AN ABSTRACT OF THE THESIS OF

KAREN REED SZULEWSKI for the degree of Master of Science

in Family Life (Child Development) presented on March 16, 1976

Title: HOMOGENEITY OF TEST ITEMS IN THE FILM

TEST FOR UNDERSTANDING BEHAVIOR

Redacted for privacy

Approved: ______________________________

Dr. J. Phillip O'Neill

The objective of this study was to determine the internal consistency rating for the Film Test for Understanding Behavior (FUB) and the three subscales within the test. An estimation formula which contrasts sums of squares generated from an analysis of variance was used to compute the internal consistency reliability coefficients for the total test score and each of the three subscale scores: Knowledge, Guidance and Sensitivity.

Data for the analyses were provided from the Film Test scores of 321 students who enrolled in a beginning child development course during the 1971-1972 and 1972-1973 academic years. The students were predominantly female (95%) undergraduates (99%) and reported a major in some area of Home Economics (61%).

The total test reliability was estimated at r = .77. The stringency of the internal consistency analysis as well as the general homogeneity of the sample would allow a conclusion that overall, the

FUB has demonstrated an adequate degree of reliability when testing students at this level. The magnitude of this coefficient is consistent with prior reliability coefficients established for the FUB using the test-retest method. The estimated reliability coefficients for the subscales were as follows: Knowledge (r=. 33), Guidance (r=.73) and Sensitivity (r=. 27).

The chronological development of reliability theory and the different methods available for estimating reliability have been reviewed, as well as the history and present status of the FUB with respect to reliability and validity.

Homogeneity of Test Items in the Film Test for Understanding Behavior

by

Karen Reed Szulewski

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

June 1976

APPROVED:

Redacted for privacy

Professor and Head of Department of Family Life
in charge of major

Redacted for privacy

Dean of Graduate School

Date thesis is presented March 18, 1976

Typed by Opal Grossnicklaus for Karen Reed Szulewski

ACKNOWLEDGEMENTS

So many of my friends and family have given me their support and encouragement during the process of completing my thesis. It is a joyous occasion for me to thank each one now for their individual contribution.

This thesis is dedicated to Barbara Mahler who as a teacher and friend has shown me new worlds to explore.

Dr. J. Phillip O'Neill as major professor has given much time and energy to aid in the completion of this project and I am most grateful. Acting as minor professor, Dr. Theodore Madden suggested the statistical device which was eventually used in this thesis and has contributed much advice during the preparation of the manuscript.

I would like to thank K. V. Rao for proofreading and Ann Burrows for her guidance and friendship which I treasure.

My parents, Paul and Emily Reed and my grandmother, Roberta Knight, have given their love and financial support which is much appreciated. Special thanks go to my sister Janet who is always there when I need her.

No words can express all the feelings I have at this time. It is the ending of an episode in my life and the beginning of a new one. Paul Szulewski, my husband and friend, has shared these experiences and our life together has been so rich.

I am so glad you are all part of my life.

TABLE OF CONTENTS

INTRODUCTION 1

REVIEW OF LITERATURE 3

- Chronological Review of Reliability Theory 3
- Theory of True and Error Scores 15
- Eclectic Concept of True Scores and Parallel Tests 17
- Domain Sampling Model 18
- Error Variance and its Effect on Reliability 19
  - Administration 21
  - Scoring 22
  - Guessing 22
- Methods of Estimating Reliability 23
  - Test-Retest Method 23
  - Parallel Forms 25
  - Subdivided Test Method 26
  - Internal Consistency Methods 27
- The Film Test for Understanding Behavior 30
  - History and Development 30
  - Administration of the Test 33
  - Standardization 33
  - Validity 34
  - Reliability 38

METHOD 40

- Subjects 40
- Procedure 43
- Analysis 44

LIST OF TABLES

Table		Page
1.	Description of the College Student Training Sequence Used in the Standardization of the FUB.	34
2.	Comparison of FUB Subscale Means for the College Student Training Sequence.	36
3.	Comparison of Professionals and College Students' (TL5) FUB Scores.	37
4.	Test-Retest Reliability Coefficients for the FUB.	38
5.	A Description of FL 225 Students by Year and Term.	40
6.	Distribution of 1971-72 and 1972-73 FL 225 Students by School and Academic Major.	42
7.	Frequency Distribution of FL 225 Students by Number of Courses in Related Areas.	43
8.	Analysis of Variance on Scores from the Film Test for FL 225 Students: Low-Medium Key.	48
9.	Analysis of Variance for Film Test Scores for FL 225 Students: Knowledge Subscale.	49
10.	Analysis of Variance for Film Test Scores for FL 225 Students: Guidance Subscale.	51
11.	Analysis of Variance for Film Test Scores for FL 225 Students: Sensitivity Subscale.	52
12.	A Summary of Reliability Coefficients with Corresponding Standard Errors for the Film Test for Understanding Behavior and its Subscales.	53

HOMOGENEITY OF TEST ITEMS IN THE FILM TEST FOR UNDERSTANDING BEHAVIOR

INTRODUCTION

Certain conditions must be fulfilled before the data obtained from a measuring instrument can be used in practical situations. First, one must provide evidence that the instrument measures what it purports to measure which is a question of validity. Second, the instrument must give a reliable measurement so that similar results will be obtained if the trait is remeasured under similar conditions.

This study is concerned with a reliability analysis of the Film Test for Understanding Behavior (Schalock and Edling, 1958) using the Analysis of Variance model suggested by Kerlinger (1973). The film test attempts to measure responses to children's behavior by incorporating some of the emotional involvement of an interpersonal situation while allowing one to measure behavioral understanding in a classroom situation rather than in a personal encounter with young children. The Analysis of Variance is chosen to estimate the reliability of the film test because it is a powerful yet simple statistical device which can be designed to provide the investigator with information beyond the reliability coefficient.

Since the idea of using motion picture films as a testing device is relatively new there is little research on the reliability of this type of measurement. Past research on the reliability of the Film

Test for Understanding Behavior (FUB) has been limited to the test-retest method using small samples. This study will provide further evidence of the level of reliability of the FUB so that future data generated from this instrument could be interpreted with a known degree of certainty. The Analysis of Variance model will provide an internal consistency reliability coefficient which will supplement the reliability information already available while setting a precedent for future departmental research on the Film Test.

The review of literature will be divided into three sections the first of which is a chronological review of the literature on reliability theory focusing on the theoretical aspects of reliability. By limiting the discussion to the conceptual rather than mathematical side of reliability the review traces the development of concepts rather than formulas, however, some equations are presented to aid in comprehension of the material. This section concludes by presenting three conceptual models which summarize the various positions of authors included in the review.

Because the concept of measurement error is basic to all conceptual models concerned with reliability the second part of the review of literature will examine the types of measurement error and their relationship to reliability. The final section of literature review traces the history and development of the Film Test for Understanding Behavior.

REVIEW OF LITERATURE

Chronological Review of Reliability Theory

Spearman developed the original assumptions and constructs of reliability theory in two articles published in 1904 and 1910. Spearman's formulation of reliability theory have been called the 'truth-error factor theory' by Tryon (1957, p. 244). This is a useful label for comparing this original formulation to the later developments in reliability theory.

According to Spearman (1910), a score on a test may be considered to contain a true component and an error component. The relationship of these components is stated in the following equation:

$$X_t = X_\infty + X_e \qquad (1)$$

X_t = observed test score

X_∞ = true score component

X_e = error score component.

This equation may be extended algebraically (Ghiselli, 1964, p. 275) to:

$$V_t = V_\infty + V_e \qquad (2)$$

Vt = observed score variance

V_∞ = true score variance

V_e = error score variance.

The formula in variance terms is more useful when comparing Spearman's truth-error model to other theoretical stances and in discussing the relationship of reliability to test variance.

The true score component (X_∞) represents the actual ability of the subject, a quantity which theoretically should remain stable from test to test assuming there is no change in the behavioral domain. True score variance (V_∞) represents the distribution of true scores around the true score mean for a particular test. This term cannot be estimated directly but it is possible to estimate the variance of the error component (V_e). The error component (X_e) is a result of various factors which cause a subject to sometimes answer correctly a question he does not know or to answer incorrectly a question he does know. Spearman assumed these errors to be of zero mean, of uniform variance, independent of each other, and independent of the true score.

Working on similar ideas Brown (1910) began by defining parallel tests rather than the true score plus error assumptions of Spearman. Brown requires parallel tests to be equivalent in content, mean, variance and intercorrelations while Spearman (1910) requires that parallel measures of trait, x, be so alike that any difference ". . . be regarded as quite accidental." (p. 274).

Brown (1910) differed with Spearman on the truth-error postulate because he believed it incorrect to label the difference in scores on two parallel tests as error variance. Brown postulated that each measure is an index of that individual's ability on a particular day. He did say, however, that there is probably a mean ability level. He considered an individual's variance from his mean ability due to individual variability rather than measurement error. Thus, Brown does not assume that errors are uncorrelated with each other, nor does he assume that they are uncorrelated with the mean ability level. Brown's concept of the mean ability level is similar to Spearman's concept of true score.

Kuder and Richardson (1937) approached the computation of reliability in a different manner, using the number of items, item difficulty and the total test variance. They arrived at an internal consistency estimate of reliability for dichotomously scored items which is referred to as KR-20. The KR-20 formula is derived under assumptions that the matrix of inter-item correlations is of rank one and that all intercorrelations are equal:

$$r_{11} = (n/n-1)\ s_t^2 - n\ \overline{p\,q}/s_t^2 \qquad (3)$$

where,

r_{11} = total test reliability coefficient

n = number of items

s_t^2 = variance of total scores on test

p = percentage of persons who pass each item

q = 1 - p.

When the test contains multi-point items rather than dichotomous items another estimation formula, KR-21, is required. The KR-21 formula requires the number of items in the test, the total test variance, and the mean of the total scores. Besides the assumptions of KR-20, it is assumed that all items have the same difficulty:

$$r_{11} = (n/n-1)[(s_t^2)-n\,\overline{p}\,\overline{q}]/s_t^2 \qquad (4)$$

KR-21 results in the same estimate of reliability as "Coefficient Alpha" an estimation formula developed by Cronbach (1951).

Rulon (1939) developed a procedure which would allow the computation of a reliability coefficient without the assumption of equal variance between subtests. Assuming that the error variance comes completely from variance in the subtests he calculates the difference between the half test scores by subtracting each individual score on subtest A from his score on subtest B. He then obtained the variance of the distribution of obtained differences which can be used to estimate a total test's reliability using the equation:

$$r_{11} = 1 - s_d^2/s_t^2 \qquad (5)$$

where,

r_{11} = total test reliability

s_d^2 = variance of the differences computed by subtracting test A from test B

s_t^2 = total test variance.

Up to this point most formulas for estimating the reliability of a test were derived from assumptions of Spearman's truth error-theory or Brown's theory of equivalent forms. In 1931 Cureton introduced a new approach based on variance ratios to estimate reliability. In this approach reliability is seen as the ratio of true-score variance to the total variance of the test. True variance then is the variance due to individual differences and not due to measurement error. Most of the early work on reliability theory based on variance ratios was done by Jackson (1939) who was concerned with assessing the precision of a measurement device.

Hoyt (1941) extended the concepts developed by Jackson to internal-consistency analysis. He began by stating Rulon's short method of estimating reliability by means of the split-half method and pointed out that depending on the nature of the odd-even split, (s_d^2) may be an underestimate of the discrepancy between the obtained variance and the true variance. Hoyt proposed the Analysis of Variance model to give a better estimate of this discrepancy. Using the least squares criterion he showed that the analysis of variance model gives the best linear estimate of the discrepancy between the obtained and true score. Further he showed that his model yields the same reliability coefficient as the KR-20 while making it possible to separate out the variance due to practice effects when two tests are administered. In this study, Hoyt's model as described by

Kerlinger (1973), is the method used to estimate the reliability of the Film Test for Understanding Behavior.

Guttman (1945) derived an equation from Rulon's original formula which gives identical results to that of Rulon's model but is easier to compute:

$$r_{11} = (2)\ \ [(1) - (s_A^2) + (s_B^2)/s_t^2] \qquad (6)$$

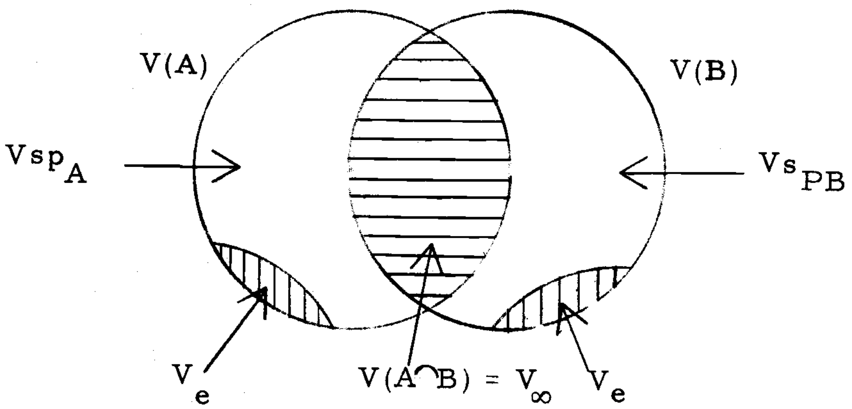
where,

r_{11} = total test reliability

s_A^2 = variance on subtest A

s_B^2 = variance on subtest B.

Tryon (1957) extended reliability theory to explain the connection between reliability and validity by developing the content-domain model. The figure below, from Kerlinger (1973a), may aid in the discussion of the relationship of reliability and validity.



(7)

Validity, like reliability may be expressed in variance terms:

$$Val = V_{co}/V_t \qquad (8)$$

$$r_{11} = V_{\infty}/V_t \tag{9}$$

where,

Val = validity

V_{co} = common factor variance

V_t = total variance of a measure

r_{11} = total test reliability

V_{∞} = true variance.

Validity is the ratio of common factor variance (V_{co}) to total test variance (V_t) with common factor variance (V_{co}) defined as the variance that two or more tests have in common. Specific variance (V_{sp}) is the systematic variance of the measure not shared with any other measure and V_e is the error variance (Kerlinger, 1973). Total test variance may be expressed as:

$$V_t = V_{co} + V_{sp} + V_e \tag{10}$$

Dividing each of the above terms by the total variance provides an equation which can be used to show the close variance relationship between reliability and validity:

$$\frac{V_t}{V_t} = \frac{V_{co}}{V_t} + \frac{V_{sp}}{V_t} + \frac{V_e}{V_t} \tag{11}$$

Validity was defined earlier as (V_{co}/V_t) and can be further defined as:

$$\frac{V_{co}}{V_t} = \frac{V_t}{V_t} - \frac{V_{sp}}{V_t} - \frac{V_e}{V_t} \tag{12}$$

Reliability may be seen as:

$$r_{11} = \frac{V_t}{V_t} - \frac{V_e}{V_t} \qquad (13)$$

If equation (12) is rewritten as:

$$\frac{V_{co}}{V_t} = \frac{V_t}{V_t} - \frac{V_e}{V_t} - \frac{V_{sp}}{V_t} \qquad (14)$$

it can be seen that reliability is equal to the first two right hand terms of (14):

$$R_{11} = \frac{V_t}{V_t} - \frac{V_e}{V_t} = \frac{V_\infty}{V_t} \qquad (15)$$

and validity equal to:

$$\frac{V_{co}}{V_t} = \frac{V_\infty}{V_t} - \frac{V_{sp}}{V_t}. \qquad (16)$$

Validity, then is the portion of the total variance of the measure that shares variance with other measures. As mentioned earlier Tryon extended reliability theory to show its relationship to validity while developing standard formulas for estimating reliability based on operations employed in objectively sampling behavior. He rejects Spearman's truth-error theory as well as Brown's model of equivalent forms because they are based on what he considers to be unnecessary and restrictive assumptions. Specifically, he does not accept the true score concept as previously defined by Brown (1910) and Spearman (1904, 1910); and states that test samples may have different variances and covariances.

In the content-domain model a test is composed of n items taken from a hypothetical domain of items. The total score (X_t) is computed as:

$$X_t = X_1 + X_2 + X_3 \ldots + X_n. \qquad (17)$$

Reliability, then, is the Pearson Product Moment Correlation between (X_t) and a comparable test, ($X_{t'}$). A comparable test according to Tryon (1957) ". . . is one whose n test-samples vary on the average as much in standard deviations and inter-correlations as do the n test-samples in the observed (X_t) composite." (p. 231). Tryon considers the correlation of comparable forms to be the "General Form" of reliability and delineates alternative computing formulas which can be used in special situations. He also developed what he calls a complementary "basic definition" of reliability which he states in variance terms:

$$r_{11} = 1 - V_{o_t}/V_t \qquad (18)$$

where,

V_{o_t} = within individual variance

V_t = total test variance.

Cronbach (1963) further refined reliability theory by incorporating the analysis of variance model into a new interpretation of reliability theory which he calls the Theory of Generalizability. According to Cronbach when an investigator is interested in the reliability of a measurement device he is really asking how well information given by one test may be generalized to other areas of performance related to the test. For example, consider the score on an essay test. The researcher may want to generalize from the scorer's judgement of the quality of this one paper to his judgement of other essay papers he grades or he may wish to generalize from the examinee's score on this essay test to other papers by the same subject. Thus, generalizability can be looked at from two perspectives, the class of observations or the measuring technique which generates the observations. When the researcher is interested in assessing the reliability of a measurement technique he conducts a "Generalizability Study" (G Study). When he wishes to make decisions based on the observations from the measuring technique he conducts a "Decision Study" (D Study). It is possible for the same study to be used for both purposes.

Generalizability theory requires that the universe being studied be clearly defined. This is necessary since the scores on a measuring device may be used to generalize to other universes. After clearly defining the universe the investigator must then specify a universe of conditions of observations over which he wishes to generalize. "Conditions is a general term referring to particular test forms or stimuli, observers, occasions, situations of observation" (Cronbach, 1963, p. 145).

The following assumptions are stated by Cronbach:

1. The universe is defined unambiguously so that the number of conditions which fall within the universe are known.

2. Conditions are experimentally independent.

3. Scores X_{pi} are numbers on an interval scale (X_{pi}) is the observed score for person p on test i) (p. 145).

Included in Generalizability Theory is the concept of a universe score, which is also referred to as a true score, (M_p) defined as ". . . the mean of X_{pi} over all conditions in the universe" (p. 145).

Where Spearman begins by assuming:

$$X_t = X_\infty + X_e \tag{1}$$

Cronbach sees (X_t) (Cronbach's X_{pi}) as the sum of four components:

$$X_{pi} = M + (M_p - M) + (M_i - M) + e_{pi}. \tag{19}$$

where,

X_{pi} = observed test score (same as Spearman's X_t)

M = mean of the M_p or the M_i (the grand mean)

M_p = true score or universe score (same as Spearman's X_∞)

M_i = the mean score on test i over all persons in the population

e_{pi} = the error term, assumed to be independently distributed

This form corresponds with the additive model in analysis of variance. Reliability then would be considered the ratio of true score variance (V_{M_p}) to observed score variance (V_i).

Using the preceding assumptions Cronbach develops formulas to estimate the interclass correlation coefficient when two tests are available for each subject and an internal consistency estimation formula which may be used when one test or rating is available for each subject. It is important to mention that Cronbach derives his formulas in such a manner to account for samples from populations while all other formulas were developed for populations only.

A two way analysis of variance leads to a between persons mean square (MS_p) and a residual mean square (MS_r). These mean squares are used to arrive at the true score variance (V_{M_p}) by subtracting MS_r from MS_p. The reliability coefficient is estimated using the following formula:

$$\rho M_x = V_{M_p}/V_x \tag{20}$$

where,

ρM_x = intraclass correlation coefficient

V_{M_p} = true score variance

V_x = total test variance

The above coefficient is used when two tests are available while the following equation is used to estimate the internal consistency reliability of a measurement when only one test or rating is available:

$$\alpha_{(ni)} = (MS_p - MS_r) / MS_p \qquad (21)$$

where,

$\alpha_{(ni)}$ = internal consistency reliability coefficient

MS_p = between persons mean square

MS_r = residual mean square

This completes the chronological review of the theoretical developments in reliability theory. Ghiselli (1964) reviewed these same theoretical developments and suggested that three theoretical models could be distinguished which would represent the different schools of thought concerning reliability theory. The three models he presents are the theory of true and error scores, the eclectic theory of true scores and parallel tests, and the domain sampling model. Each of these models will be briefly reviewed in the following sections.

Theory of True and Error Scores

Basic to this model is the equation stated earlier in presenting Spearman's original work in reliability theory.

$$X_t = X_\infty + X_e \qquad (1)$$

where,

X_t = observed or fallible score earned on one administration of a test

X_∞ = true score

X_e = error score

Ghiselli (1964) states three assumptions which he believes necessary in this model:

1. The individual possesses stable characteristics or traits that persist through time

2. Errors are completely random

3. Fallible or observed scores are the result of the addition of true and error scores (p. 221).

In this model parallel tests would measure the same true score, consequently the means and standard deviations of true scores are the same on all parallel tests. Since the errors are completely random the error scores for any one individual equal the distribution of error scores for any other individual. This means when the number of parallel tests (k) is large, the standard deviation of the error scores of each person over k tests is equal to the standard deviation of the error scores for every other person and the standard deviations of all parallel tests are then equal. Also, since errors are assumed to be random, they are assumed to have a mean of zero. It can be proved algebraically that the mean of the fallible scores on all parallel tests are equal to the mean of the true scores on these parallel tests (Ghiselli, 1964, p. 233-234).

In the true and error score model reliability is seen as the variation in an individual's scores over a series of parallel tests and is representative of Spearman's (1910) original formulation of reliability theory.

Eclectic-Concept of True Scores and Parallel Tests

While the true and error score model is representative of Spearman's conceptualization of true score and error concepts the Eclectic model closely resembles Brown's (1910) work on reliability theory.

True scores in the Eclectic model are seen as the average score of an individual on an infinite number of tests. In this model parallel tests measure the same trait and have precisely the same pattern of correlations with a wide variety of other tests. It is assumed that all parallel tests have equal means and standard deviations but it is not stated clearly what these assumptions are based on. The parallel tests can be made to have equal means and standard deviations by standardizing scores but tests may be parallel without meeting these criteria.

In this Eclectic model, as in the true and error score model, reliability is seen as the proportion of individual or true score variance to the total test or fallible score variance (Ghiselli, 1964).

Domain Sampling Model

According to this model when a trait is defined, in effect one is describing a domain of behavior, that is, categories of behavior, which all have some property in common. To measure the trait an instrument is developed which prompts behaviors in this behavior domain. The measuring instrument does not tap all the behaviors in the domain but merely a sample of them. Reliability is seen as the extent of variation among an individual's scores on a number of different comparable samples of items or situations. It may also be seen as the correlation between scores on two or more samples from a behavior domain. In this model it is not necessary to assume that parallel tests have equal means and standard deviations to establish reliability but it is assumed that parallel tests do generally meet these criteria.

True scores are defined as the sum or average of an individual's scores on an infinitely large number of representative samples or parallel tests. The mean of true scores is equal to the mean of fallible scores on any one parallel test and the variance of true scores is equal to the variance of fallible scores times the reliability coefficient.

In all three models true scores are defined differently and each is developed from different assumptions, yet in each case the same basic definition of reliability results. This definition being the ratio of true score variance to fallible or observed score variance. True scores are seen as having the same mean as fallible scores in all three models. The various ways of estimating reliability of measurement may appear unnecessary as they seem to arrive at the same basic definition of reliability. However, they are based on different assumptions and disagree on the definition of important concepts such as true score and error.

Because the concept of measurement error is basic to all conceptual models concerned with reliability this concept will be discussed as well as the types of measurement error and their relationship to reliability.

Error Variance and its Effect on Reliability

In this section of the review of literature the discussion will center on the relationship of reliability to error of measurement along with the discussion of three major sources of error: administration, guessing, and scoring. Error of measurement is considered to be anything which causes an individual to answer a question correctly that he does not know or to answer incorrectly a question he does know (Magnusson, 1967).

The error component for a certain individual can be seen as the sum of a number of different error components which are the result of a number of sources of error. These error components are assumed to be independent and hence uncorrelated. For a number of individuals we can form a distribution for a given source of error. The variance of the distribution in which every individual's total error component is included will be equal to the variance of the sum of the distributions for each error source (Magnusson, 1967).

It is not possible to compute the error variance directly; thus it must be estimated. The computation of the reliability coefficient (r_{11}) is based on the estimate of the variance of the distribution of the total error component (s_e^2). The following formula states the relationship between reliability and error variance:

$$r_1 = 1 - s_e^2/s_t^2 \tag{22}$$

where,

r_{11} = total test reliability

s_e^2 = error variance

s_t^2 = total test variance

For perfect reliability $R_{11} = 1$, that is, the error variance is zero.

Different practical methods of estimating reliability arrive at different estimates of the size of the error variance. This is due to the fact that each method is affected by different sources of error (Nunnally, 1970). In some of the methods a true component is included in the error variance which results in an overestimate of s_e^2 and an underestimate of reliability. In other methods part of the error component serves as true score, leading to an underestimate of s_e^2 and an overestimate of the reliability of the measurement device being evaluated. For this reason, the assessment of a reliability coefficient depends on the methods used and the type of variance the particular method includes in error variance (Nunnally, 1970).

Error variance is a composite of many sources of error all of which contribute to a test's unreliability. The three major sources of error, administration, guessing, and scoring are discussed in the following paragraphs.

Administration

The administrator of a test can influence the degree of measurement error introduced during the testing process (Magnusson, 1967; Nunnally, 1970; Aiken, 1971). This is more likely in the case of individually administered tests than group administered tests. When working in a one to one situation, the test administrator can make conditions conducive to optimal performance by adapting his behavior. The extent of outside disturbance and the type of surroundings in which the test is taken may also introduce measurement error. Further, the instructions given to an individual may also introduce error if they are ambiguous since the testee may interpret the questions differently on different occasions. It is also important to give the same instructions to all individuals taking the same test so that the scores from the test may be comparable because the subjects were given equivalent information. Today most standardized tests include a set of specific instructions which are to be read to the testee before taking the test to insure the comparability of scores (Magnusson, 1967; Nunnally, 1970; Aiken, 1971).

Scoring

Error variance is introduced in scoring when the correctness of a response is left to the judgement of an individual. This is more often a problem with individual rather than group testing since most group tests are designed as multiple choice and the correctness of the answer is not left to the judgement of an individual. In individual testing unique responses from the testee must be subjectively rated by the testor (Cronbach, 1970).

Guessing

In multiple choice tests several alternative answers are given for questions, one of which is correct. If an individual does not know the correct answer, he still has a chance of guessing the right answer. Because of guessing the person taking the test has a possibility of obtaining a plus when he should have scored zero. This is an example of pure measurement error (Magnusson, 1967; Nunnally, 1967; Cronbach, 1970).

This concludes the discussion of the major sources or error involved in testing. Each method of estimating a reliability coefficient controls for different sources of error and in the following section each method will be discussed in terms of the error sources it includes.

Methods of Estimating Reliability

This discussion will focus on the specific methods of computing reliability namely, test-retest, parallel forms, subdivided test, internal consistency estimates and analysis of variance. When presenting the error sources each method includes, Magnusson's (1967) notation is used since it provides a common notation which aids in comparison.

Test-Retest Method

This method involves the administration of the same test on two occasions. The correlation of the two scores is then an estimate of the reliability coefficient (Nunnally, 1970).

Assuming that the terms are uncorrelated, the composition of the total test variance for the test-retest method can be stated as:

$$s_t^2 = s_T^2 + s_{e(m)}^2 + s_{e(adm)}^2 + s_{e(g)}^2 + s_{e(subj)}^2 + s_{T(fl)}^2 \qquad (23)$$

where,

s_t^2 = total test variance

s_T^2 = true variance

$s_{e(m)}^2$ = error variance due to memory

$s_{e(adm)}^2$ = error variance due to administration effects

$s_{e(g)}^2$ = error variance due to guessing

$s_{e(subj)}^2$ = error variance due to lack of agreement between scorers or raters

$s_{T(fl)}^2$ = variance due to fluctuation in true scores from one testing to the next (Magnusson, 1967, p. 106).

If the same person administers the test on both occasions systematic error rather than error variance would occur and the term $s_{e(adm)}^2$ would not be included in the total error component when the reliability is computed.

The test-retest method is advantageous when time and funds do not allow the construction of a second test to be used as a parallel measure. It may also be used when there is reason to believe that memory will not have a significant effect in making the two scores similar. It is especially appropriate to use this method when sampling of content is not the issue but rather the repeatability of scores, such as a speed test (Wessman, 1968; Nunnally, 1970).

It is interesting to note that the fluctuation in true scores increases while the variance due to memory decreases as the length of time between testing increases. Since both of these are undesirable measurement error, the length of time between testing should be arranged to minimize both (Magnusson, 1967).

Parallel Forms

In this method two tests are constructed with an attempt made to satisfy the conditions of parallelism as defined by the person constructing the tests. The correlation of these two tests is the estimate of the reliability coefficient. It is difficult, if not impossible to construct truly parallel tests, so often the parallel tests that are used measure slightly different true scores. The two tests can have high correlations which suggest that on the whole they measure the same true score but each test will measure a true component which is not measured by the other. This true component which is measured by one test but not by the other is treated as measurement error and thus does not contribute to the reliability coefficient. The total test variance for one of the two tests may be written as:

$$s^2_t = s^2_T + s^2_{e(m)} + s^2_{T(equ)} + s^2_{e(adm)} + s^2_{e(g)} + s^2_{e(subj)} + s^2_{T(fl)} \qquad (24)$$

where,

$s^2_{T(equ)}$ = the true component measured by one of the parallel tests but not by the other and all other terms are previously defined under the test-retest method (Magnusson, 1967, p. 107).

As in the test-retest method, if the same person administers the test, the term $s^2_{e(adm)}$ would not be included in the computation of the reliability coefficient.

The parallel test method is almost always preferable to the test-retest method because memory can have only a slight effect and it will give information on all the sources of error found in the test-retest method plus an indication of the amount of error due to content sampling (Nunnally, 1970). The time interval between testing is a problem with parallel forms as it was for the test-retest method. The first test may have some effect on the scores of the second test which results in measurement error.

Subdivided Test Method

If the time interval between administrations of two parallel tests is undesirable, the test can be reconstructed by alternating items from both forms. The odd items on this test would be considered test 1 and the even items test 2. The number of rights on each test would be calculated for a total score. The correlation of these two scores would yield a reliability coefficient which estimates the reliability of both tests (Magnusson, 1967).

The same principle operates in what is commonly called the split-half method of estimating reliability. Instead of constructing two parallel forms one test is divided in half (Magnusson, 1967; Nunnally, 1970). According to Magnusson (1967) it is best to score the test first and then place the items in a score matrix in order of frequency of correct response. One parallel test is then made up of even numbered items and one of odd numbered items insuring the two tests will be equal in difficulty and equally differentiating by having equal means and variances. This procedure insures that the two subtests measure as much of the same true score as possible. Total test variance for the subdivided test method may be expressed as (Magnusson, 1967, p. 112):

$$s_t^2 = s_T^2 + s_{T(equ)}^2 + s_{e(adm)}^2 + s_{e(g)}^2 + s_{e(subj)}^2 \qquad (25)$$

where,

all terms are as defined previously.

The two subtests would be correlated to achieve a reliability coefficient but a statistical correction must be made to estimate the reliability of the whole test which is known as the Spearman-Brown Prophecy Formula.

The subdivided test method determines some error due to the sampling of content but not as well as the parallel form method. This method does show error due to instability over time and for this reason is usually an overestimate of reliability (Nunnally, 1970).

Internal Consistency Methods

These methods are concerned with the estimation of the homogeneity within one test. Equations concerning homogeneity or the internal consistency of a test estimate the correlation between an existing test and a hypothetical equivalent form. According to Nunnally (1970) this requires two assumptions:

> 1. The average correlation between items within the existing test would be the same as the correlation between items in the hypothetical equivalent form.
>
> 2. The average correlation between items in the two forms would be the same as the average correlation within the existing form (p. 550).

Three existing formulas for estimating internal-consistency reliability will be discussed; the KR-20 (Kuder and Richardson, 1937), Coefficient Alpha (Cronbach, 1951), and the Analysis of Variance Model (Hoyt, 1941; Cronbach, 1963; Kerlinger, 1973).

The KR-20 formula is the most familiar method of estimating internal consistency reliability for dichotomous items, i. e. items scored either zero or one. The KR-20 alleviates the problem of subdividing a test since it gives the mean of all possible split-half correlations between subtests (Kuder and Richardson, 1937). When computing the KR-20, the true variance is determined by the size of the covariance terms for a given number of items. The size of the covariance terms in turn is determined by the intercorrelations and standard deviations of the items (Magnusson, 1967). The internal consistency coefficient obtained from KR-20 is therefore directly dependent on the correlations between the items in the test, or on the extent to which the items measure the same variable. The more homogenous the items are, the greater the numerical value of KR-20 (Nunnally, 1973). Kuder and Richardson (1937) also developed a formula for the estimation of the reliability of a test with weighted item scores which is called KR-21 and yields the same reliability coefficient as Coefficient Alpha (Cronbach, 1951).

Cronbach (1951) developed Coefficient Alpha which is a more general form for the estimation of the reliability of a test with weighted items. He considers KR-21 to be a special case of his more general formula.

The Analysis of Variance model may be used to establish an intraclass correlation coerficient when two tests are available (Cronbach, 1963) or when only one test or rating is available an internal consistency reliability coefficient. The information from a sum of squares table is inserted in an estimation formula to arrive at the reliability coefficient.

The KR-20 and Analysis of Variance method result in the same numerical value for the reliability coefficient but the Analysis of Variance method is considered to be statistically more powerful and provides more information regarding sources of measurement error (Cronbach, 1963; Kerlinger, 1973). In summary, internal consistency estimates provide an indication of the tests homogeneity and are useful when only one test per subject is available.

The Film Test for Understanding Behavior

History and Development

The Film Test for Understanding Behavior (FUB) was created by Schalock and Edling (1958) as a method of measuring responses to children's behavior ". . . which attempts to incorporate some of the emotional involvement that is encountered in an interpersonal situation, yet maintains sufficient simplicity to make its administration feasible." It allows one to measure behavioral understanding of young children in a classroom situation rather than in a personal encounter with young children.

Ten film episodes of children s behavior in the preschool were selected from footage taken at Orchard Street Child Development Laboratory, Oregon State University, Corvallis, Oregon. The ten episodes, each approximately one minute in length, were selected as representative occurrences throughout the day in a demonstration laboratory school. The subjects of each episode included in the film test are (1) a child sitting and watching what is occurring around him in the nursery school, (2) a child playing with paint, (3) a child taking part in rhythm activities, (4) a child dressing, (5) a child painting leaves outside, (6) a child eating, (7) a situation in which two children confiscate the property of another child with its consequence, (8) a motor sequence, (9) a sequence involving aggression, and (1) an episode enabling comparative judgements of mental abilities.

After the film episodes were selected an item pool was generated for each episode. The items were concerned with knowledge of child development and behavior, knowledge of guidance principles and awareness of how the child was feeling in the situation being filmed.

Once the item pool was completed, the items were presented to people outside the field of child development and psychology for evalaution of their clarity and readability. After this process was completed 130 items were left in the item pool. Each item was then assigned a response weight by a group of five persons holding advanced degrees in child development, preschool teaching, or psychology. Each person was asked to respond individually to each item. He was then asked to justify his response in writing, listing in detail the basis for his answer. If he was responding to a behavioral cue, he was asked to identify the behavior and outline the assumptions made in interpreting the behavior.

The group of five then met to compare their individual responses to the items. After viewing the film episodes and consulting child development literature, a consensus was achieved on the ordering of the items. The items are weighted on a five point scale from +2 to -2. The five possible responses to each item are ordered according to degree of correctness. The most correct response scores +2, the second most +1, the third 0, the fourth -1, and the least correct response scores -2.

The FUB was then administered to three groups of students with varying levels of knowledge in child development and psychology. Group I had completed zero courses in child development or psychology, Group II had completed two terms in psychology and one in child development, while Group III had completed two terms in psychology and one term in child development plus participation in the preschool. Each group consisted of 20 students.

The Likert Method of item analysis was used to find which items discriminated most between high and low scores on the test. The 36 items which discriminated Group I from Group II and Group III were included in a low-medium key which is used in scoring FUB tests taken by individuals having little or no background in child development or psychology. The high scoring key also contains 36 items which were found to discriminate Group III from Group I and II. Fourteen items discriminate between all three groups and are found on both scoring keys.

Three subscales were developed for the FUB (O'Neill, 1960) by asking authorities in the field to separate items as they related to the areas of normal child development and behavior, knowledge of guidance principles, and awareness of the child's feelings in the situation being filmed. This resulted in the Knowledge Scale (score range +14 to -14), Guidance Scale (score range +18 to -18) and the

Sensitivity Scale (score range +11 to -11).

Administration of the Test

The FUB may be administered individually or in a group setting. Materials required for administration include a 16 mm film projector, movie screen, film test, test booklet and score sheet. The test booklet contains an introduction to the test, directions, and background information for each episode, as well as the test items for each episode.

The person administering the test must be trained in the operation of the projector and should stand behind the projector which needs to be placed near the back of the room. This procedure was introduced after a study by Owen (1968) showed that facial expression and other non-verbal reactions of the administrator influenced the subject's response to the film episodes.

The projector is stopped after each episode to allow the subject to respond to the items regarding the episode. There is no time limit for the test although the total testing time is usually 40 minutes.

Standardization

The FUB was standardized (Family Life Department, 1974) on a predominantly female sample of 1500 college students from Oregon State University who were enrolled in a sequence of courses in child development or preschool teaching. Each of these courses were assigned a training level in the following table which also provides a description of the sample.

Table 1. Description of the College Student Training Sequence Used in the Standardization of the FUB.

Course sequence number	Testing time	Training level (TL)	Class level	N
1	Pretest	TL1	Freshman	113
2	Pretest	TL2	Freshman	723
3	Pretest	TL3	Soph/Junior	260
4	Pretest	TL4	Junior/Senior	271
5	Posttest	TL5	Junior/Senior	201
			TOTAL	1568

(Family Life Department, 1974)

The standard scores are computed for each training level using the following formula:

$$\text{standard score} = [(15/SD_r)\,(x-\bar{x})] + 100 \qquad (26)$$

where,

SD_r = standard devaition of raw score distribution for the training level

$\bar{x}$ = raw score mean of the training level

x = raw score

Validity

Content validity is the representativeness or sampling adequacy of the content for a particular subject (Kerlinger, 1973). Every psychological property has a theoretical universe which contains all that is possible to be said or observed regarding a property. Theoretically, if a test is representative of this content universe then the test is said to have content validity. This universe exists in theory only. Thus, the assessment of a measuring instrument's content validity must be done by judges with knowledge of this content universe and the ability to judge the degree to which a test covers all the possible areas of content. The method used in developing items for the FUB and the selection of film episodes to be included in the test establish the content validity. When selecting items care was given to choose all types of items related to knowledge of child behavior and guidance principles. When choosing film episodes an attempt was made to get examples of common occurrences in the daily routine of a laboratory preschool. Thus, in the opinion of the judges, each of whom held an advanced degree in a related field, the FUB has a high degree of content validity.

The predictive validity of the FUB may be seen by examining the increases in the mean standard score for each training level shown in Table 2. The college student standardization sample showed consistent increases in the mean score at each training level with significant increases occurring at Training Level (TL) 2, TL 3, TL 5. The Knowledge Subscale showed significant increases at all training while

Table 2. Comparison of FUB Subscale Means for the College Student Training Sequence.

		TOTAL				KNOWLEDGE		
TL	Mean	Mean diff	S.E.M.	t	Mean	Mean diff	S.E.M.	t
1	8.92	--	--	--	1.32	--	--	--
2	15.64	6.72	1.44	4.66**	2.76	1.44	0.39	3.69**
3	20.74	5.10	1.00	5.10**	3.87	1.11	0.26	4.27**
4	22.51	1.77	1.05	1.69	4.95	1.08	0.30	3.60**
5	23.56	1.05	.59	1.78*	5.47	0.52	0.26	2.00**
		GUIDANCE				SENSITIVITY		
1	4.47	--	--	--	3.11	--	--	--
2	8.31	3.84	0.96	4.00**	4.57	1.46	0.48	3.04**
3	10.41	2.10	0.67	3.13**	6.45	1.88	0.34	5.53**
4	11.58	1.17	0.67	1.75	5.97	-0.48	0.37	-1.30
5	11.56	-0.02	0.87	-0.02	6.52	0.55	0.25	2.20**

**Significant .01 Level

*Significant .05 Level

(Family Life Department, 1974)

the Guidance subscale increased from TL 1 to TL 4 but had an insignificant decrease at TL 5. The Sensitivity scale is less predictive with significant increases between TL 1-TL 3, a decrease at TL 4 and a significant increase at TL 5. Data supporting the predictive validity of the FUB comes from scores of teachers of preschool children and parents of young children. Table 3 shows a comparison of FUB scores for teachers and college students. Teachers had a significantly higher mean FUB score for the Total, Guidance and Sensitivity scales but had significantly lower scores on the Knowledge subscale.

Table 3. Comparison of Professionals and College Students' (TL5) FUB Scores.

	Professional mean	Student mean	Mean diff.	S. E. D.	t	p
Total	31. 06	23. 56	7. 50	2. 06	3. 64**	. 001
Know.	4. 15	5. 47	1. 32	0. 37	3. 57**	. 001
Guid.	17. 06	11. 56	5. 50	1. 41	3. 90**	. 001
Sens.	9. 84	6. 52	3 32	0. 82	4. 05**	. 001

(Family Life Department, 1974)

Construct Validity: In developing the subscales for the FUB, three theoretical dimensions of teacher competence were delineated:

1. Knowledge of normal child behavior and development
2. Understanding of guidance principles in specific applications
3. Sensitivity to the thoughts and feelings of young children.

It was postulated that these competencies were related to teacher effectiveness and were acquired through training and experience with young children.

Reliability

As mentioned previously reliability studies on the FUB have been completed using the test-retest method. Sugawara (1972) conducted one such study with a sample of 48 college students in TL 1. The students were retested after four or seven weeks. The four week sample contained 16 females and five males while the seven week sample contained 20 females and five males. The results of this analysis are shown in the following table.

Table 4. Test-Retest Reliability Coefficients for the FUB.

	Combined (N=48)	One month (N=21)	7 weeks (N=27)
Total	.63	.64	.57
Know.	.75	.49	.81
Guid.	.60	.57	.63
Sens.	.74	.61	.75

(Family Life Department, 1974)

Other supportive test-retest studies have been done as part of other research using the FUB. Karuven (1960) established a reliability coefficient of . 78 with a sample of seven college students. The time interval between testings varied from seven days to five months with an average of two months.

The other evidence of reliability comes from a reliability coefficient of . 73 established by O Neill (1963). The sample consisted of nine college students who were retested after a one month time interval.

METHOD

This study establishes an internal consistency reliability coefficient for the Film Test for Understanding Behavior (FUB) using results of an Analysis of Variance. A description of the subjects who provided the FUB scores and the procedure by which the data were collected will be discussed in this section as well as a description of the analysis used to arrive at the reliability coefficient.

Subjects

Data for this reliability estimate were collected from a total of 321 students who were enrolled in a beginning child development course, FL 225, during the academic years of 1971-72 and 1972-73. This total is a combination of all students tested during seven different terms. Table 5 provides a breakdown of the subjects by year and term.

Table 5. A Description of FL 225 Students by Year and Term.

Year	Term	n
1971	Fall	81
1972	Winter	38
1972	Spring	37
1972	Fall	40
1973	Winter	37
1973	Spring	65
1973	Summer	30
	TOTAL	328*

*Seven did not complete the Film Test and were not included in the analysis.

In addition to responding to the Film Test the subjects were asked to report on their status with respect to selected background characteristics. These self-reports provide the basis for the descriptions and classifications in Table 6. Of the total 321 subjects, 95% were female and 99% were undergraduates. The school affiliation and major area are also reported in Table 6.

Students from the School of Home Economics constitute a clear majority of the subjects. At approximately 61%, they are four times larger in number than the next highest contriubotry (15%) the School of Liberal Arts. These two are followed by the School of Education at 10%. The representation from the other five schools represents but 8% of the total sample. The remaining students, approximately 6% reported that they were undecided on a major at the time of testing or they did not respond to the question at all. Clearly, then, the subjects were primarily female undergraduate students and a majority of these were in the school of Home Economics.

One additional aspect of the subjects' background is of interest to this study and that deals with the number of related courses in academic subject matter areas which also deal in some way with human behavior. Knowledge of this type is helpful in assessing the potential astuteness of the subjects in knowledge of human behavior. Table 7 provides information about the number of courses in related fields the subjects had taken when they were administered the Film Test.

Table 6. Distribution of 1971-72 and 1972-73 FL 225 Students by School and Academic Major.

School, Major	f	%	Total %
AGRICULTURE	(2)		0.6
Ag. General	1	0.3	
Ag. Economics	1	0.3	
BUSINESS	(11)		3.4
Bus. Administration	11	3.4	
EDUCATION	(32)		10.0
Elementary	17	5.3	
General	12	3.7	
Physical therapy	2	0.6	
Physical Ed. and Health	1	0.3	
FORESTRY	(1)		0.3
General	1	0.3	
HOME ECONOMICS	(195)		60.7
General	99	30.8	
Family Life	27	8.4	
Foods and Nutrition	18	5.6	
Clothing and Textiles	19	5.9	
Home Ec. Education	32	10.0	
LIBERAL ARTS	(48)		15.0
Art	1	0.3	
English	1	0.3	
History	2	0.6	
Journalism	2	0.6	
Liberal Studies	20	6.2	
Political Science	1	0.3	
Psychology	6	1.9	
Sociology	11	3.4	
Speech Communications	4	1.2	
PHARMACY	(2)		0.3
General	2	0.3	
SCIENCE	(12)		3.7
Nursing	7	2.3	
Physiology	1	0.3	
Pre-Dentistry	1	0.3	
Pre-Medicine	3	0.9	
UNDECIDED	(8)		2.5
NOT RETURNED	(10)		3.1
TOTAL	321		

Table 7. Frequency Distribution of FL 225 Students by Number of Courses in Related Areas.

Related areas	Number of Courses									
	0	1	2	3	4	5	6	7	8	9
Family Life	208	94	17	1	1	0	0	0	0	0
Sociology	157	47	62	46	5	1	0	0	0	3
Anthropology	238	52	14	10	4	0	1	1	0	1
Education	245	28	22	8	7	1	3	1	0	6
Psychology	156	117	23	19	4	1	1	0	0	0

Procedure

The Film Test for Understanding Behavior (FUB) was administered to all students enrolled in FL 225 during each of the seven terms of the 1971-72 and 1972-73 academic years. The test was administered in a group setting within the first two weeks of the respective terms. Additional demographic data needed to describe pertinent background characteristics of the subjects were collected at the same time by means of a single sheet handed out with the FUB (Appendix B).

This three credit course is considered a basic core course for the School of Home Economics and is required for all majors. It has only one section per term and meets regularly for one hour on a MWF schedule. Because of the University registration procedures classes on the MWF schedule meet only on Wednesday and Friday during the first week of classes. In FL 225 it is necessary to use these class periods for orientation to the course and for visits to the Child Development Laboratory which is the site of the required observation of preschool children. The testing was therefore scheduled on the earliest convenient day during the second week of classes and always before lecturing on the human growth and development content of the course had begun. All data collection was completed with ease during one 50 minute class period.

All subjects were informed that participation in this research was strictly on a volunteer basis and their choice to participate or not would have no effect on their eventual grade for the course. This volunteer participation represents a long standing departmental policy as well as compliance with the more recently established policy of the University regarding the protection of human subjects.

Analysis

Previously reliability was defined as the ratio of true score variance to total or individual variance. It was also stated that the value of the true score variance could not be calculated directly but the value of the error variance could be. Calculation of the error variance would permit the estimation of true score variance because the following relationship exists between error and true scores:

$$X_t = X_\infty + X_e \qquad (1)$$

where,

X_t = Total score

X_∞ = True score component

X_e = Error score component

This equation may be extended algebraically to:

$$V_t = V_\infty + V_e \tag{2}$$

where,

V_t = Total test or individual variance

V_∞ = True score variance

V_e = Error score variance

By calculating the error variance and subtracting it from the total variance, the true variance may be estimated:

$$V_\infty = V_t - V_e \tag{27}$$

The Analysis of Variance provides the total (individual) variance in the form of the Individual Mean Square and the error variance as the Residual Mean Square. In Analysis of Variance terms reliability may be stated as:

$$r_{11} = \frac{MS_{Ind} - MS_r}{MS_{Ind}} \tag{28}$$

where,

r_{11} = Total test reliability

MS_{Ind} = Individual Mean Square

MS_r = Residual Mean Square

Homogeneity of the Sample and Reliability

Other things being equal, the more heterogenous the group, the higher the reliability. The reason for this can be explained by examining this definition of reliability:

$$r_{11} = 1 - s_e^2/s_t^2 \tag{22}$$

where,

r_{11} = Total test reliability

s_e^2 = Error variance

s_t^2 = Total test variance.

S_e^2 is the variance of a person's observed score about his true score, there is no reason to expect the observed score to vary as a result of group characteristics. But s_t^2 increases with heterogeneity. If s_e^2 remains constant and s_t^2 increases, r_{11} increases.

The sample of subjects in this study is very homogenous. Thus, the reliability coefficient should be evaluated as a conservative estimate (Mehrens and Lehmann, 1973).

RESULTS AND CONCLUSIONS

The purpose of this study was to establish an internal consistency coefficient of reliability for the Film Test for Understanding Behavior (FUB) from data provided by an Analysis of Variance.

The subjects for the study were 321 college students enrolled in a basic course in child development offered by the Family Life Department of Oregon State University. The testing took place during Fall, Winter, and Spring terms of the 1971-1972 school year and Fall, Winter, Spring, and Summer terms of the 1972-1973 school year as part of an on-going research project of the Family Life Department.

Table 8 presents the results of the Analysis of Variance of Film Test scores and the calculation of the reliability coefficient for the total test which was estimated at .77. Designed as a measure of behavioral understanding, this instrument has demonstrated a moderately high level of reliability.

The reliability of the Knowledge Subscale is calculated from the data presented in Table 9. This subscale contains seven items and was designed to measure a subject's knowledge of child development and normal child behavior. The reliability of this subscale was estimated at .33 with the small number of items contributing to the low reliability coefficient.

Table 8. Analysis of Variance on Scores from the Film Test for FL 225 Students: Low-Medium Key.

Source	df	s. s.	m. s.	F
Items	35	3994. 29	114. 12	73. 6258*
Individuals	320	2114. 29	6. 61	4. 2645*
Residual	11200	17354. 27	1. 55	
TOTAL	11555	23462. 85		

*significant at . 01

Reliability Coefficients Estimated from Variances
Calculated by Analysis of Variance

$r_{11} = 1 - V_e/V_t$

$r_{11} = 1 - 1.55/6.61 = .7655$

rounded = . 77

where,

r_{11} = Total test reliability

V_e = Error variance

V_t = Total test variance

For perfect reliability $r_{11} = 1$, that is, the error variance is zero.

The Standard Error of Measurement of the reliability coefficient is 1. 245.

Table 9. Analysis of Variance for Film Test Scores for FL 225 Students; Knowledge Subscale.

Source	df	s. s.	m. s.	F
Items	6	1048. 47	174. 68	110. 556*
Individuals	320	757. 41	2. 37	1. 5*
Residual	1920	3035. 64	1. 58	
TOTAL	2246	4841. 12		

*significant at .01

Reliability Coefficient Estimated from Variances
Calculated by Analysis of Variance

$r_{11} = 1 - V_e/V_t$ $r_{11} = 1 - 1.58/2.37 = .3320$

rounded = .33

where,

r_{11} = Total subscale reliability

V_e = Error variance

V_t = Total subscale variance

For perfect reliability $r_{11} = 1$, that is, the error variance is zero.

The Standard Error of Measurement of the reliability coefficient is 1. 26.

Table 10 shows the calculation of the reliability coefficient for the Guidance subscale of the film test. With 18 items the Guidance subscale contains almost twice as many items as the other two subscales which may partly account for the .73 reliability coefficient since increasing the number of items in a test generally increases its reliability (Kerlinger, 1973).

The Guidance subscale includes items which measure an individual's knowledge of guidance principles.

The Sensitivity Subscale was designed to measure the subject's empathy with and awareness of the child's feeling in the situation being filmed. The reliability of this subscale was estimated at .27 and was calculated from data presented in Table 11. This subscale has low reliability but as with the Knowledge Subscale the small number of items in the Sensitivity Subscale (n=11) may account for this low value.

Table 10. Analysis of Variance for Film Test Scores for FL 225 Students: Guidance Subscale.

Source	df	s. s.	m. s.	F
Items	17	2171. 62	127. 74	86. 89*
Individuals	320	1740. 04	5. 44	3. 70*
Residual	5440	8002. 82	1. 47	
TOTAL	5777	11914. 49		

*significant at . 01

Reliability Coefficient Estimated from Variances
Calculated by Analysis of Variance

$r_{11} = 1 - V_e/V_t$ $r_{11} = 1 - 1.47/5.44 = .7295$

rounded = . 73

where,

r_{11} = Total subscale reliability

V_e = Error variance

V_t = Total subscale variance

For perfect reliability $r_{11} = 1$, that is, the error variance is zero.

The Standard Error of Measurement of the reliability coefficient is 1. 21.

Table 11. Analysis of Variance for Film Test Scores for FL 225 Students: Sensitivity Subscale.

Source	df	s. s.	m. s.	F
Items	10	755. 96	75. 60	46. 380*
Individuals	320	713. 15	2. 23	1. 368*
Residuals	3200	5219. 50	1. 63	
TOTAL	3530	6688. 60		

*significant at . 01

Reliability Coefficient Estimated from Variances
Calculated by Analysis of Variance

$r_{11} = 1 - V_e / V_t$

$r_{11} = 1 - 1.63/2.23 = .2681$

rounded = . 27

where,

r_{11} = Total subscale reliability

V_e = Error variance

V_t = Total subscale variance

For perfect reliability $r_{11} = 1$, that is the error variance equals zero.

The Standard Error of Measurement of the reliability coefficient is 1. 27.

Table 12. A Summary of Reliability Coefficients with Corresponding Standard Errors for the Film Test for Understanding Behavior and its Subscales.

Scale	r_{11}	$SE_{meas.}$
Total	.77	1.25
Knowledge	.33	1.26
Guidance	.73	1.21
Sensitivity	.27	1.27

Suggestions for Further Study

The present reliability coefficients established for the Film Test are most accurate in estimating the reliability of the FUB when it is used in testing freshman and sophomore females with little or no experience or training with children. These individuals are scored on the low-medium key. Past research has shown that the high-medium key should be used to score tests taken by students at Training Level 3 and above. It would be useful to establish a reliability coefficient for the FUB at Training Level 3 where the student begins a practicum experience with children. Eventually, it would be advisable to establish a reliability coefficient for each training level so that scores from students at each training level could be assessed more precisely.

REFERENCES

Aiken, L. R., Jr. Psychological and educational testing. Boston: Allyn and Bacon. 1973.

Brown, W. Some experimental results in the correlation of mental abilities. British Journal of Psychology, 1910, 3, 296-322.

Brownell, W. A. On the accuracy with which reliability may be measured by correlating test halves. Journal of Experimental Education, 1933, , 204-215.

Burrows, A. L. Relationship of self-concept to behavioral understanding in prospective preschool teachers. Unpublished master's thesis, Oregon State University, 1972.

Burt, C. Test reliability estimated by analysis of variance. British Journal of Statistical Psychology, 1955, 8, 103-118.

Coombs, C. H. The concepts of reliability and homogeneity. Educational and Pyschological Measurement, 1950, 10, 43-56.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L J.; Rajaratnam, N.; Glesser, G. C. Theory of generalizability: a liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 15, 137-163.

Cronbach, L. J. Essentials of psychological testing (3rd ed.). New York: Harper, 1970.

Cureton, E. The definition and estimation of test reliability. Educational and Pyschological Measurement, 1958, 18, 715-738.

Family Life Department. Preliminary Analysis: Film Test for Understanding Behavior. Unpublished Manuscript, Oregon State University, 1974.

Ghiselli, E. E. Theory of psychological measurement. New York: McGraw-Hill, 1964.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Gulliksen, H. Theory of mental tests. New York: John Wiley and Sons, 1950.

Guttman, L. A basis for analyzing test-retest reliability. Psychometrika, 1945, 10, 255-282.

Harrison, R. L. Assessment of a selected group of high school girls experience in a child observation center. Unpublished master's thesis, Oregon State University, 1970.

Hollingshead, A. B. Two factor index of social position. Unpublished manuscript, 1957.

Horst, P. A. A generalized expression for the reliability of measures. Psychometrika, 1949, 14, 21-31.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Isaac, S. and Michael W. W. Handbook in research and evaluation. USA: Robert Knapp, 1971.

Jackson, R. W. B. Reliability of mental tests. British Journal of Psychology, 1939, 29, 267-287.

Karuven, M. M. The effect of coursework in child development and psychology on understanding the behavior of preschool children. Unpublished master's thesis, Oregon State University, 1960.

Kerlinger, F. N. Foundations of behavioral research (2nd ed.). New York: Holt, Rinehart and Winston, 1973.

Kuder, G. F., and Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

LaForge, R. Components of reliability. Psychometrika, 1965, 30, 187-195.

Lyerly, S. B. The Kuder-Richardson formula 21 as a split-half coefficient, and some remarks on its basic assumptions. Psychometrika, 1958, 23, 267-270.

Magnusson, D. Test Theory. London: Addison Wesley, 1966.

Mehrens, W. A. and Lehmann, I. J. Measurement and evaluation in education and psychology New York: Holt, Rinehart and Winston, 1973.

Mosier, C. I. A note on item analysis and the criterion of internal consistency. Psychometrika, 1936, 1, 275-282.

Nunnally, J C. Educational measurement and evaluation. New York: McGraw-Hill, 1967.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Nunnally, J. C. Introduction to psychological measurement. New York: McGraw-Hill, 1970.

O'Neill, J. P. Development of Subscales for the Film Test for Understanding Behavior. Unpublished Manuscript, Family Life Department, Oregon State University, 1960.

O'Neill, J. P. Test-Retest Analysis of the Film Test for Understanding Behavior. Unpublished Manuscript, University of Illinois, 1963.

Owen, S. L. The effects of different instructions and instructional levels on the scores of a film test for understanding children's behavior. Unpublished master's thesis, University of Illinois, 1968.

Rulon, P. J. A simplified procedure for determining the reliability of a test by split-halves. Harvard Educational Review, 1939, 9, 99-103.

Schalock, H. D. and Edling, J. The film test for understanding behavior. Unpublished manuscript, Oregon State University, 1958.

Smith, M. M. Correlates of college students' understanding of children's behavior. Unpublished master's thesis, Oregon State University, 1960.

Spearman, C. The proof and measurement of association between two things. American Journal of Psychology, 1904, 15, 72-101.

Spearman, C. Correlations calculated from faulty data. British Journal of Psychology, 1910, 3, 271-295.

Sugawara, A. Unpublished Manuscript, Family Life Department Oregon State University, 1972.

Tyler, L. E. Tests and measurements (2nd ed.). New Jersey: Prentice-Hall, 1971.

APPENDICES

APPENDIX A

OREGON STATE UNIVERSITY

FAMILY LIFE DEPARTMENT

THE FILM TEST FOR UNDERSTANDING BEHAVIOR

(FORM II)

INSTRUCTIONS

The statements in this booklet are statements about the episodes of behavior you will observe in the film. Some of these statements have to do with how a child feels; some with ways of handling what is happening; and some with general principles of development and behavior.

After observing an episode of behavior, you are asked to respond to the items pertaining to that episode. Generally speaking, you are to indicate whether you agree with an item, you disagree with it, or whether you are uncertain as to your agreement or disagreement about it. Specifically, your response to each item will be made in terms of one of five categories.

A	Ah	U	DH	D
Agree	Agree, but with some hesitation	Uncertain, due to insufficient evidence in the film to judge or due to insufficient knowledge in the field to judge.	Disagree, but with some hesitation	Disagree

Thus, if you clearly agree with a statement, you select "A" for your response. If you generally agree with a statement, but realize that it is likely that there will be exceptions to it, you select Ah for your response. The reverse is true for indicating disagreement. If there is insufficient evidence presented in the film for making an agreement-disagreement decision, or if you feel that the knowledge available in the field of human development and behavior is insufficient to permit an agreement-disagreement decision, you select "U" for your response.

In all cases your response to an item is to depend only on what you see in the film, coupled with what you know generally about the behavior and development of children. Insufficient knowledge of a particular child and his background should not be considered a basis for your response to any item.

You are to indicate your response to each item by blackening the appropriate space on the accompanying answer sheet. Please do not write on this booklet.

DO NOT TURN THE PAGE AND READ THE STATEMENTS ABOUT THE EPISODE BEFORE OBSERVING THE EPISODE. Be sure to read, however, "Information Needed in Observing Episode 1, 2, etc.", which appear on the page preceding the statements which go with a particular episode.

Information Needed in Observing Episode 1.

The boy sitting, facing the camera is the subject of episode 1. He has just reached his third birthday, and this is his second day in nursery school.

EPISODE 1

1. If an adult has helped the child take part in the activities around him, rather than just letting him sit and watch, the child would have adjusted to the situation more quickly.

2. An adult should have suggested that the child move to a place where he was less distracted.

3. Although the child was interested in the activity around him, he really wasn't ready to take a more active part in it.

4. One of the things this child will gain from going to nursery school is more confidence in himself when he enters a new situation.

5. Within a week or so it is likely that the child will play freely with other children.

6. One would judge this child's adjustment to be more adequate had he entered the situation with less hesitation.

7. It is likely that this child hadn't played with many children before entering nursery school.

8. An adult at least should have talked to the child or asked him if there was anything they could do to help.

9. Throughout the elementary school years, this child is apt to sit back and watch for a while whenever he enters a new situation.

10. It is likely that the child will not be a leader in school.

Information Needed in Observing Episode 2.

The girl using the paints is the subject of episode 2. She is nearing her fourth birthday, and has been in nursery school for nearly a year.

EPISODE 2

16. An adult should have shown the child how to use the paint more constructively.

17. If this child is allowed to continue to be messy with paint and other things at nursery school, she will want to be messy at home.

18. Using paint in this way has little value as an art experience.

19. The child probably was seeing how messy she could be with the paints before an adult stopped her.

20. It is likely that this child isn't allowed to be messy at home.

21. The child seemed to be more concerned about getting paint on her clothes than she was with getting it on her hands and arms.

22. If this child is allowed to be messy at nursery school, but not at home, she soon will not be sure where she can be messy and where she can't.

Information Needed in Observing Episode 3.

The boy that the camera opens on is the subject for episode three. He is just past three years of age, and he is in only his second week at nursery school.

EPISODE 3

31. The child seemed to feel guilty about not doing as the others were doing.

32. The child seemed to be a well-adjusted child.

33. The child probably was less interested in rhythms than he was in what the children on the ground were doing.

34. An adult should have helped the child stand on the board.

35. An adult should have helped the child do something else.

Information Needed in Observing Episode 4.

The girl putting on her trousers is the subject for episode 4. She is four and a half years old, and has been in nursery school about six months.

EPISODE 4

46. This is a good example of an adult helping a child when the child really didn't need help.

47. The child was becoming upset over not being able to get her trousers on by herself.

48. The adult should have used this situation to point out to the child how to get into her trousers by herself rather than helping her.

49. The next time the child has a problem in dressing she is apt to want help from an adult.

50. The child probably would have become upset if the adult had not helped her when she did.

Information Needed in Observing Episode 5.

The child you will see in the film is nearing four years of age. He has been in nursery school for nearly a year.

EPISODE 5

61. This would have been a more valuable experience for the child had he made a good print of the leaf.

62. An adult should have shown the child how to be less messy in his painting.

63. Apparently, the child didn't care that his picture was a messy one.

64. An adult should have helped the child make a better print.

65. The child shouldn't have been left by himself to do such a complicated task.

Information Needed in Observing Episode 6.

The child you will see in the film is three and a half years old and has been in nursery school for about six months.

EPISODE 6

76. The child should not have been allowed to eat with his fingers.

77. It seemed to be easier for the child to eat with his fingers than with his fork.

78. The adult should be sure that the child finishes the food on his plate before he leaves the table.

79. This child has to learn that mealtime is a time for eating rather than a time for playing or just looking around.

80. Most children of this age would not let their attention wander from their eating as much as this child did.

81. When allowed to eat like this at nursery school the child is likely to eat in much the same way at home.

82. The child seemed to resent the adult telling him what to do.

Information Needed in Observing Episode 7.

The girl putting leaves in the wagon is the subject of Episode 7. She is nearly four years of age, and has been in nursery school for nearly a year.

EPISODE 7

91. Apparently, the girl is a friendly, sociable child.

92. An adult should have helped the girl keep the boys from taking the leaves.

93. Most children of this age would not have felt so strongly about losing some leaves as this girl did.

94. The boys who took the leaves from the wagon should have been reprimanded.

95. It is likely that these boys are trouble makers in the nursery school.

96. Someone should help the girl realize that she should not cry over something as unimportant as this.

Information Needed in Observing Episode 8.

The larger boy on the bars is the subject of Episode 8. He is three and a half years old and has been in nursery school for about six months.

EPISODE 8

106. It is likely that the child is well adjusted since he is so free and confident in his body movements.

107. As an adolescent, it is likely that the child will excell in athletics.

108. An adult should have been near the child when he was playing on the bars.

109. This child probably wouldn't be interested in such things as painting or listening to music.

110. It is likely that this child is a bully.

Information Needed in Observing Episode 9.

Two girls are the subjects in this episode. The girl the camera opens on will be called the "1st" girl. The girl who enters the episode later will be called the "2nd" girl. Both are four and a half years old, and have been in nursery school for about six months.

EPISODE 9

121. Leaving the girls to settle their differences by themselves was a good idea.

122. The "lst" girl is likely to be assertive throughout childhood.

123. The "lst" girl probably is an insecure child.

124. An adult should have comforted the "2nd" girl.

125. The "lst" girl should have been reprimanded for taking the toy away from the "2nd" girl.

126. The "lst" girl is a selfish child.

127. After the "lst" girl took the cup away, an adult should have helped the "2nd" girl get started in another activity.

128. The "2nd" girl probably is an insecure child.

Information Needed in Observing Episode 10.

A girl and a boy are the subjects in Episode 10. Both have just passed their third birthday, and have been in nursery school about a month.

136. The girl seemed to be upset by not being able to work the puzzle.

137. It is likely that the difference in the ability of the two children to work puzzles is due to something other than intelligence or the opportunity to practice.

138. An adult should have helped the girl work the puzzle.

139. Even though the girl didn't work the puzzle well, she should have been praised for her effort.

140. An adult should have given the boy a puzzle that was harder for him to work.

APPENDIX B

Background Information

Please fill out the information below. It will be held in strictest confidence and in no way will it affect your class grade. Please read and follow the directions carefully. If you have any questions, please ask them. Thank you.

A: Name ________________ Age ______ Sex ______ College Major ______

Marital Status ______ No. of Children ______ Year in School (number) ______

Grade point average (based on A=4) ______

Occupation: Father ________________

Mother ________________

Husband ________________

Wife ________________

Yourself ________________

Years (number) of schooling completed by: Father ______

Mother ______

Husband ______

Wife ______

Yourself ______

Number of brothers ______ Ages of brothers ______

Number of sisters ______ Ages of sisters ______

B: List additional courses you have taken or are now taking at Oregon State University in the appropriate space below. From that list and the one shown (1) Circle courses you have taken, and (2) place a rectangle around those courses in which you are now enrolled.

Family Life:

FL 222
FL 223
FL 225
FL 311
FL 312
FL 322
FL 413
FL 421
FL 423
FL 425
FL 426
FL 427
FL 428
FL 430
FL 435
FL 481

List other courses:

Sociology:

Soc 204 Soc 205 Soc 206

List other courses:

Anthropology:

Anth 207 Anth 208 Anthe 106X

List other courses:

Education:

List courses:

Psychology:

Psy 200

C: List any previous experiences you have had with children (i.e. baby sitting, teaching, etc.)