1969

A Reliability and Validity Assessment of Scales Measuring the Perception of the Evaluations of Others

John A. Vonk
Western Michigan University

A RELIABILITY AND VALIDITY ASSESSMENT OF SCALES MEASURING THE PERCEPTION OF THE EVALUATIONS OF OTHERS

by

John A. Vonk

A Thesis
Submitted to the
Faculty of the School of Graduate
Studies in partial fulfillment
of the
Degree of Master of Arts

Western Michigan University
Kalamazoo, Michigan
1969

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to my committe members (Drs. Lewis Walker, Chairman, Edsel L. Erickson and James A. Schellenberg) for their advice, guidance and encouragement throughout this thesis. I am especially indebted to Dr. Erickson for making available part of the data from a larger study. This study was entitled Scales and Procedures for Assessing Social-Psychological Characteristics of Visually Impaired and Hearing Impaired Students (Cooperative Research Project No. 6-8720), supported by the Department of Health, Education and Welfare (U.S. Office of Education).

Finally, I would like to thank Dr. Lee M. Joiner for his methodological contributions and the students for their responses to the instruments used in this study.

TABLE OF CONTENTS

CHAPTER I
INTRODUCTION

Statement of the Problem

There have been many studies which have examined the various social-psychological characteristics of the educational process of students regularly placed in the school system. There have not been many studies, however, which have focused upon the educational process of the exceptional child. This has been considered a serious problem by a few scholars in the discipline of sociology as well as in the field of special education.

In 1960, sociologists, psychologists and research specialists in the area of the deaf[1] called attention to this problem as members of a national conference on the Research Needs in the Vocational Rehabilitation of the Deaf.

These research specialists emphasized the need for the development of instruments designed to measure certain social-psychological characteristics of hearing impaired children. The conference called for investigations concerning parental attitudes and educational attainment;

---

[1]Rogers, Merrill and Quigley, Stephen P., "Research Needs in the Vocational Rehabilitation of the Deaf." American Annals of the Deaf, Vol. 105, No. 4 (Sept. 1960), 3-5

and research as to how the individual relates to the school, home, and community, and his concept of his position within each. The conference also gave high priority to sociological and psychological studies investigating the status of the family with deaf members. Moreover, the members of the conference stressed the need for the development of revised instruments which would yield data comparable to that gained from instruments used on non-hearing impaired students. One of the methods that may be used to attain this would be a modification of existing instruments so that they will be applicable to hearing impaired students.

In the past, it has been very difficult to compare sociological, psychological, and social-psychological data gained from populations or samples differing on salient variables, for example, hearing ability. This is due,[1] at least in part, to the necessity for different measurement instruments, different theories, and different research designs. This problem has been noted for the field of sociology by Wilbur B. Brookover[2] and his associates. Brookover is one investigator who has

---

[1]Geer, William C.,and Deno, Evelyn D., "CEC and Legislation-Now and in the Future." Exceptional Children, Vol. 32, No. 3 (November 1965), 187-194.

[2]Brookover, et al, Relationship of Self Concept To Achievement in High School, U.S. Office of Education Cooperative Research Project No. 2831 (East Lansing, Michigan State University, to be published in September 1967).

studied several social-psychological characteristics of the educational process of students regularly placed in the school system. He has contended however, that his theoretical models, employed in several studies of students without known impairments, will be of questionable value until they are applied in studies of exceptional children. The problem Brookover notes is applicable to nearly all sociological, social-psychological and psychological studies.

As previously mentioned, one of the difficulties of obtaining comparable data from hearing impaired students and students without known hearing impairment is the problem of instrumentation. If an instrument is designed to measure a social-psychological construct for non-hearing impaired students, the language or wording used may not be comprehensible to the hearing impaired student. Moreover, if the instrument is modified, or the concept translated into sign and finger language, it may render the concept incomprehensible. This increases the probability of error and hinders the ability to reach accurate conclusions. If however, the instrument is redesigned to cope with the limits of the exceptional child, there is no assurance that it will yield reliable and valid data. For example, if the wording or language is changed to meet the limitations of hearing impaired children, it may not

yield data comparable to the data gained from non-hearing impaired children. The reliability and validity of revised instruments then, must be established before using these instruments to study exceptional children.

The topic considered in this study is based upon instruments developed and standardized under co-operative research projects #845, #1636 and #2831, directed by Brookover[1] at Michigan State University. These projects directed by Brookover were carried out upon students without known impairments. That is, the students were regularly placed in the public school system and not in a special education curriculum. Some of the important findings of these projects pertinent to this study are: (1) ones self definition of ability is significantly related to achievement; that is, the academic achievement levels of the students is impeded or facilitated by a low or high self concept of academic ability; (2) a student's self definition of ability is dependent upon his definitions of the evaluations others have of him, and (3) change or stability in the definitions of evaluations of others is related to change or stability in the students' self definition.

---

[1] see Appendix A.

If one is to engage in this type of analysis of the educational process with exceptional children, problems of instrumentation will most certainly arise. One can not be confident that the modified scales will yield data comparable to the data gained from the original scales. For this reason, a methodology study focusing on an assessment of the reliability and validity of the modified instruments should be the first step in a study of the educational processes of exceptional children.

The problem considered in this research is an assessment of the reliability and validity of an instrument designed to measure a student's definition of how others (primarily parents, teachers, and friends) evaluate his academic ability. More specifically, the purpose of this study will be to determine the reliability and validity of instruments purporting to measure these definitions of the evaluations of others with hearing impaired students.[1]

The instrument under analysis consists of three sub-scales: (1) the first sub-scale is designed to measure the students' definition of parental evaluation of his ability, (2) the second sub-scale is designed to measure the students' definition of his teachers

---

[1]The scales under investigation were originally developed and standardized by Wilbur B. Brookover and his associates under Cooperative Research Projects #845, #1636 and #2831.

evaluation of his ability, and (3) the third subscale is designed to measure the student's definition of his friend's evaluation of his ability. These definitions of the evaluations of others refer to how the student perceives his academic ability is being appraised by his parents, teachers, and friends. For example, the student is asked to respond to the question: "Does your best friend think you could graduate from college?"

Edsel Erickson, Lee Joiner, and Wilbur Brookover have modified this instrument to the extent that it is now translatable into sign and finger language for use with hearing impaired students.[1]

To summarize, this investigation is designed as a methodological study, which will assess the reliability and validity of two supposedly parallel instruments to determine if, indeed, they are assessing the same phenomenon. It is hoped that through this study a more definitive statement may be made concerning the reliability and validity of the instrument purporting to yield comparable data of two differing populations.

---

[1]See Appendix A for scales designed for regularly placed students and Appendix B for the scales designed for hearing impaired students.

RELATED LITERATURE

A study of the reliability and validity of the instruments is the necessary first stage in a study of exceptional children. Without a knowledge of the reliability and validity of the instrument, one can not rely upon the results obtained in the study, or the conclusions drawn from the results. Logically, in a study of this type, an assessment of the reliability would be a starting point, since reliability is a necessary condition for validity.[1]

Reliability

In assessing the reliability of an instrument, one seeks to determine how reliable or unreliable it is by determining the degree to which errors of measurement are present. To the degree that errors of measurement are present, it is unreliable. Kerlinger[2] states that reliability can be defined as "...the relative absence of errors of measurement in a measuring instrument. Reliability is associated then, with random or chance error."

---

[1]Thorndike, Robert L., and Hagen, Elizabeth, Measurement and Evaluation in Psychology and Education: Second Edition, New York: John Wiley and Sons, Inc., 1961. 185.

[2]Kerlinger, Fred N., Foundations of Behavioral Research. New York: Holt, Rinehart and Winston, Inc., 1964. 430.

There are three basic types of reliability with corresponding or appropriate techniques for computing each which are accepted by most methodologists interested in reliability analysis.[1] The first type of reliability is concerned with the comparability of items within an instrument. Basically, it asks the question: Do these items in the instrument all measure the same behavioral phenomena? This type of analysis yields a coefficient of internal consistency and is based upon an internal analysis of the data from a single testing. The second type is an assessment of the equivalence of forms of the measuring instrument. This type of reliability inquires whether a modified form of the test will result in a similar or identical ordering of the individuals as the original form of the same instrument. This type of analysis yields a coefficient of equivalence in that it is a correlation coefficient between two forms of the same instrument administrated at the same time. The third type of reliability assesses the stability of measurement of the instrument over time. This type seeks to answer the question: Will the same results be obtained at a later testing?

---

[1]For a more complete discussion of these reliability coefficients, see American Psychological Association, "Technical Recommendations for Psychological Test and Diagnostic Techniques." Psychological Bulletin, No. 2 part 2 (special supplement, March 1954), 51.

This coefficient of stability is a correlation coefficient between two administrations of the same test administered over a period of time. A complete analysis of the accuracy and precision of an instrument should employ all three types of reliability assessment.

Validity

It is possible for a measuring instrument to measure with the greatest of precision or accuracy and yet be invalid for the purpose intended.[1] For example, one may take a tape measure and accurately measure head size, but if the results of this measure are supposed to reflect intelligence, it will be an invalid measure. In this example, the measurement device would be reliable, but not valid. This should indicate that whenever one is concerned with validity, he is concerned with something above and beyond the mere precision of the measuring instrument. However, the accuracy or precision of the instrument is an integral part of its validity. Thorndike and Hagen[2] state that "...only to the extent that a test measures something accurately, can it measure validity." They

---

[1]Thorndike and Hagen, op.cit., p. 185.

[2]ibid.

state in a statistical theory of reliability that the theoretical limit for a validity coefficient is the square root of the reliability coefficient of that test. For example, if a test has a reliability coefficient of .81, the theoretical ceiling for the validity coefficient would be .90.[1] They are implying here that one must have reliability before one has validity, that is, reliability is a necessary condition for validity, and it is determinant of the theoretical limit of validity. Validity is concerned with the basic question: "Does the scale actually measure what it claims or purports to measure?"

There are four types of validity outlined by the American Psychological Association in their Technical Recommendation for Psychological Test and Diagnostic Techniques. These are: (1) Content Validity, (2) Predictive Validity, (3) Concurrent Validity, and (4) Construct Validity.[2]

---

[1] ibid.

[2] As stated, these are the types outlined by the American Psychological Association. There are different, but related ways of describing approaches to validation. For example, Claire Selltiz, et. al., in Research Methods in Social Relations, 1963, call predictive validity pragmatic validity. They consider both predictive and concurrent validity under pragmatic validity, since both predictive and concurrent validity may be considered as criterion oriented. In this approach to validity, the investigator is concerned with the usefulness of the test as a predictor of

Content validity or logical validity is probably the most commonly used method of validation. A logical analysis of the content of the items in an instrument is the essence of content validity. One examines the items in the measurement instrument to determine if they appear to be related to what is going to be measured. This type of validation then, is basically judgemental. The items of the test must be studied and judged according to its presumed relevance to the criterion being measured.[1] This logical validation or face validity "...is almost always used because it automatically springs from the careful definition

---

[2](con't) some criterion, or as an indicator of some criterion. In one form of pragmatic validity, the researcher may be interested in proposing one form of test as a substitute for another. In the other form, the researcher may predict from an independant variable to a dependant variable. Goode, William J., and Hatt, Paul K., in Methods in Social Research, 1952, also discuss different approaches to validation. They discuss four approaches which they label as: (1) Logical Validation or face validity, (2) Jury Opinion, which seems to be just another method of logical validation, (these approaches correspond to what the APA has called content validity), (3) Known Groups, which corresponds to concurrent validity, and (4) Independent Criteria. In general, these different approaches to validation correspond to those given by the American Psychological Association, or those discussed by Cronbach and Meehl in "Construct Validity in Psychogical Tests", Psychological Bulletin. (July 1955) 281-302.

[1]Kerlinger, op. cit., p. 447.

of the continuum and the selection of items.[1] It is, however, of the lowest power in validation, because whether or not a set of questions measured a given phenomena, cannot be answered by logic alone, but should be empirically determined.[2]

Predictive validity does not require that a certain outcome will occur in a future state of affairs. It could predict something in the future, at the present, or even something that happened in the past. Kerlinger[3] states that in predictive validation, one predicts from an independent variable to a dependent variable. Seltiz and associates[4], state that in the pragmatic approach to validity, the investigator is concerned with predicting some other behavior of the individual.

Concurrent validity should be considered whenever it is proposed that one test be substituted for another.[5] Validation by this method is an assessment

---

[1]Goode and Hatt, op. cit., p. 237

[2]ibid.

[3]Kerlinger, op. cit.

[4]Seltiz, et. al., op. cit., p. 157

[5]Cronbach, Lee J., and Meehl, Paul E., "Construct Validity in Psychological Test." Psychological Bulletin. Vol. 52 (July 1955) 282.

of the extent to which both forms predict the same event. Selltiz[1] states that there must be well established (reliable and valid) indicators with which the results may be compared. This method of validation appears to be similar to the assessment of the reliability of equivalent forms. The emphasis for validation, however, is not based upon the relationship between the two forms, but is upon the common relationship to the hypothesized dependent variable.

According to Kerlinger[2] , construct validity is involved whenever a hypothesized relationship is empirically studied. This is a limited test of construct validity. The test in construct validity is not only the accuracy of the test, but is a test of the construct. Stated more broadly, it involves a validation of the theory in which the construct is embedded; a test of the whole theory, not just the validity of the measuring instrument.[3] For example, an examination of construct validity would involve an assessment of the relationship, between the definitions of the evaluations of others and self concept, between self concept and achievement, and between the definition of the evaluation of others and achievement.

---

[1] op. cit.

[2] Kerlinger, op. cit., p. 452.

[3] Selltiz, op. cit., p. 160.

CHAPTER II

THE SAMPLES AND METHODS OF ANALYSIS

The principal task of this chapter will be to present the sample descriptions and the techniques used to assess the reliability and the validity of the two forms of an instrument designed to ascertain students' definitions on the specific dimensions.

The Samples

The design of this study calls for a group to which two forms of the test will be administered along with a second group to which only one form of the instrument will be administered at two points in time. The latter group will be used to determine the index of stability in the reliability analysis, thus, longitudinal data are necessary. Two samples were necessary because the data for this study were dependent upon two somewhat unrelated studies.

The non-hearing impaired group to which the two forms were administered was drawn from a public high school population in Grand Rapids, Michigan. This group was a racially hetergeneous group, consisting of 185 non-hearing impaired students, 16 years of age, in the eleventh grade, who came mainly from the

lower class and were not involved in any special education program.

A sample had to be selected that could complete both forms of the instrument in order to carry out a systematic analysis concerning reliability and validity. Thus, the non-hearing impaired group was given the original and revised forms, because of the incomprehensibility of the original form for hearing impaired students.

The hearing impaired sample consists of all hearing impaired students who attend the Indiana School for the Deaf, in Indianapolis, as residential students. These hearing impaired students are in a regular high school program, grades eight through twelve, and are between the ages of 12 and 19 years old. The number of this population is approximately 87 students. The students in this population were not controlled by age, area, or grade, because this would reduce the size of the sample considerably and meaningful comparisons could not be made.

As mentioned before, this population was selected because of the longitudinal data that could be gained. Longitudinal data for the deaf scales were not obtained from the Grand Rapids sample. In order for an index of stability in the reliability analysis, longitudinal data are necessary.

METHOD OF ANALYSIS

Reliability

It was stated in the previous chapter, that in order to be able to depend on their measurement, the instruments used by social scientists, must be consistent, stable, accurate and predictable.[1] To assess this reliability, one must determine the degree to which systematic and error variance enter into the scores that are obtained when a measurement device is administered.[2] The total variance of scores may be due to several factors, usually these fall into two categories: (1) Where X_t = an obtained score, X_t may be thought of as a sum of ($X\alpha$), the true score under perfect conditions of measurement, and (X_e) an error component. $X_t = X\alpha + X_e$. (2) Reliability may also be thought of in variance terms, where (V_t), the total variance of scores, equals ($V\alpha$) the sum of true variance and (V_e), the variance due to error

---

[1]Stouffer, S., et. al., Measurement and Predictions. Princeton: Princeton University Press, 1950.

[2]Kerlinger, Fred N., Foundations of Behavioral Research. New York: Holt, Rinehart, and Winston, Inc., 1964.

of measurement. $V_t = V\alpha + V_e$.

It seems that a condition where there is no error of measurement rarely exist in the behavioral sciences, that is, where $X_t = X\alpha$ and $V_t = V\alpha$.[1]

The methods employed to assess this reliability will correspond to the types of reliability stated in the related literature section of the previous chapter. This will include an assessment of Internal Consistency, Equivalence and Stability.

---

[1] ibid., for the weaknesses of this statistical explanation of reliability, see Jane Loevinger, "A Systematic Approach to the Construction and Evaluation of Test of Ability," Psychological Monograph. Vol. 61 No. 285 (1947). In the statistical explanation of reliability stated by Kerlinger, the basic assumption is that an individual's score can be expressed as the sum of a true score and error component. Loevinger quotes Thurstone as saying, "The true score in the test is assumed to be the average score that a subject would make in an infinite number of parallel forms of the test. Of course, the true score can never be actually obtained because the number of parallel forms that can be given to a person is finite and hence, there will always be a residual of chance error even if we ignore the large systematic errors of fatique and boredom, which an attempt would necessarily invite. But, theoretically, the concept of a true measurement as the mean of an infinite number of repeated measurement is a useful one. Evidently, when a test is given to a subject, we want to ascertain as nearly as possible his true score." The Systematic Errors of fatique, boredom, practice and learning (systematic in the sense that if they would be constant for all levels of ability) are a more serious objection. It can hardly be conceived that all of the subjects measured would react or be affected the same way by practice, boredom, or the learning affect operating in repeating measurement. According to Loevinger, "The objection, be it noted, is not that the true score cannot be computed, but that it

Assessment of Internal Consistency

There are four analytical methods for determining internal consistency. These methods attempt to show the existence of a common sharing of characteristics or in other words, homogeneity.

1. Analysis of Internal Consistency by Face Value

This is the most typical and most frequently used method of assessing the integrity of items. The question one asks with this method is: Is it apparent that the items are measuring the same behavioral phenomena?[2] For example, in the parental evaluation scale, two of the questions are:

1. Think of your mother and father. Do your mother and father say you can do school work better, the same, or poorer than your friends?

2. Would your mother and father say you would be with the best, average, or below average students when you graduate from high school?

---

[1](con't) does not exist, it is defined in terms of operations, but operations which in the nature of things cannot be performed, namely the averaging of repeated tests, where there is not effect of repetition." This theoretical statement of reliability will be applied, however, even though it is pragmatically difficult to assess, it remains logically sound.

[2]See Appendix B for scales measuring a student's definition of the evaluations of others.

The question to be asked now is: Are these items both measuring the same phenomena?

Decisions as to homogeneity of items on the basis of face value are the lowest level of scale analysis, and these have been made in the construction of the scales for both the hearing impaired and the non-hearing impaired student.

2. Analysis of Internal Consistency by Item Analysis

One method that has been used to estimate internal consistency with the original scales is Hoyt's Analysis of Variance. This method will also be used in this study. Hoyt's Analysis of Variance utilizes the theoretical definition of reliability presented in Foundations of Behavioral Research by Kerlinger;[1] in that, reliability is the ratio of the variance of true scores to the variance of obtained scores. More specifically, this test is used to determine whether or not the ratio of error variance or random error to individual variance is appropriately small.

3. Analysis of Internal Consistency by a Rational Ordering of Items

The rational ordering of items with respect to

---

[1]See the Introduction to Reliability in this chapter.

one another is commonly referred to as scaling. It is a type of internal consistency analysis often not considered in test analysis. This ordering of items with respect to one another will be employed to ascertain whether each item in the scale has a similar meaning.

Guttman, according to Goode and Hatt,[1]

> "...considered an area scalable if responses to a set of items in that area arranged themselves in certain specified ways. In particular, it must be possible to order the items such that, ideally, persons who answer a given question favorable all have higher ranks than persons who answer the same question unfavorable. From a respondents rank or scale score, we know exactly which items he endorsed. Thus, we can say that the response to any item provides a definition of the respondent's attitude. This quality of being able to reproduce the responses to each item, knowing only the total score, is called reproducibility, which is one of the tests as to whether or not a set of items constitutes a scale in Guttman's sense."

This ability to reproduce a subjects' complete response pattern from a knowledge of his total score and the order of difficulty is called reproducibility.[2]

Guttman's definition of reproducibility is similar to Loevinger's notion of homogeneity. Loevinger defines homogeneity as, "In a perfectly homogeneous test,

---

[1] Goode and Hatt, op. cit., p. 286-287.

[2] White, Benjamin W., and Saltz, Eli, "Measurement of Reproducibility." Psychological Bulletin, Vol. 54 (March 1957), 83.

when the items are arranged in the order of increasing difficulty, if any item is known to be passed, the probability is unity of passing all previous items."[1] Thus, it can be seen from these two definitions that the perfectly reproducible test and the perfectly homogeneous test are identical. Thus, reproducibility or homogeneity attest to the accuracy of an instrument as determined by the magnitude of the reliability coefficient.

There are several techniques of scale analysis available to the researcher. Thurstone's technique is used when initially constructing a scale. The selection of items and scaling proceeds simultaneously. In the construction of the modified form of the definitions of others' evaluation scale, it was not necessary to select items from a universe of items. The problem was not in selecting the best items, but one of determining if the modified items retained their "goodness" or accuracy. Hence, for this study, the Thurstone technique is inappropriate.

Guttman scalogram analysis[2] has been applied to the original scales, and will be applied to the

---

[1] loc. cit., p. 87

[2] For the weaknesses and criticisms of Guttman scalogram analysis, see Jane Loevinger, "The Technique of Homogeneous Test compared with some aspects of 'Scale Analysis' and 'Factor Analysis'". Psychological Bulletin, Vol. 45, (1948), 507-529.

scales for hearing impaired students, because the technique is well known and understood. Guttman recommends a coefficient of reproducibility of .90 or better before an instrument is accepted as scalable.

In addition to Guttman's scalogram analysis, Green's method of scale analysis will also be employed, because it is considered a more conservative index and it allows the analyst to set the confidence limits of the scale score. Green calls his coefficient an "index of consistency" and recommends a coefficient of .50 before the instrument can be acceptable as internally consistent. The reason for the difference between the magnitude of the two coefficients is that Guttman is only concerned with reproducibility, while Green takes into account the probability of the items being related by chance and the difficulty of the items.

4. Analysis of Internal Consistency by Inter-Test Correlation

Another method of internal consistency analysis is the inter-test correlation, which cannot be used in this study because of the limited number of items. Goode and Hatt state that "Each half scale must contain sufficient items to be reliable itself." A minimum number for this is probably 8 to 10, so the

entire scale should not be shorter than 16 to 20 items.[1] The scales under investigation in this analysis fall short of this minimum number of items set by Goode and Hatt, as they consist of 5 items.

Assessment of Equivalence

A second method of assessing the reliability of an instrument is an "estimate of equivalence", in that an attempt is made to estimate the extent to which different instruments, applied to the same individuals, yield similar results.

According to Thorndike and Hagen, "if we have two forms of a test, we may give each pupil first one form and then the other." They may follow each other immediately if we are not interested in stability over time, or may be separated by an interval, if we are. The correlation between the two forms will provide an appropriate reliability coefficient.[2]

In an assessment of equivalence, the time interval between the two administrations must be short enough so that it is reasonable to expect that the

---

[1] Goode and Hatt, op. cit., p. 236.

[2] Thorndike and Hagen, op. cit., p. 78.

characteristic being measured has not changed. In addition to a pearson product moment correlation coefficient computed between the alternate forms, the complete equivalence analysis will include a comparisons of the Kurtosis and Skewness of the distributions.

Assessment of Stability

According to Selltiz[1] et. al., "the appropriate method for determining stability is comparison of the results of repeated measurements. This is true whether the source of instability is genuine fluctuation in the characteristic being measured or random error due to inadequacies of the measuring procedure." Thorndike's statement (quoted in the section on equivalence) also applies to an assessment of the stability of a measurement instrument. Here, a correlation between the scores from the first and second administration of the scales for use with hearing impaired students provides an appropriate reliability coefficient. In this type of analysis, however, if there is not a high correlation, it is impossible to tell if this is a result of genuine changes in the characteristic being measured, or is due to the inadequacies of the measuring instrument.

---

[1]Sellitz, et. al., op. cit., p. 169.

Validity

A complete validity analysis will involve an assessment of the four types of validity discussed in the previous chapter.

Assessment of Content Validity

As stated earlier, content or logical validity is one of the most commonly used methods of validation, because it stems from the definition of what is to be measured and the selection of items. This appears similar to decisions as to the face value of items in reliability analysis, but differs in the questions asked. In the reliability assessment the question raised is, "Do all of the items measure the same thing?" In an assessment of validity the question raised is, "Do these items all measure what they purport to measure?"

Assessment of Predictive Validity

An assessment of the predictive validity of a scale involves decisions as to the existence of an association between hypothesized independant and dependant variables. Evidence of predictive validity will be obtained if there is a satisfactory correlation between the scale purporting to measure one's "definition of how others evaluate his academic ability" and his "self-concept of his academic ability."

Assessment of Concurrent Validity

Concurrent validity refers to the extent to which alternate forms of a test predict the same event. This appears similar to the reliability assessment of the equivalence of forms. The emphasis, however, is upon the common relationship to the hypothesized dependant variable. Evidence of concurrent validity will be obtained if the scale for use with hearing impaired students correlates with his self concept as well as the scale for use with non-hearing impaired students. If the reliability of the scale for hearing impaired students is higher than that of the scale for non-hearing impaired students, the relationship should be higher. If the reliability coefficient is lower, the validity coefficient should be lower.[1]

Assessment of Construct Validity

Construct validity or theoretical validity, according to Kerlinger,[2] occurs whenever hypothesized relationships are empirically tested. This type of validity, however, is more than the instrument alone.

---

[1]For a discussion of this statistical limit of validity, see the related literature section of Chapter I, p. 11.

[2]Kerlinger, op. cit., p. 452-53.

Construct validity is an assessment for the theoretical rational for the constructs under investigation. Basically, it is a test of the complete theory underlying the investigation. In Research Methods in Social Relations, Campbell and Fiske[1] are quoted as suggesting the basic kind of evidence necessary in construct validation. This is: evidence that different measuring instruments of the same construct will yield similar results.[2]

This correlation should provide evidence for a partial assessment of construct validity. Only a partial assessment of construct validity can be achieved in this investigation, since the scales under analysis constitute only a small part of the theory underlying the construct.

Throughout this discussion of the methodology, the emphasis has been upon the reliability analysis, because there has been considerable validity analysis done with the original scales. There has only been limited reliability analysis on the original scales. Therefore, the primary concern of this investigation will be upon the reliability of these scales.

---

[1]Selltiz, op. cit., p. 161.

[2]Evidence of this will be demonstrated in a examination of concurrent validity.

CHAPTER III

FINDINGS

This chapter is devoted to the findings of this research. It is concerned with the reliability and validity of the three scales purporting to measure a student's definition of how others, (primarily parents, friends and teachers) evaluate his academic ability. The results of this research will be discussed in the same order as it was presented in the methodology section of the previous chapter.

Reliability

The purpose of reliability analysis is to determine the extent to which consistency and chance error are present in a measuring instrument. The reliability analysis in this chapter will be an assessment of Internal Consistency, Equivalence, and Stability.

Assessment of Internal Consistency

Analysis of Internal Consistency by Face Value

Determination of the internal consistency of the items by face value is very similar to an assessment of the face validity of the items. The question

one asks in reliability analysis is: Do all of the items measure the same behavioral phenomena?[1] A perusal of the scale should indicate that the items in the scale all appear to be measuring the same behavioral phenomena.

Definition of Parental Evaluation - Deaf

1. Think of your mother and father. Do your mother and father say you can do school work better, the same, or poorer than your friends?

   a. better
   b. the same
   c. poorer

2. Would your mother and father say you would be with the best, average, or below average students when you graduate from high school?

   a. the best
   b. average
   c. below average

3. Do you think you could graduate from college?

   a. yes
   b. maybe
   c. no

4. Remember, you need more than four years of college to be a teacher or doctor. Do your mother and father think you could do that?

   a. yes
   b. maybe
   c. no

---

[1]The question one asks in the validity analysis is: Do these items all appear to measure a student's definition of how others evaluate his academic ability. This will be discussed in a later section of this chapter.

5. What grades do your mother and father think you can get?

   a. A's and B's
   b. B's and C's
   c. D's and E's

The scales measuring the evaluations of friends and teachers are very similar. Rather than the words mother and father appearing in each question, the words friend and teacher occur in the appropriate scale.[1]

Analysis of Internal Consistency by Item Analysis

Another approach to the assessment of internal consistency is to determine the equivalence of items within a scale by various statistical techniques. For example, Hoyt's Analysis of Variance, or the Kuder-Richardson Formula #20.

The Kuder-Richardson Formula #20 was ruled out as a technique for determining the internal consistency of the scales under investigation. However, a discussion of this technique is necessary because of the relevance of their assumptions to Hoyt's Analysis of Variance and this approach to reliability.

The Kuder-Richardson Formula #20[2] yields a co-

---

[1] See Appendix B for an example of these scales.

[2] Remmers, H. H., Gage, N. L. and Rummels, J. Francis, A Practical Introduction to Measurement and Evaluation. New York: Harper and Row, 2nd Edition, 129-130.

efficient equal to the mean of all possible split half coefficients of the test under examination. Their formula where (n) is the number of items in the test, (St) is the standard deviation of the total test scores, (p) is the proportion of persons passing each item and (q) is 1-p is:

$$\frac{n}{n-1} \qquad \frac{St^2 - pq}{St^2}$$

The Kuder-Richardson Formula was considered inappropriate for the scales under consideration in the present investigation, since it can be demonstrated that this formula only applies "to a case of no importance." Jackson and Ferguson[1] point out that in the derivation of this formula, Kuder and Richardson explicitly assume that all items are of equal difficulty and also make the assumption that all items have equal standard deviations. Making these assumptions is equivalent to assuming that "... there are at most two degrees of difficulty of items, that is, the number passing any item must equal either the number passing, or the number failing any other item.[2] In addition, it can be demonstrated that perfect (1.00) item inter-

---

[1]Jackson, Robert and Ferguson, George, "Studies on the Reliability of Test," Bulletin No. 12, Department of Educational Research, University of Toronto, (1941).

[2]Loevinger, Jane, op. cit., p. 11.

correlation is also a necessary condition in order to obtain a perfect (1.00) reliability coefficient, and any deviation from this lowers the reliability coefficient." From the statement, the reliability will equal one only if all the items are perfectly correlated and equal in difficulty. It is only one step to the statement that the reliability will equal one only if everyone has a score of zero or perfect.[1] If this is the case, one could obtain exactly as good results by just giving one item rather than giving the whole test. In addition, if a scale is designed to be reproducible, there cannot be equal item difficulty or perfect inter-item correlation, since it would be impossible to rank the items within the scale.

Cyril Hoyt[2] derives the same formula with two new sets of assumptions, but since his "results" have the same "consequences", his "derivations are suspect of harboring the original or equally bad assumptions." Hoyt's definition of reliability is defined as the ratio of true score variance to obtained score variance.[3] Since this analytic technique has been used so frequently

---

[1] ibid.

[2] ibid.

[3] For a discussion of Hoyt's definition, see footnote 25 in Chapter II.

in the studies of Michigan State University by Wilbur Brookover and associates, a discussion of the limitations of this method, is pertinent to this investigation.

Loevinger[1] argues that:

> His (Hoyt's) initial assumption is that the error component for each person on each item is normally distributed with the same variance as the error component in every other item. The error component is defined as the difference between the actual score and the true score of the person on the item. The true score is a constant based on the difficulty of the item and the ability of the person. Since the actual score on the item is either one or zero, and the true score is a constant, the error component must equal either one minus the true score, or simply minus the true score. The error component for any one person and any one item has only two possible values, which is a far departure from the normal curve. Moreover, the variance of the error component depends solely on the probability of the person passing the item, so the assumption of a constant variance for the error component is equivalent to the assumption that the probability of any person passing the item is a constant. Hoyt's assumptions are worse than Kuder and Richardson's. Rather than simply restricting consideration to an unimportant special case, Hoyt has considered an impossible case, for his assumptions are mutually contradictory.

In spite of the limitations of this type of analysis, when employed with an instrument scored right or wrong (one or zero) and designed to be a reproducible scale, analysis of variance will be calculated in order that the present findings may be compared with those of Brookover and his associates.

---

[1]Loevinger, Jane, op. cit.

The data presented in Table I are Hoyt's analysis of variance calculated on the high school students in Michigan State University studies done by Wilbur Brookover and his associates. These data are presented here in order that the analysis of variance calculated on the Grand Rapids data might be compared.

TABLE I

Hoyt's Analysis of Variance

	DPEv	DFEv	DTEv
8th grade	.838	.755	.918
9th grade	.846	.880	.927
10th grade	.742	.869	.901
11th grade	.828	.859	.929
12th grade	.849	.871	.912

It appears that each of the items on each scale have an acceptable amount of shared variance and are accepted as reliable scales.[1]

TABLE II

Hoyt's Analysis of Variance

	DPEv	DFEv	DTEv
Grand Rapids 10th grade	.791	.815	.836
Grade Rapids Deaf	.689	.783	.809

It appears that each item on the original scale has an acceptable amount of shared variance, and is

---

[1]Brookover, Wilbur B., Erickson, Edsel L., and Joiner, Lee M., Self Concept of Ability and School Achievement. East Lansing, Michigan: Michigan State University, 1967. Vol. III. Chapter II.

comparable to the data from the Michigan State Studies.

Whereas these reliability coefficients are not extremely high, they are relatively consistent. Definitions of teacher evaluations are consistently higher in each case, but this is probably due to the fact that teachers are constantly evaluating students. According to Guilford,[1] these reliability coefficients are probably an understimation of the reliability, since it is a short test. A perusal of Table II above will reveal a lesser amount of shared variance on the deaf scale. This is probably due to the smaller range of response on this scale. On the original scale, an individual could score from 1-5 on each item, and from 5-25 on the total complete scale. On the deaf scale, however, an individual could score from 1-3 on each item and from 5-15 on the total scale.

Analysis of Internal Consistency by a Rational Ordering of Items

According to Cureton,[2] "The most important re-

---

[1]Guilford, op. cit., p. 383.

[2]Cureton, Edward, "Quantitative Psychology as a Rational Science." Psychometrika, Vol. XI (1946), 191-196, as quoted by Jane Loevinger in "The Technic of Homogeneous Test Compared with Some Aspects of 'Scale Analysis." Psychological Bulletin, Vol. VL (1948), 507-529.

quirement for a test whose scores are to be interpreted as measurements would seem to be that test items all draw upon the same sets of abilities or traits." Tests of this type have been called reproducible scales, unidimensional test, or unified test. This type of analytic device seems more appropriate for the type of scales under investigation. Since Mead's theory is composed of rather obscure concepts, the operationalizing of these concepts could result in tapping several obscure dimensions of that concept. For example, a scale designed to measure a student's definition of the evaluation of others might be tapping expectations of others, surveillance of others or self definitions of academic ability if it is untested for unidimensionality. Conventional internal consistency analysis (eg: Hoyt's Analysis of Variance, or split halves) could result in a judgement of adequate reliability, when it is in fact not unidimensional.

If the above example were true, however, it would be inaccurate to refer to the concept operationalized in the scale as a single variable. One could not be precise in defining exactly what the scale measures. To avoid this pit fall, there have been several analytic techniques devised to determine whether or not a scale is unidimensional as required by Cureton's criterion.

As stated in the previous chapter, Guttman's

index of reproducibility is probably the most widely used and best known device used to determine unidimensionality. It is, however, subject to one deficiency. It is deficient to the extent that the value of the Reproducibility coefficient (R) is subject to the level of difficulty of the items (measured by persons passing or failing the item). Guttman recommends a reproducibility coefficient of .90 before a test should be considered scalable. The value of this coefficient may vary from test to test. Consider for example two tests, one in which 25% of the subjects pass all of the items, another in which 60% of the subjects pass all of the items. In the second instance, the minimum reproducibility is much higher than the first. Therefore, the value of the reproducibility coefficient is much less.

In spite of these limitations of Guttman's coefficient of reproducibility, an (R) value was computed.

TABLE III

Guttman's (R)	
DPEv	.93
DFEv	.95
DTEv	.95

This (R) value was computed even though not all of Guttman's recommendations were met. Guttman re-

commends that there be as many items below the .50 level of difficulty as above that level. He also recommends that there be about ten items. The "pass-fail" distribution also did not meet with Guttman's recommendations. Most of these limitations cannot be determined until after a Guttman type analysis has begun. Since these shortcomings detract from the value of (R), Green's index was also computed. Green's method is not affected by the level of difficulty of the items or the "pass-fail" distribution. This method also allows one to determine the extent to which the items might scale by chance, and it allows the investigator to set the confidence limits of the scale scores.

The general formula for the "obtained reproducibility" (Rep) is:

$$\text{Rep} = 1 - \frac{1}{NK} \sum_{i=1}^{K-1} n_{\overline{i-1}} + 1 - \frac{1}{NK} \sum_{i=2}^{K-2} n_{i-1, \overline{i}, i+1, i+2}$$

Where N is the number of subjects, K is the number of items in the test. The symbols n-i+1 is the number of subjects who fail the i^{th} item and pass the next most difficult item i+1. There are K-1 number of such pairs. The quantity $n_{\overline{i-1}, \overline{i}, i+1, i+2}$ is the number of subjects who have failed both item i-1 and $\overline{i}$ and passed both items i+1 and i+2. There are K-3 number of such terms.

Green has developed a method to determine the reproducibility that would be expected if the items exhibited zero covariance, i.e., were mutually independent. He labels this coefficient Rep_{ind}. By combining Rep and Rep_{ind} in his summary statistic, the perfectly reproducible test will have a value of 1.00 and for a test in which the items exhibit zero covariance, the value will be 0.00.

Green's formula for obtaining Rep_{ind} is:

$$Rep_{ind} = 1 - \frac{1}{N^2K} \sum_{i=1}^{K-1} n_i, n_i + 1 - \frac{1}{N^4K} \sum_{i=2}^{K-2} n_i\, n_{i+1},\, n_i+2$$

$$n_{i+1}$$

Green has developed a summary statistic to determine whether a test should be considered scalable. The formula for Green's summary statistic (I) is:

$$\frac{Rep - Rep_{ind}}{1.00 - Rep_{ind}}$$

Green recommends a value of .50 for (I) before a test should be considered scalable. The average discrepancy between Green's summary statistic (I) and the exact reproducibility of ten scales was .002,

according to an article cited by White and Saltz.[1]

Green's method also provides an approximation to the standard error of the obtained reproducibility (Rep). The formula for the standard error of Rep is:

$$\sigma_{Rep} \approx \sqrt{\frac{(i-Rep)\quad(Rep)}{NK}}$$

With this standard error, it is possible to determine the confidence limits within which the true value of Rep occurs. If a standard error of measurement value is appropriately small, it means that the sample statistics are close to the population parameter. When statistics are based upon a sample, it is possible that these sample values will be different than the actual value for the entire population. Hence, the need for an approximation of the standard error of measurement.[2]

TABLE IV

Green's Index of Reproducibility

	Rep	Rep_{ind}	σ Rep	I
DPEv	.920	.103	.009	.911
DFEv	.944	.113	.008	.937
DTEv	.939	.113	.009	.932

---

[1] White, Benjamin W. and Saltz, Eli, "Measurement of Reproducibility." Psychological Bulletin, Vol. 54, No. 2, (1957), 90.

[2] op. cit.

Definitions of Parental Evaluation Values from Table IV

The obtained reproducibility score of this scale (.920) is sufficiently above the minimum reproducibility (.50) recommended by Green. The value Rep_{ind} refers to the reproducibility value that would occur if the items were mutually independent. The standard error of Rep is referred to as σRep. This value means that chances are 99 out of 100 that a true score value for the population will be between .902 and .938.

Green's summary statistic (I), which takes into account both Rep and Rep_{ind} is .911. This value also greatly exceeds the minimum reproducibility (.50) recommended by Green. These scale values indicate that the items in Dpev-D are uni-dimensional.

Definition of Friends Evaluation Values from Table IV

The obtained reproducibility values of the Friends evaluation scale (.944) is also sufficiently above the minimum (.50) reproducibility. The standard error of Rep for this scale (.008) means that chances are 99 out of 100 that the true score value for the population will be between .921 and .953. Green's summary statistic (I) also greatly exceeds the minimum set up by Green. It appears that the items in this scale are also uni-dimensional.

Definitions of Teachers Evaluation Values from Table IV

The obtained reproducibility values from the scale (.939) and the summary statistic (.932) greatly exceed minimum requirements of (.50). The standard error of measurement value (.009) means that chances are 99 out of 100 that the true scale value for the population will be between .914 and .950. It appears that the items in this scale are also uni-dimensional.

Assessment of Equivalence

To assess the equivalence of the two forms of the scales under examination, a Pearson product moment correlation was computed and an examination of the distribution of the scores was made. The mean scores were not compared, since the range of scores differed for the different scales. The range of scores was 5-25 for the original scale and 5-15 for the revised scale for use with hearing impaired students.

TABLE V

Equivalence Correlations

DPEv	x	DPEv-D	r=.682
DFEv	x	DFEv-D	r=.744
DTEv	x	DTEv-D	r=.701

By squaring these coefficients of equivalence, it is possible to determine the percent of variance in either form of the scales that is associated with or predictable from measures of either variable. For Parental evaluations the explained variance is 46.5%, for Friends evaluations it is 55.4%, and for Teachers it is 49.1%. All of the above correlations were significant at the .05 level of significance.

To complete the equivalence analysis, an examination of the distribution of scores for each form was made.

In this assessment, a comparison was made to determine the extent to which each distribution of scores forms a mesokurtic distribution, i.e., a normal curve.

TABLE VI

Distribution of Scores: Parental Evaluations

	DPEv	DPEv-D
Range of scores	5-25	5-15
Mean	19.50	12.51
Median	19.07	13.26
Standard Deviation	3.12	1.72
Skewness	.003	.006
Kurtosis	.000	.001

The most meaningful comparisons between these two scales consist of a comparison of the skewness and kurtosis of the two distributions. A skewness value of .003 for the original scale and .006 for the revised scale indicates that the two distributions are almost symmetrical. There is slightly more positive skewness with the DPEv scale (more higher scores, the left tail of the curve extended), but the difference is negligable.

The skewness and kurtosis index shows that both instruments measuring the definition of Parent's evaluations, result in a normal distribution.

TABLE VIII

Distribution of Scores: Friends Evaluations

	DFEv	DFEv-D
Range of scores	5-25	5-15
Mean	18.66	12.23
Median	17.62	12.38
Standard Deviation	2.86	1.94
Skewness	.048	.000
Kurtosis	.000	.000

A skewness index of .048 for the original scale, and .000 for the revised scale, indicate that the distribution of scores are once again nearly symmetrical. The skewness and kurtosis values of each scale are nearly identical and reveal once again that definitions of Friends evaluations as measured by either scale, result in a normal distribution.

TABLE VIII

Distribution of Scores: Teachers Evaluations

	DTEv	DTEv-D
Range of scores	5-25	5-15
Mean	18.82	12.49
Median	18.25	12.07
Standard Deviation	2.95	1.94
Skewness	.007	.010
Kurtosis	.001	.002

Skewness values of this magnitude (.007 and .010) indicate that this distribution of scores is once again nearly symmetrical. There is slightly more positive skewness with the revised scale. However, the difference once again is to slight to warrant criticism.

The skewness kurtosis index shows that both of these scales measuring the definitions of Teacher evaluation result in a normal distribution. Thus, it is concluded that a student's definition of the evaluation of others is a normally distributed variable.

Assessment of Stability

An assessment of stability (test-retest reliability) by a correlational routine is very likely the worst index of reliability. In order to determine the stability of a measuring instrument, it is necessary to correlate two or more sets of scores obtained with the same instrument over a period of time. When engaging in this type of analysis, it is impossible to determine if a low correlation is the result of an unstable instrument, or if there was a change in the phenomena being measured. A high correlation may be the result of testing effect, i.e., the subject remembering the answers from the previous test. When some things are known about the population however, there may be some plausible explanations made about the stability coefficients. For example, if it is known that the phenomena being measured will not change from the first test to the second test, a correlation between the two tests may be more meaningful.

In addition to the possibility of a change in the phenomena being measured, there can also be a change in the testing procedure. This is the case with the instruments under investigation. The stability measures were done with the hearing impaired population and it was a new testing situation for the researchers, as well as for the subjects.

On the first testing, the pages on the test were not numbered and this presented quite a problem. With the hearing impaired sample, the subjects progressed through the test page by page. Whenever it was necessary to refer to the location in the questionaire, problems would arise. When the original form of the test was given to non-impaired subjects, it was not necessary to number the pages, since each student would answer the questions from beginning to end independently. One of the results of this oversight was to create considerable confusion and to lengthen the time required to complete the questionaire. For the second testing, the pages were numbered which may have reduced the confusion a great deal.

A second problem occuring on the first testing was the motion of the proctors. During the first testing, the proctors moved up and down the room to insure that the subject was responding to the proper item. The result of this motion was to visually distract the subjects. It was decided that many of the subjects "lost their place" because of this motion. For the second testing, the proctors placed themselves along side of the group where they could maintain eye contact, and moved about as little as possible. A lenghtening of the time required to complete the questionaire may also have been caused by this distraction. This length of time required to complete

a questionaire is a crucial point, since among some groups of exceptional children, there is a shortness of attention span.

It should also be noted that the hearing impaired students were not familiar with mass testing situations. In the first administration of the questionaire, complete testing took approximately one hour. In the second administration, complete testing took only 35 minutes. It is also possible that this quicker administration of the second test is due in part to the practice received by the subjects with the first test.

These stability coefficients should be accepted with caution, since these changes in the testing situation could be expected to have an adverse effect on the stability estimates.

In the following table are the stability coefficients found by correlating the results of the first testing with the results of the second testing.

TABLE IX

SCALE	TEST 1x TEST 2
DPEv-D	r = .000
DFEv-D	r = .571*
DTEv-D	r = .738*

*Significant at .05 level

The correlation between the first week and the second week on the DTEv-D scale is sufficiently high. However, one might surmise that these scales should evidence stability. One of the tasks of teachers is to evaluate the students. Teachers are regularly evaluating the students and informing them of how they are being evaluated by means of a "report card". For this reason, the student is more likely to evidence stability in his definitions of how his teachers are evaluating his academic ability.

The correlation between the DFEv-D scales from the first testing to the second testing is somewhat lower, but still indicates some stability. For the hearing impaired, it is much more of an effort to engage in a conversation with their friends than it is for non-hearing impaired students, since they must use sign and finger language. Also, the deaf and hearing impaired have a very limited vocabulary and probably do not discuss how they evaluate each other with their friends. This would result in a student holding a rather vague definition of how his friends evaluate his academic ability. It is also quite possible that the first testing was a new social-psychological experience for these students, and after being tested, they asked their friends how they were being evaluated. This would have the effect of lowering the stability

coefficient.

There is a zero correlation between the DPEv-D scales from the first week to the second week. This is also probably due to the restricted vocabulary of the hearing impaired student. In most cases, the amount of discussion between the hearing impaired child and his family on any subject, is extremely limited. The family does not have the special training required to communicate freely with the hearing impaired child, hence, a lack of discussion. It is also possible that parents do not evaluate the students academic ability the same as a teacher or another hearing impaired friend. A teacher or friend evaluates the students in relation to other hearing impaired students, but it is quite possible that a parent evaluates the hearing impaired child in relation to a non-hearing impaired population. This could also result in the student holding a vague definition of how his parents evaluate his academic ability. After being asked how their parents evaluate their academic ability in the first test, the student might have given this question a lot of thought and reappraised his definition of his parents' evaluations.

Since the changes in the testing procedure from the first administration of the test to the second administration would have an adverse effect on the

stability indices, one must conclude that the DTEv-D and the DPEv-D scales indicate some stability. The DPEv-D scale appears to be unstable, but it is impossible to determine if this low stability index is the result of the changes in the testing procedure due to some other phenomena, such as the plausible explanation posited above, or if it is actually the fault of the test.

VALIDITY

The purpose of validity analysis is to determine the extent to which a measuring instrument measures what it purports to measure. The validity analysis in this chapter will consist of Content Validity, Predictive Validity, Concurrent Validity, and partial evidence of Construct Validity. Less emphasis was placed on the validity analysis since there was considerable validity analysis done with the original scales.[1] There has been only limited reliability analysis done with the original scales. Hence, the emphasis on the reliability of these scales.

Assessment of Content Validity

In assessing the content validity or face validity of a scale, one asks the question: Do the items in this scale appear to measure what the researcher intends for them to measure? A close examination of the following scale should indicate an acceptable degree of face or content validity.

Definition of Parental Evaluation - Deaf

1. Think of your mother and father. Do your mother and father say you can do school work better, the same, or poorer than your friends?

---

[1] See the Brookover studies mentioned in Chapter I.

a. better
b. the same
c. poorer

2. Would your mother and father say you would be with the best, average, or below average students when you graduate from high school?

a. the best
b. average
c. below average

3. Do they think you could graduate from college?

a. yes
b. maybe
c. no

4. Remember, you need more than four years of college to be a teacher or doctor. Do your mother and father think you could do that?

a. yes
b. maybe
c. no

5. What grades do your mother and father think you can get?

a. A's and B's
b. B's and C's
c. D's and E's

The scales measuring the definitions of the evaluations of friends and teachers remain the same, except for the words mother and father. One would judge these other scales as exhibiting face or content validity.[1]

Assessment of Predictive Validity

Evidence of predictive validity is established

---

[1]See Appendix B for the scales measuring the students definition of his friends and his teachers evaluations.

with the existence of a significant relation between hypothesized independent and dependent variable.

According to Median theory, a student's definition of how others (primarily parents, friends, and teachers) evaluate his academic ability should be significantly related to his definition of his academic ability. The following table presents the stability correlations between the three scales under investigation and a scale designed to measure a student's self concept of academic ability (SCA).

TABLE X

Predictive Validity Correlations

DPEv-D	x	SCA	----------	r	=	.496
DFEv-D	x	SCA	----------	r	=	.585
DTEv-D	x	SCA	----------	r	=	.569

By squaring these validity coefficients, it is possible to determine the amount of variance in either testing with the scales that is associated with or predictable from either measure. For the SPEv-D scale, the explained variance is 24.6%, for Friends evaluations, it is 34.2%, and for Teachers evaluations, it is 32.4%. All of the above correlations were significant at the .05 level of confidence.

The above correlations are statistically significant.

However, they are lower correlations than one receives when the original scales are correlated with the self concept scale. This may be due to the reduced number of possible responses for each item. The original scale has five possible responses for each item, whereas, the revised scale has only three possible responses for each item. If the scale for use with hearing impaired students has less reliability, this could also have the effect of lowering the validity coefficients. If one accepts Hoyt's analysis of variance as an index of reliability for this scale, this revised scale will have a lower theoretical ceiling for the validity coefficient. For example, the reliability coefficient as determined by Hoyt's analysis of variance on the DPEv-D scale was .689. According to Thorndike and Hagen, the theoretical ceiling for a validity coefficient on the DPEv-D scale would be .830.[1]

Assessment of Concurrent Validity

Evidence of concurrent validity is established when there is a significant correlation with alternate forms of a test to a hypothesized dependent variable. With the present study, both the original and the

---

[1]See page 9, Chapter I for a discussion of the theoretical limit of validity, and page 36, Chapter III for the reliability coefficient used in the example above.

revised form of the test correlate with a student's self concept of academic ability. The emphasis is upon the common relationship to the SCA scale.

TABLE XI

Concurrent Validity Correlation

SCALE	COEFFICIENT
DPE x SCA	r = .621
DFEv x SCA	r = .630
DTEv x SCA	r = .630
DPEv-D x SCA	r = .496
DFEv-D x SCA	r = .585
DTEv-D x SCA	r = .569

*All significant at .05 level

It is evident that the original scale, for use with non-hearing impaired students, has a higher correlation with the self concept of ability scale. The revised form of the scale has a lower correlation with the SCA scale. The correlation between the Definitions of Parental evaluations and self concept of ability leave 61.5% of the variance unexplained with the original scale and 75.4% unexplained with the revised forms.

It is apparent from these data that the revised form of the test is somewhat less valid than the original

form of the test. This reduction in the validity coefficient could be due, once again, to the reduced number of possible responses for each item.

Assessment of Construct Validity

Construct Validity is an assessment of the complete theory underlying the construct under investigation. The construct under examination is basically drawn from a theory of George Herbert Mead. This theory may be stated in one sentence: An individuals self definition arises out of interaction with others and functions to direct his behavior.

This theory has been modified by Wilbur Brookover and others, and may be stated as follows: An individuals self definition of Academic Ability arises from a students definition of how others evaluate his academic ability and functions to direct his academic behavior.

From either statement of the theory, there are dozens of postulates that may be derived. A complete assessment of construct validity is a topic worthy of an entire study. Hence, for this investigation, an assessment of construct validity was limited to a restatement of concurrent validity and predictive validity.

A re-examination of these validity coefficients should reveal that the first part of the theory is

supported by empirical research. The theory remains valid, however it appears that the revised instrument used to test this theory is less valid than the original instrument.

CHAPTER IV

SUMMARY AND CONCLUSIONS

The principle task of this chapter will be to summarize this investigation and present the conclusions drawn from the analysis. Some suggestions for further research will also be presented.

Summary Problem Statement

In 1960, many sociologists, psychologists, and research specialists requested that high priority be given to the development of instruments for measuring social psychological factors of the deaf. They stressed the need for instruments which would yield comparable data with both hearing impaired and non hearing impaired children.

Heretofore, this has been a difficult task because of the differing populations, instruments, and research designs. This problem has been noted for the field of sociology of education by Wilbur Brookover and associates. Brookover contends that the value of his research is limited until it is tested upon exceptional children, since his research was based mainly on children without known impairments. One of the difficulties that arise when testing his theoretical models with

exceptional children is the problem of instrumentation. If an instrument is designed to measure a social psychological construct with non hearing impaired students, the language or wording used may not be comprehensible to the hearing impaired student. If the instrument wording is modified to cope with this limitation, there is no assurance that it will yield reliable and valid data.

Since the language or wording was changed for use with the hearing impaired children, the problem was a problem of translation or equivalence of statements. Do the questions tap the same phenomena when translated to another language?

One could argue that translation is involved whenever the researcher is not a member or a participant in the culture he is investigating. Also, according to R. Bruce W. Anderson,[1] "translation is involved whenever research requires asking the 'same' questions of people with differing backgrounds." This is true if the researcher is studying the same phenomena in two nations speaking the same language (such as the United Sates and Canada) or between subcultural

---

[1]Anderson, R. Bruce W., "On the Comparability of Meaningful Stimuli in Cross Cultural Research." Sociometry, Vol. 30 (June, 1967), 124-136 See also John Useem, "Notes on the Sociological Study of Language." Social Science Research Council Items, (September, 1963), 29-31.

groupings within one society (such as hearing impaired children and non hearing impaired children).

Brookover, Erickson, and Joiner have modified their original instruments to cope with this limitation. This revised instrument under analysis in this investigation consists of three sub scales designed to measure a students definition of the evaluations others have of him. These three scales, specifically are designed to measure: (1) a students definition of parental evaluations of his ability (2) a students definition of his teachers evaluation of his ability and (3) a students definition of his friends evaluation of his ability.

The problem covered specifically in this study is an assessment of the reliability and validity of these revised instruments. If these two supposedly parallel instruments are indeed assessing the same phenomena, it is hoped that a more definitive statement may be made concerning Brookover's theoretical models.

---

[1](con't) See also Herbert P. Phillips, "Problems of Translation and Meaning in Field Work." Human Organization, Vol. 18 (Winter 1959-60), 184-192. Joyce O. Hertzler, A Sociology of Language. Randon House: New York (1965), 128-131.

Summary of the Methods

An assessment of the reliability of an instrument involves determining the extent to which systematic and random errors are present in the scores of the measurement device. A situation where there is no error of measurement is a condition rarely found in behavioral research. Usually the errors of measurement fall into categories: (1) Where X_t = an obtained score, X_t may be thought of as a sum of (X_α), the true score under perfect conditions of measurement, and (X_e) an error component. $X_t = X_\alpha + X_e$ (2) Reliability may also be thought of in variance terms where (V_t), the total variance of scores, equals (V_α) the sum of true variance and (V_e), the variance due to error of measurement. $V_t = V_\alpha + V_e$.

There are three main types of reliability coefficients used in this investigation. These are: (1) an assessment of internal consistency, (2) an assessment of equivalence, and (3) an assessment of stability.

When constructing an instrument designed to measure an obscure and rather vague concept, one should be concerned that the instrument be a uni-dimensional test. In the scales under analysis in the present investigation, this test of uni-dimensionality is a crucial aspect.

There are four general approaches used to test this internal consistency. These are: face value, item analysis, ordering of items, and inter-test correlation.

Assessment of Internal Consistency

1. Analysis by Face Value

One indication of internal consistency is an examination of their face value. Do all the items tap the same underlying phenomena? This type of analysis is the lowest level of analysis and is almost always done because it springs from the nature of test construction.

2. Analysis by Item Analysis

One method that has been used with the original scales is Hoyt's Analysis of Variance. This was also used in the present investigation, despite some drawbacks to this approach. The object of this test is to determine whether or not the ratio of error variance is appropriately small.

3. Analysis by a Rational Ordering of Items

There were several types of scaling available to this researcher. However, several types were ruled out as being inappropriate. Thurstone's techniques were ruled out, since it was used in construction of

the original test. The problem was not one of selecting items, but determining if the items retain their "goodness". Guttman's technique was used for the sake of comparability. However, Green's method was also used because it appears to be a stronger index of internal consistency.

4. Analysis by Inter-Test Correlation

An assessment of internal consistency by split half techniques was considered inappropriate because of the scale characteristics. The scale consisted of only five items. This is too short for a split-half analysis according to Goode and Hatt.

Assessment of Equivalence

In an assessment of the equivalence of forms of the two parallel instruments, both forms were given to a non-hearing impaired group and a correlation was calculated. A comparison of means and standard deviations was ruled out because of the difference in the range of scores of the two instruments. Complete equivalence analysis involved a Pearson product moment correlation, and a comparison of the skewness and kurtosis of the distributions.

Assessment of Stability

In an assessment of stability or test-retest

reliability, there is an interval of time between the two administrations of the test. This is perhaps the worst type of reliability analysis, since if the correlations are not perfect, one cannot be sure if the test is unstable, or if there was a change in the phenomena being measured. Also, if the correlation is very high, one cannot be sure that "testing effect" was not operating.

An analysis of validity of an instrument involves an assessment of the degree to which the instrument really measures what it purports to measure. This study attempted to answer three types of validity and give partial evidence of a fourth type. The four types of validity investigated were: (1) content validity (2) predictive validity (3) concurrent validity (4) construct validity.

Content validity or face validity is also of the lowest power in validity analysis, but is almost always done because this also springs from the nature of test construction. This form of validity, while it is almost always done, should always be used with some form of empirical assessment.

Predictive validity refers to the prediction of a relationship between an independent variable and a hypothesized independent variable.

Concurrent validity is an assessment of the extent

to which alternate forms of an instrument predict the same event. An assessment is made between both forms and a common dependent variable. Both predictive and concurrent validity are sometimes classified as pragmatic validity.

Construct validity occurs whenever hypothesized relationships are empirically tested. This is a limited assessment of construct validity. In this analysis, the complete theory is to be evaluated. In the present study, however, an assessment of construct validity was limited to a re-examination of predictive validity and concurrent validity.

SUMMARY OF THE RESEARCH FINDINGS

Reliability

The analysis of internal consistency by face value seems to show that all of the items are measuring the same phenomena, and were accepted as being internally consistent based on this limited analysis.

The analysis of internal consistency by analysis of variance were consistent however, they were lower in every case for the revised scale. The coefficients were:

	DPEv	DFEv	DTEv
Non-hearing impaired	.791	.815	.836
Hearing impaired	.681	.783	.809

The reproducibility of the scales were well above the minimum suggested by both Guttman (.90) and Green (.50).

	Guttman's (R)
DPEv	.93
DFEv	.95
DTEv	.95

Green's Index of Reproducibility

	Rep	Rep_{ind}	σRep	I
DPEv	.920	.103	.009	.911

DFEv	.944	.113	.008	.937
DTEv	.939	.113	.009	.932

The coefficients of equivalence ranged from a low of .682 on parental evaluations, .701 on teacher's evaluations, and a high of .744 on friend's evaluations. The skewness and kurtosis also showed that both instruments form a normal curve.

The coefficients of stability ranged from a low of .000 on parental evaluations to .571 for friend's evaluations and a high of .738 for teacher's evaluations.

Validity

The analysis of content validity or face validity indicates that all of the items in the instrument are measuring a student's definition of the evaluations of others.

The coefficients of predictive validity were .496 for DPEv, .569 for DTEv, and .585 for DFEv in predicting self concept of academic ability.

Concurrent Validity Correlation

SCALE	COEFFICIENT
DPE X SCA	r = .621
DFEv x SCA	r = .630
DTEv x SCA	r = .630

DPEv-D	x	SCA	r	= .496
DFEv-D	x	SCA	r	= .585
DTEv-D	x	SCA	r	= .569

The limited analysis of construct validity is a re-examination of the predictive and concurrent validity coefficients. Campbell and Fiske maintain that an assessment of concurrent validity is the basic aspect of construct validity.

Suggestions for Further Research

This study was designed to evaluate the reliability and validity of two supposedly parallel instruments. The data generally supported the hypothesis that they were both reliable and valid. However, one of the major drawbacks to the stability analysis was the lack of familiarity of mass testing with hearing impaired students. In future research, more attention should be paid to test administration.

The revised scale might also be made comparable to the original scale. The correlation coefficients might be low because of the nature of the scales. For each item in the original scale, there was a range of 1 to 5 and a range for the whole test of 5 to 25. On the revised form, there was a range of 1 to 3 for each item and a range for the whole test of 5 to 15. If the range for each test were the same, it is possible that the correlations might be higher.

BIBLIOGRAPHY

American Psychological Association, "Technical Recommendations for Psychological Test and Diagnostic Techniques". Psychological Bulletin, VXI, no. 2 part 2 (special supplement, March, 1954).

Anderson, R. Bruce W., "On the Comparability of Meaningful Stimuli in Cross Cultural Research." Sociometry, XXX, (June, 1967), 124-136.

Bechtoldt, H., "Construct Validity a Critique." American Psychologist, XIV, (1959), 619-629.

Brookover, Wilbur, Erickson, Edsel L., and Joiner, Lee M., Self Concept of Ability and School Achievement. III, (East Lansing, Michigan, Michigan State University, 1966), Chapter II.

Brookover, et al Relationship of Self Concept to Achievement in High School. U.S. Office of Education Cooperative Research Project No. 2831, (East Lansing, Michigan State University, to be published in September, 1967).

Cronbach, L., Essentials of Psychological Testing. 2nd ed., New York: Harper and Row, 1960. p. 121.

Cronbach, L. J., "Response Set and Test Validity". Educational Psychology Measurement, VI, (1946), 475-494.

Cronbach, L., and Meehl, P. L., "Construct Validity of Psychological Test." Psychological Bulletin, LII, (1955), 281-301.

Cureton, Edward E., and Lindquist, E.F., (Ed), "Validity." Education Measurement, Washington D.C., American Council on Education, (1951), Chapter 16.

Cureton, Edward, "Quantitative Psychology as a Rational Science". Psychometrika, XL, (1946), 191-196.

Festinger, Leon, Katz, and Daniel, Research Methods in the Behavioral Sciences. New York: Holt, Rinehart, and Winston, Inc., 1953.

Geer, William C., and Deno, Evelyn D., "CEC and Legislation Now and in the Future." Exceptional Children, XXXII, (November, 1965), 187-194.

Goode, William J., and Hatt, Paul K., Methods in Social Research. McGraw-Hill, 1952.

Guilford, J.P., Fundamental Statistics in Psychology and Education. 2nd ed., New York: McGraw-Hill, 1950.

Guilford, J., Psychometric Methods. 2nd ed., New York: McGraw-Hill, 1954.

Jackson, Robert, Ferguson, and George, "Studies on the Reliability of Test." Bulletin No. 12, (Department of Education Research, University of Toronto, 1941).

Kerlinger, Fred N., Foundations of Behavioral Research. New York: Holt, Rinehart, and Winston, Inc., 1964.

Kerlinger, F., and Kaya E., "The Predictive Validity of Scales Constructed to Measure Attitudes Towards Education". Education and Psychological Measurement, XIX, (1959), 305-317.

Loevinger, Jane, "Technique of Homogeneous Test compared with some aspects of 'Scale Analysis' and 'Factor Analysis'". Psychological Bulletin, IVV, (1948), 507-529.

Loevinger, J., "A systematic Approach to the Construction and Evaluation of Test of Ability." Psychological Monograph, LXI, No. 285, (1947).

Remmers, H.H., Gage, N.L., and Rummels, J. Francis, A Practical Introduction to Measurement and Evaluation. 2nd ed., New York: Harper and Row, Pp. 129-130.

Rogers, Merrill, Quigley, and Stephen P., (eds), "Research Needs in the Vocational Rehabilitation of the Deaf." American Annals of the Deaf, CV, No. 4, (September, 1960), 335-370.

Selltiz, Claire, Marie Jahoda, Morton Deutsch, and Cook, Stuart W., Research Methods in Social Relations. New York: Holt, Rinehart, and Winston, Inc., 1963.

Stouffer, S., et al. Measurement and Predictions. Princeton: Princeton University Press, 1950.

Thorndike, Robert L., and Hagen, Elizabeth, Measurement and Evaluation in Psychology and Education. New York: John Wiley and Sons Inc., 1961.

White, Benjamin w., and Saltz, Eli, "Measurement of Reproducibility." Psychological Bulletin, (March, 1957), 83.

APPENDIX A

Please answer the following questions as you think your PARENTS would answer them. If you are not living with your parents, answer for the family with whom you are living.

Circle the letter in front of the statement that best answers each question.

1. How do you think your PARENTS would rate your school ability, compared with other students your age?

   a. Among the best
   b. Above average
   c. Average
   d. Below average
   e. Among the poorest

2. Where do you think your PARENTS would say you rank in your high school graduating class?

   a. Among the best
   b. Above average
   c. Average
   d. Below average
   e. Among the poorest

3. Do you think that your PARENTS would say you have the ability to complete college?

   a. Yes, definitely
   b. Yes, probably
   c. Not sure either way
   d. Probably not
   e. Definitely not

4. In order to become a doctor, lawyer, or university professor, work beyond four years of college is necessary. How likely do you think your PARENTS would say it is that you could complete such advanced work?

   a. Very likely
   b. Somewhat likely
   c. Not sure either way
   d. Somewhat unlikely
   e. Very unlikely

5. What kinds of grades do you think your PARENTS would say you are capable of getting in general?

   a. Mostly A's
   b. Mostly B's
   c. Mostly C's
   d. Mostly D's
   e. Mostly E's

APPENDIX A

Think about your closest friend at school. Now answer the following questions as you think this FRIEND would answer them.

Circle the letter in front of the statement that best answers each question.

1. How do you think this FRIEND would rate your school ability compared with other students your age?

   a. Among the best
   b. Above average
   c. Average
   d. Below average
   e. Among the poorest

2. Where do you think this FRIEND would say you would rank in your high school graduating class?

   a. Among the best
   b. Above average
   c. Average
   d. Below average
   e. Among the poorest

3. Do you think that this FRIEND would say you have the ability to complete college?

   a. Yes, definitely
   b. Yes, probably
   c. Not sure either way
   d. Probably not
   e. Definitely not

4. In order to become a doctor, lawyer, or university professor, work beyond four years of college is necessary. How likely do you think this FRIEND would say it is that you could complete such advanced work?

   a. Very likely
   b. Somewhat likely
   c. Not sure either way
   d. Somewhat unlikely
   e. Very unlikely

5. What kind of grades do you think this FRIEND would say you are capable of getting in general?

a. Mostly A's
b. Mostly B's
c. Mostly C's
d. Mostly D's
e. Mostly E's

APPENDIX A

Think about your favorite teacher--the one you like best; the one you feel is most concerned about your schoolwork. Now answer the following questions as you think this TEACHER would answer them.

Circle the letter in front of the statement which best answers each question.

1. How do you think this TEACHER would rate your school ability compared with other students your age?

   a. Among the best
   b. Above average
   c. Average
   d. Below average
   e. Among the poorest

2. Where do you think this TEACHER would say you would rank in your hugh school graduating class?

   a. Among the best
   b. Above average
   c. Average
   d. Below average
   e. Among the poorest

3. Do you think this TEACHER would say you have the ability to complete college?

   a. Yes, definitely
   b. Yes, probably
   c. Not sure either way
   d. Probably not
   e. Definitely not

4. In order to become a doctor, lawyer, or university professor, work beyond four years of college is necessary. How likely do you think this TEACHER would say it is that you could complete such advanced work?

   a. Very likely
   b. Somewhat likely
   c. Not sure either way
   d. Somewhat unlikely
   e. Very unlikely

5. What kind of grades do you think this TEACHER would say you are capable of getting in general?

a. Mostly A's
b. Mostly B's
c. Mostly C's
d. Mostly D's
e. Mostly E's

APPENDIX B

Pretend you are your mother or father. Answer like they would. Pick one. Circle their answer.

1. Think of your mother and father. Do your mother and father say you can do school work better, the same, or poorer than your friends?

   a. better
   b. the same
   c. poorer

2. Would your mother and father say you would be with the best, average, or below average students when you graduate from high school?

   a. the best
   b. average
   c. below average

3. Do they think you could graduate from college?

   a. yes
   b. maybe
   c. no

4. Remember, you need more than four years of college to be a teacher or a doctor. Do your mother and father think you could do that?

   a. yes
   b. maybe
   c. no

5. What grades do your mother and father think you <u>can</u> get?

   a. A's and B's
   b. B's and C's
   c. D's and E's

APPENDIX B

Pretend you are your best friend. Answer like he or she would. Pick one. Circle their answer.

1. Think of your best friend. Would your best friend say you can do school work better, the same. or poorer than other people your age?

   a. better
   b. the same
   c. poorer

2. Would your best friend say you would be with the best, average, or below average students when you graduate from high school?

   a. the best
   b. average
   c. below average

3. Does your friend think you can graduate from college?

   a. yes
   b. maybe
   c. no

4. Remember, you need more than four years of college to be a teacher or doctor. Does your best friend think you could do that?

   a. yes
   b. maybe
   c. no

5. What grades does your best friend think you can get?

   a. A's and B's
   b. B's and C's
   c. D's and E's

APPENDIX B

Pretend you are your teacher, the one you like best. Answer like he or she would. Pick one. Circle their answer.

1. Think of your teacher. Would your teacher say you can do school work better, the same, or poorer than other people your age?

   a. better
   b. the same
   c. poorer

2. Would your teacher say you would be with the best, average, or below average students when you graduate from high school?

3. Does your teacher think you could graduate from college?

   a. yes
   b. maybe
   c. no

4. Remember, you need more than four years of college to be a teacher or doctor. Does your teacher think you could do that?

   a. yes
   b. maybe
   c. no

5. What grades does your teacher think you can get?

   a. A's and B's
   b. B's and C's
   c. D's and E's