Quality of paper?
   d. Are test blanks free from distractions, norms, directions to examine, etc.?
2. Is the time required for giving as small as is consistent with reliable measurement?
3. Is the cost in keeping with the amount, scope, and reliability of the results yielded?
4. Is good test service provided by the publisher?
5. Kind of new-type questions used?

APPENDIX B

SCHEDULE FOR EVALUATING A TEST

1. GENERAL REFERENCE INFORMATION

   a. Name of test.
   b. Author's name (and position, if available)
   c. Publisher
   d. Date of publication
   e. Cost
   f. Time for administration.

2. VALIDITY

   a. Evidence from the Plan for the Test. What were the procedures for determining the scope of the test? For determining the particular content to be covered? For determining the functions and processes to be represented? How adequate do these appear to be? How closely do the test objectives correspond to objectives that you are interested in for your school?

      What provisions were made for editorial review of the test materials? How adequate do these appear?

   b. Evidence from the Test Blank Itself. Do the test items appear appropriate for the objectives that you are trying to evaluate? Do the test items appear to be well constructed? Are they free from ambiguity? Do they have attractive wrong-answer choices?

   c. Evidence from Statistical Studies of the Test in Use. With what concurrent measures has the test been correlated? For what sort of groups? How substantial are the correlations?

      With what later criterion measures has the test been correlated? For what sorts of groups?

      How does the evidence on statistical validity compare with that for other tests?

      How accurate a prediction does it give of significant outside criteria? How do these results compare with those of other tests that try to measure the same trait?

   d. Evidence from Outside Authority. What have reviewers and critics said about the validity of the test?

3. RELIABILITY

a. How adequately are Data Reported? Do the authors indicate size and nature of groups for which data are reported? Do they indicate type of reliability coefficient computed? Do they give mean and standard deviation for the groups? Do they report reliabilities for single age and grade groups?

b. What are the facts on Reliability? What actual data on reliability are reported? (Indicate, as far as given, the age or grade, size of groups, mean and standard deviation, procedures by which reliability was computed, and resulting values obtained.) How do the data compare with other competing tests?

4. PRACTICAL CONSIDERATIONS IN ADMINISTRATION AND USE OF TEST

a. Factors in Administration

1. Adequacy of manual.
2. Complexity of procedures.

a. Complexity of process required of students.
b. Adequacy of instructions and practice exercises.
c. Complexity of process required of examiner. Timing, giving instructions, and interpreting responses of subjects examined.

3. Time requirements.
4. Legibility, attractiveness, and convenience of format.

b. Factors in Scoring.

1. Time required (i.e. form of answer, type of key, etc.).
2. Special skills required (subjective scoring and qualitative interpretation).

c. Factors in Interpretation.

1. Type of norms. Appropriateness to uses, completeness, representativeness of sample. How readily may raw scores be converted into derived scores?
2. Aids to interpretation provided by manual.

d. Factors in Continued Use.

1. Are there comparable forms? How many? How well is comparability established?
2. Cost. Does this permit routine continued use?

APPENDIX C

SCHEDULE FOR EVALUATING A TEST

I. General Reference Information
   - A. Name of test.
   - B. Author's name and position, if available.
   - C. Publisher.
   - D. Date of Publication.
   - E. Cost.
   - F. Miscellaneous information.

II. Is the Test Valid?
   - A. Evidence From the Plan for the Test.
   - B. Evidence from the Test Blank Itself.
   - C. Evidence From the Statistical Studies of the Test.

III. Is the Test Reliable?
   - A. How adequately are Data Reported?
   - B. What are the Facts on Reliability?

IV. Is the Test Practical?
   - A. Factors in Administration.
     1. Adequacy of Manual.
     2. Complexity of Procedures.
        - a. Complexity of process required of students.
        - b. Adequacy of instructions and practice exercises.
        - c. Complexity of process required of examiner. Timing, giving instructions, etc.
     3. Time Requirements.

   - B. Factors in Scoring.
     1. Time required. (form of answer, type of key, etc.).
     2. Special skills required (subjective scoring and qualitative interpretation.).

   - C. Factors in Interpretation.
     1. Type of norms and appropriatness to user.
     2. Completeness of norms.
     3. Suggestions of remedial program.
     4. Diagnostic or survey.
     5. Aids to Interpretation provided by manual.

   - D. Factors in Continued use.
     1. Comparable forms available.
     2. Cost.

   - E. Format.
     1. Arrangement of printed matter.
     2. Legibility of the Type
     3. Quality of paper.
     4. Freedom from distractions.

APPENDIX D

OTIS SCORE CARD FOR RATING STANDARDIZED TESTS

Item	Stand. No. Points	Name of Tests				
1. Manual	7					
2. Validity	20					
3. Reliability	10					
4. Reputation	3					
5. Ease of Administration (20) a. Little special preparation b. Adequate detailed directions c. Time limits clearly stated d. Alternate forms available	 4 6 6 4					
6. Ease of Scoring (15) a. Objectivity . . . . . . b. Convenient form of Key . c. Time required	 8 4 3					
7. Ease of Interpretation (20) a. Types of norms . . . . . b. Directions for . . . . . c. Class Record Sheet . . . d. Remedial Program	 10 3 2 5					
8. Typography and Makeup . .	5					
Total.....	100					

APPENDIX E

BIBLICAL TEST

Name _____

Grade & Teacher _____

THE LETTER OF JAMES[1]

Directions: Answer the question by circling a T for True and an F for False. DO NOT GUESS. LEAVE BLANK IF YOU DO NOT KNOW THE ANSWER.

1. James was an important person in the early Christian Church..... T F
2. The New Testament says nothing at all about Jesus' younger brother James..... T F
3. Jesus had more than one brother and sister..... T F
4. When Jesus started to preach his family and his old friends supported him 100%..... T F
5. Jesus appeared to his brother James after the Resurrection..... T F
6. James wrote to Jesus the letter we are studying..... T F
7. James never said anything about rich men and poor men..... T F
8. According to James the thing that causes war is selfishness..... T F
9. James says that it is all right to swear if you do not use God's name in doing so..... T F
10. James never makes any reference to Old Testament characters in this letter..... T F

Directions: Write the number of the answer in the space at the side of the paper which most accurately completes the statement.

1. According to James it is the (1. poor, 2. wealthy, 3. sick, 4. prayerful) who will be rich in faith and heirs of God's Kingdom..... ____

2. James says, "Love your neighbor as yourself" and you will fulfill (1. the Ten Commandments, 2. the English Common Law, 3. the Royal Law, 4. the Labor Law) ..... ____

3. According to James, a man has faith who (1. prays for, 2. feeds, 3. finds lodging for, 4. takes to church) a needy man..... ____

4. James uses (1. Moses, 2. Jacob, 3. Elijah, 4. Abraham) as an example of a "man of faith."..... ____

---

[1]H. Keith Beebe, Religious Education, March-April 1951, p. 99.

5. James compares the rudder of a ship to (1. the tongue, 2. the mind, 3. the body, 4. the eyes)..... ____

6. James also compares the tongue with things in (1. philosophy, 2. modern literature, 3. school, 4. nature)..... ____

7. James suggests that self-control help you to (1. to speak cleanly, 2. to influence others, 3. to earn a better living, 4. to make friends easily)..... ____

8. "If you (1. give money, 2. are humble, 3. worship every Sunday, 4. are proud) God will exalt you," says James..... ____

9. The (1. rich, 2. religious, 3. poor, 4. unselfish) have laid up treasure on earth, according to James..... ____

10. James refers to (1. Elijah, 2. Samuel, 3. David, 4. Isaiah) who prayed for rain in an Old Testament story..... ____

11. James is angry because the rich have held back (1. the clothes, 2. the water rights, 3. the wages, 4. the privilege of worshipping) from the poor..... ____

12. James was (1. an elder, 2. a deacon, 3. a minister, 4. a trustee) in the early Christian Church..... ____

APPENDIX F

Test Your "S. S. I. Q."[1]

Here's a quiz planned to test your teacher's I. Q. in Sunday school knowledge. So, if you want to spring something different at your next teacher's meeting, pull this sheet out of your pocket and use it to make your staff brush up on its Sunday school fundamentals. And, incidentally, it's one thing to know the correct answers, and it's another thing to find a Sunday school that follows correct procedure. The answers are below.

1. When uniform teaching material is used it means that:

    a. the same Bible lesson or theme is taught throughout the entire school.
    b. the same Bible lesson is taught to each class throughout a department.
    c. closely graded material is provided for pupils of every age as is done uniformly in public school.

2. The offering is best collected:

    a. in each class.
    b. in the departmental worship service.
    c. outside the front door.

3. Teachers and officers may be initially contacted by departmental leaders but they should be officially appointed or dismissed by:

    a. the pastor or superintendent.
    b. the Christian Education Board.
    c. the pupils involved.

4. The absent pupil is most likely to return if he is followed-up by:

    a. a personal visit in the home.
    b. three phone calls.
    c. ten mailings.

5. A child coming to Sunday school at the age of three is considered:

    a. a Primary pupil.
    b. a Nursery pupil.
    c. a Beginner pupil.

6. Memory Work is best taught Juniors by:

    a. hitting them on the head.

---

[1]Eunice Fischer, Christian Life Magazine, July 1950, Vol. 12, no. 3, p. 54.

b. visualizing the Bible passage.
c. asking them to learn the passage at home.

7. When an adult class has been known as the "Young Women's Class" for twenty years and you want to split it into two classes, the best way is to:

a. divide the group alphabetically.
b. let the older ones remain in the group but take out the younger ones to form a new class having a new name.
c. ask to see their birth certificates.

8. Teaching materials for the Sunday school should be selected by:

a. each teacher who knows the needs of the individual class.
b. each departmental superintendent who consults with workers in the department.
c. a committee composed of pastor, departmental leaders and Christian Education Board who carefully study materials from more than one publisher, and then prayerfully choose one curriculum that combines sound Bible teaching with pedagogical teaching methods.

9. The ideal method of teaching is:

a. the lecture method-the teacher does all the talking.
b. lecture plus quiz-teacher reads questions from quarterly, and the pupils answer.
c. pupil participation-teacher encourages discussion and questions by the students.

10. The ideal size of children's classes is:

a. six to eight pupils.
b. ten to fifteen pupils.
c. twenty to twenty-five pupils.

Answers:
1a, 2b, 3b, 4a, 5b, 6b, 7b, 8c, 9c, 10a.

APPENDIX G

METHOD OF INQUIRY INTO CONTEMPORARY RELIGIOUS EXPERIENCE

A.

1. Study something in motion from one condition to another and/or in encounter. Not just a static condition.

   An enterprise in development, change. We are studying how to work with people on some need or growth; not just to study them.

2. Tend toward study of persons in realistic conditions - rather than paper and pencil situations.

   (e.g. the inquiry situation itself may be more revealing than the recorded content.)

B.

3. Work from within an "I-Thou" situation as a participant servant.
   A) The exploration is of service to the person or situation being explored; and to some degree under their control. We are not prying, or violating the privacy of their inner personal region. (These should reduce the need of the person to fabricate answers; put on protective coloration, clam up.)

   B) Both are looking at it; not just for the investigators use. To some degree it will always be a discovery-of-self and better working solutions of problems. Research is connected with something they already have to do anyway.

C.

4. We study religious experience thru the conscious and willed revelation which people make - their endeavor to make explicit to themselves and to us their experiences, how they see and feel and handle "the something" being studied.

   Such study enterprise is limited by factors such as -
   (a) people may lack words to express their situations, and the words they have catch only a small part of their total feeling and experience. Given other tools, other material might emerge.

(b) Also many of the most important are below the quickly recalled level; and may be brought up only thru crisis or important decision experience.

Thus some "stimulus" which causes behavior, stabs up their reservoir of feeling, awakens, "where they live and stand" is very useful - if not perceived as an attack. The absolutely open-end interview may penetrate at the same level, if we can skillfully follow feelings.

5. The sensitivity of the researcher, to the significant questions on which we need insight; and to the human person is the critical factor. To become a good researcher is therefore largely a matter of then increasing our own sensitivity - which partly means becoming more aware of our own feelings and experiences.

6. Hypotheses representing our present best understanding and hunches are aids to sensitivity, (if not used in a wooden way). At least the conceptual tools (aesthetic also) we use in seeing and analyzing need to be defined and sharpened continually....so that we can see more and integrate more productively what we see.

Further some concept of the overall process we are studying (e.g. a morality of sensitivity instead of legalism) seems to be helpful. We don't see, unless we already know enough to recognize and give selective attention.

7. Evidence will be largely the report of a sensitive personality rather than primarily in forms that can be mathematically treated.

D.

8. Objectivity is secured (a) by this statement of the tools hypotheses, sensitivities and perspectives with which we look and categories by which we analyze.
(b) by the cooperative use of the person "studied" and other experts in this situational field.
(c) by tape recording which bypasses wishful memory distortion, preserves the spoken word and mood. Our problem is to secure "an objectivity of subjectivity."

9. There are therefore some advantages to team exploration so that the evidence can be looked at from different frames of understanding.

E.

10. In addition to the usual record, the material should also be written

up into a form of the first person internal frame of reference. (Write it up as if the person himself were speaking in first person, trying to explain his Self situation and faith relationship.)

Such a form and effort has these values -

(A) The "inquirer" is more sympathetic, less judgmental of the person or persons being studied. There is less treatment of them as "a case."
(B) The very form of the task tends to produce more "entering into" in us of the other person. And therefore improves our relationships in inquiry.
(C) The write-up has more impact; more intense subjectivity, and yet less likely to be on basis "I liked that about him."
(D) Helps to keep our ideas and analysis tentative.
(E) We have to be concerned with the depth, and not just the external symptoms.

(Both an objective report of external behavior and this internal frame-report is desirable.)

F.

11. Without unduly extending ourselves and the material, and exercising too much dogmatism or malignant simplification, some thought, idea or insight finally organizes the material, tries to indicate it's center, structure, and significance (relationship to other experience and ideas.)

This is the creative imagination at work; (which may have no discipline without No. 10) and comes as we mull over, let gestate the hard work we have been doing. (See theory of creativity.)

APPENDIX H

Copy of a letter to Mr. Watkins from The Federated Theological Faculty of The University of Chicago, Chicago, Illinois.

Dear Mr. Watkins:

Your observation is correct that not much has been done for quite a while in the field of tests and measurements in religious education. Partly, everybody has been so busy trying to find new foundations and design new shapes of program. I suppose once some grasp of these is managed, people will begin to ask "how do we know if we are reaching the goals we have set?"

Another factor is our discovery that maybe what we can find out by tests--even attitude tests; leaves us with pretty surface information. And there is very little either student or teacher can do. In order to find depth, paper and pencil tests are misleading. We have to come closer to the methods of a therapeutic interview; we have to find ways whereby a "startling encounter" with a situation or person's faith can take place, and then see what this awakens in a person. This is particularly true in the field of religious development.

Certainly all the devices invented and used in college placement tests can be adapted to test a persons knowledge of facts; and his mental power to do external thinking about those facts. This probably we should try to develop. But not kid ourselves that we are getting at the real article.

We have done nothing along the lines of measurement since Dr. Chave retired. We have been doing considerable open-ended and depth

interviewing, using projective tests that are adaptations of the TAT, are fooling around some with Osgood's semantic differential ; all in all trying to find ways that the learner can be helped to look at his experiences and himself at greater depth...and for his own use rather than for some teacher to then try to shape him in the way he is not yet. Ernest Ligon makes the greatest pretense at scientific measurement: but I think it is mostly pretense. We also use a number of instruments in group work, by which the group tries to analyse its own behavior and progress.

Sincerely,

(signed) Ross Snyder.