DOCUMENT RESUME

ED 190 608 TM 800 400

AUTHOR Saunders, Joseph C.: Huynh, Huynh
TITLE Consideration for Sample Size in Reliability Studies for Mastery Tests. Publication Series in Mastery Testing.
INSTITUTION South Carolina Univ., Columbia. School of Education.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO RM-80-3
PUB DATE Mar 80
GRANT NIE-G-78-0087
NOTE 14p.: Paper presented at the Annual Meeting of the Eastern Educational Research Association (Norfolk, VA, March 5-9, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cutting Scores: Difficulty Level: *Error of Measurement: *Mastery Tests: Primary Education: *Scoring Formulas: Test Items: *Test Reliability
IDENTIFIERS Comprehensive Tests of Basic Skills: *Sample Size: Test Length

ABSTRACT

In most reliability studies, the precision of a reliability estimate varies inversely with the number of examinees (sample size). Thus, to achieve a given level of accuracy, some minimum sample size is required. An approximation for this minimum size may be made if some reasonable assumptions regarding the mean and standard deviation of the test sccre distribution can be made. To facilitate the computations, tables are developed based on the Comprehensive Tests of Basic Skills. The tables may be used for tests ranging in length from 5 to 30 items, with percent cutoff scores of 60%, 70%, or 80%, and with examinee populations for which the test difficulty can be described as low, moderate, or high, and the test variability as low or moderate. The tables also reveal that for a given degree of accuracy, an estimate of kappa would require a considerably greater number of examinees than would an estimate of the raw agreement index. (Author)

PUBLICATION SERIES IN MASTERY TESTING
University of South Carolina
College of Education
Columbia, South Carolina 29208

Research Memorandum 80-3
March. 1980

CONSIDERATIONS FOR SAMPLE SIZE IN RELIABILITY STUDIES FOR MASTERY TESTS

Joseph C. Saunders
Huynh Huynh

University of South Carolina

Presented at the annual conference of the Eastern Educational Research Association, Norfolk, Virginia, March 5-9, 1980.

ABSTRACT

In most reliability studies, the precision of a reliability estimate varies inversely with the number of examinees (sample size). Thus, to achieve a given level of accuracy, some minimum sample size is required. An approximation for this minimum size may be made if some reasonable assumptions regarding the mean and standard deviation of the test score distribution can be made. To facilitate the computations, tables are developed based on the Comprehensive Tests of Basic Skills. The tables may be used for tests ranging in length from five to thirty items, with percent cutoff scores of 60%, 70%, or 80%, and with examinee populations for which the test difficulty can be described as low, moderate, or high, and the test variability as low or moderate. The tables also reveal that for a given degree of accuracy, an estimate of kappa would require a considerably greater number of examinees than would an estimate of the raw agreement index.

This work was performed pursuant to Grant No. NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator.

1. INTRODUCTION

In many applications of educational and psychological testing, an empirical demonstration of the reliability of the measuring instrument is desirable. Such demonstration is most meaningful when the estimate for the reliability has been cbtained with a reasonable degree of accuracy. That is, the standard error of estimate must be within some acceptable limit. In most instances, the standard error is a decreasing function of the number of examinees (sample size) to be included in the reliability study. Thus, some minimum sample size is needed to achieve a given level of precision. The purpose of this paper is to illustrate how this sample size can be assessed in estimating the reliability of mastery tests.

The paper consists of three major parts. The first part presents an overview of the procedures for estimating two reliability indices for mastery tests by using data collected from one test administration. The use of the estimation process to determine the minimum sample size is illustrated in the second part. Finally, a set of tables is developed to facilitate the determination of the minimum sample size in reliability studies for mastery tests.

2. OVERVIEW OF SINGLE-ADMINISTRATION ESTIMATES FOR RELIABILITY

Mastery tests are commonly used to classify examinees into two achievement categories, usually referred to as mastery and nonmastery. The reliability of such tests is often viewed as the consistency of mastery-nonmastery decisions. It may be quantified via the raw agreement index (p) or the kappa index (κ). The p index is simply the combined proportion of examinees classified consistently as masters or nonmasters by two repeated testings using the same form or two equivalent forms of a mastery test. The kappa index, on the other hand, takes into account the level of decision consistency which would result from random category assignment. It expresses the extent to which the test scores improve the consistency of decisions beyond the chance level.

Though both p and κ are defined in terms of repeated testings, there are many practical situations in which they may be estimated from the scores collected from a single test administration (Huynh, 1976). The estimation process assumes that the test scores conform to a beta-binomial (negative hypergeometric) model, and may be carried out via formulae, tables, and a computer program reported elsewhere (Huynh, 1978; 1979). The data reported by Subkoviak (1978) and by Huynh and Saunders (1979) tend to indicate that the beta-binomial model yields reasonably accurate estimates for p and κ in situations involving educational tests such as the Scholastic Aptitude Test and the Comprehensive Test of Basic Skills.

The beta-binomial model also provides asymptotic (large sample) standard errors for the estimates. Simulation studies indicate that the asymptotic standard errors tend to underestimate the actual standard errors when the sample size is small (Huynh, 1980). The degree of underestimation is not substantial when the sample has sixty or more examinees. Since the beta-binomial model will be used throughout the remaining part of this paper, a minimum sample size of sixty examinees will be assumed to hold uniformly for all cases under consideration.

3. ILLUSTRATIONS FOR SAMPLE SIZE DETERMINATION

The standard error (s.e.) of estimates for p and for κ are functions of sample size m. The quantity G = s.e. × $\sqrt{m}$ is asymptotically (i.e., in large samples) a constant, however. This constant depends only on the number of items (n), the mean (μ) and standard deviation (σ) of the test scores, and the cutoff score (c). Given the availability of these parameters, the value of G may be determined via the tables or the computer program presented elsewhere (Huynh, 1978). Once G is determined, a minimum sample size m can be calculated which will restrict the standard error of estimate to whatever tolerable range is required.

Suppose, for example, that an estimate of κ is needed for a short ($n = 6$ items) test to be used with a particular population of students. Passing or mastery on the test is to be granted if an examinee attains a score of 5 or 6. Further, suppose that we want the standard error of this estimate to be smaller than 10% of κ, that is, s.e. $(\kappa) \leq .10\kappa$.

What sample size would be needed to obtain the specified degree of accuracy in the estimate? To answer this question using the above mentioned Huynh procedure, a preliminary knowledge of the test mean and standard deviation is needed. Suppose past data suggest that the students are generally well-prepared on the content of the test in question and can be expected to be fairly homogeneous in achievement. We might suppose that in the population the mean will be 5.0 and the standard deviation will be 1.2. Using these values, and the cutoff score of 5, a value of G can be read from the tables (or computed): $G(\kappa) = .7390$. If the population mean and standard deviation are as given, then, assuming the beta-binomial model, the population value of κ is .3778. These results are then used to estimate the sample size needed to bring the standard error of estimate with the desired limits (i.e. less than $.10\kappa$).

Since the standard error of estimate is approximately $G/\sqrt{m}$, the standard error must be such that

$$\frac{G(\kappa)}{\sqrt{m}} \leq .10\kappa$$

or, equivalently,

$$m \geq [G(\kappa)/.10\kappa]^2.$$

For this example, then,

$$m \geq [.7390/(.10)(.3778)]^2 = 382.62.$$

Thus, to have no more than 10% relative error requires that at lease 383 examinees be tested to estimate κ.

A similar computation can be made for s.e. $(p) \leq .10p$ when the above assumed population values hold. Thus, using the tables,

$$G(p) = .3210,$$

$$p = .7532,$$

and

$$m \geq [G(p)/.10p]^2 = 18.16.$$

Because of the previously mentioned problems of underestimation in small samples, a sample size of at least sixty is recommended regardless of the above computation.

It might be disheartening to note that a much larger sample size is needed to keep the standard error of the κ estimate within the desired limits than is required when an estimate of p is used. However, the standard error for κ is much larger than that of p (Huynh, 1978). Thus, for the same relative size of errors of estimation, larger samples are needed to estimate κ than to estimate p. It could be argued that the same degree of accuracy of estimation is not required. If so, then a less accurate estimate of κ would allow a smaller sample size.

The above illustration presumes that the mean and standard deviation of the test scores can be projected prior to the real test administration. In a number of instances involving the use of standardized tests for a heterogeneous group of students, reasonable assumptions may be made, which will yield projected values for both μ and σ. For example, when an n-item multiple-choice is built to maximize the discrimination among individual examinees, it is not unreasonable to assume that the test mean is half way between the expected chance score and the maximum score n, and that the standard deviation is about one-sixth of the test score range from 0 to n. (If there are A options per item, the expected chance score is n/A.) In other words, it is not unreasonable to presume that

$$\mu = (n+n/A)/2$$

and

$$\sigma = n/6.$$

For example, consider a test consisting of 10 four-option items. Then A = 4, and the projected mean and standard deviation are

$\mu = 6.25$ and $\sigma = 1.66667$. Presuming a cutoff score of $c = 6$, it may be found that $p = .6140$, $G(p) = .3661$, $\kappa = .1118$, and $G(\kappa) = .8213$. If a relative error of 5% is acceptable for p, then a sample of at least $[.3661/(.05\times.6140)]^2 = 143$ students would be needed. On the other hand, a relative error of 25% for kappa would require $[.8213/(.25\times.1118)]^2 = 864$ students.

4. PRACTICAL CONSIDERATIONS IN SETTING SAMPLE SIZE IN BASIC SKILLS TESTING

Some general formulae are given for expressing the relationships among s.e., G, m, p, k, and the proportion of sampling error desired in an estimate. These general expressions will then be used in a series of simulations designed to explore their typical numerical values for real tests. Tables are developed to help the practitioner decide on the sample size needed to obtain estimates of p and κ for various degrees of precision.

General expressions

Since $G = \text{s.e.} \times \sqrt{m}$ is a constant for large samples, this expression forms the basis for the formulations in this section. In the previous section .10 and .05 were used as examples of desired degrees of precision for a sample estimate of p. In general, we will call this quantity γ, using γ_p and γ_κ to distinguish precisions desired for p and κ, respectively. Thus, the general expressions for minimum sample size are:

$$m \geq \left[\frac{G(p)}{\gamma_p p}\right]^2$$

and

$$m \geq \left[\frac{G(\kappa)}{\gamma_\kappa \kappa}\right]^2$$

A further simplification is to let $R(p) = [G(p)/p]^2$ and $R(\kappa) = [G(p)/\kappa]^2$. The above expressions for minimum sample size, m, become

$$m \geq R(p)/(\gamma_p)^2$$

and

$$m \geq R(\kappa)/(\gamma_\kappa)^2.$$

These expressions will allow minimum sample size to be determined from knowledge of two quantities, R and γ.

Determining typical values of R(p) and R(κ)

In practical applications, the values R(p) and R(κ) depend on a test score distribution which is not yet available. So, as in the previous section, conjectures must be made regarding the mean and standard deviation of the test score in order to project the minimum sample size.

In this section, typical values for R(p) and R(κ) will be reported for practical testing situations involving the assessment of basic skills. Several combination of test length, difficulty, variability, and cutoff scores will be used. To arrive at the values of R(p) and R(κ) reported in Tables 1-3, the following series of steps was taken.

First, a series of subtests was developed, using items found in the Comprehensive Test of Basic Skills (CTBS), Form S, Level 1. The items composing each subtest were randomly selected from one of five CTBS content areas, to reflect a variety of subjects and skills. For each content area, subtests were constructed with 5, 10, 15, 20, 25, and 30 items, producing a total of 30 subtests.

Second, the administration of the subtests was simulated using actual student responses. Data for the simulation came from 5,543 students, comprising a systematic sample (every tenth case) of the third grade students tested using Level 1 of the CTBS by the 1978 South Carolina Statewide Testing Program. From the students' responses to each item in the CTBS, raw scores were generated for each student on all 30 subtests.

Third, values of the mean and standard deviation of raw scores on each test were obtained. District means and standard deviations were calculated for each school district with 40 or more students in the sample. For each of the 30 subtests, means and standard deviations were plotted in a bivariate scatter diagram. The scatter-plots were divided into areas representing different categories of test difficulty and variability. Then districts were selected with means and standard deviations considered to be typical of six categories of difficulty and variability. These six categories (tests of low, moderate, and high difficulty, with low and moderate variability) were chosen to represent types of test score distributions typically encountered in mastery testing.

Fourth, the typical values obtained in the previous step were used to determine R(p) and R(κ). For each of the 30 subtests, the computer program described elsewhere (Huynh, 1978) was used to obtain estimates of G(p), p, G(κ), and κ when the cutoff scores were equivalent to 60%, 70%, and 80%. These data were used to calculate R(p) and R(κ) in each case.

Finally, the values of R(p) and R(κ) obtained above were averaged over the five CTBS content areas and the resulting values were compiled in tabular form. Tables 1, 2, and 3 provide values of R(p) and R(κ) for percent cutoff scores of 60%, 70%, and 80%, respectively.

The data needed to enter the tables are: (1) test length (n), (2) an idea of test difficulty (high, moderate, or low), (3) test variability (low or moderate), and (4) percentage cutoff score (60%, 70%, or 80%). The minimum sample size needed is simply R/γ^2, that is, the value of R obtained from the tables divided by the square of the acceptable proportion of sampling error in the estimate.

Numerical example

Suppose a study is planned to assess the reliability of a twenty-item test ($n = 20$) using the kappa index when a cutoff score of 14 ($c = 70\%$) is employed. The students for whom the test is

TABLE 1

Values of R for p and κ for Six Categories of Tests at the Percent Cutoff Score of 60%

Test Category			Number of Items					
(diff)	(var)		5	10	15	20	25	30
High	Low	(p)	0.219	0.075	0.050	0.031	0.023	0.018
		(κ)	5.349	1.623	0.666	0.391	0.307	0.209
High	Mod	(p)	0.164	0.061	0.036	0.025	0.018	0.014
		(κ)	2.589	0.908	0.327	0.280	0.209	0.139
Mod	Low	(p)	0.244	0.085	0.056	0.032	0.025	0.020
		(κ)	5.809	1.485	0.613	0.367	0.269	0.200
Mod	Mod	(p)	0.148	0.068	0.036	0.027	0.021	0.015
		(κ)	2.215	0.838	0.312	0.266	0.198	0.126
Low	Low	(p)	0.199	0.095	0.044	0.031	0.025	0.020
		(κ)	5.502	1.345	0.560	0.365	0.247	0.186
Low	Mod	(p)	0.142	0.068	0.032	0.024	0.020	0.016
		(κ)	2.371	0.770	0.298	0.249	0.176	0.128

intended are known to be a homogeneous group of relatively high ability. Thus, it might be expected that the test would be of low difficulty (i.e., easy), with low variability. Let us say that a fairly precise estimate of κ is desired, so γ_κ is set at .05. Entering Table 2, in the row corresponding to low difficulty and low variability, it if found that R(κ) for n = 20 items is .362. The minimum sample size needed to estimate kappa with 5% allowable error is then computed as $m = R(\kappa)/\gamma_\kappa^2 = .362/(.05)^2 = 144.8$. Thus, a sample of at least 145 students is necessary to achieve the desired degree of precision. If reliability is to be determined via the raw agreement index p, a similar procedure is followed using R(p) and γ_p. Again, at least 60 students should be used in the sample, even if it is found that m < 60.

TABLE 2

Values of R for p and κ for Six Categories of Tests at the Percent Cutoff Score of 70%

Test Category			Number of Items					
(diff)	(var)		5	10	15	20	25	30
High	Low	(p)	0.219	0.075	0.046	0.029	0.022	0.017
		(κ)	5.349	1.623	0.776	0.455	0.410	0.272
High	Mod	(p)	0.164	0.061	0.033	0.023	0.017	0.013
		(κ)	2.589	0.908	0.360	0.324	0.276	0.178
Mod	Low	(p)	0.244	0.085	0.053	0.031	0.023	0.019
		(κ)	5.809	1.485	0.646	0.396	0.322	0.242
Mod	Mod	(p)	0.148	0.068	0.035	0.026	0.019	0.014
		(κ)	2.215	0.838	0.321	0.289	0.237	0.149
Low	Low	(p)	0.199	0.095	0.050	0.031	0.024	0.019
		(κ)	5.502	1.345	0.512	0.362	0.265	0.203
Low	Mod	(p)	0.142	0.068	0.036	0.023	0.019	0.015
		(κ)	2.371	0.770	0.280	0.254	0.190	0.137

Some observations on the tabled values

In every case R(κ) > R(p). This fact implies that the sample size necessary to estimate kappa will be larger than that needed to estimate p, for any fixed degree of precision, γ. As noted previously, practical limitations may require that larger proportions of error be tolerated when estimating kappa than when estimating p.

R-values for the case of low variability are larger than those for moderate variability. If there is doubt about the expected degree of variability, the value of R for the low variability case would produce the more conservative estimate of m.

R decreases as the number of test items increases. The relationship between R and n is not linear, however. Hence, linear interpolation would not be appropriate for determining R for non-

TABLE 3

Values of R and p and κ for Six Categories of Tests at the Percent Cutoff Score of 80%

Test Category			Number of Items					
(diff)	(var)		5	10	15	20	25	30
High	Low	(p)	0.132	0.063	0.032	0.021	0.018	0.013
		(κ)	7.076	2.805	1.494	1.055	0.887	0.660
High	Mod	(p)	0.098	0.045	0.024	0.018	0.015	0.011
		(κ)	3.510	1.678	0.608	0.717	0.568	0.404
Mod	Low	(p)	0.174	0.064	0.038	0.025	0.020	0.015
		(κ)	6.831	2.283	1.087	0.812	0.640	0.558
Mod	Mod	(p)	0.113	0.047	0.026	0.021	0.017	0.012
		(κ)	2.633	1.337	0.484	0.571	0.458	0.311
Low	Low	(p)	0.189	0.060	0.044	0.029	0.022	0.017
		(κ)	5.849	1.906	0.652	0.611	0.471	0.417
Low	Mod	(p)	0.122	0.046	0.029	0.023	0.018	0.014
		(κ)	2.675	1.113	0.348	0.430	0.325	0.248

tabled values of n. The value of R listed for the largest tabled n less than the actual number of items should yield a conservative estimate for m. For example, suppose the test considered in the numerical example above actually contained 22 items. The tabled value of R corresponding to n = 25 would produce an underestimate of m, and the resulting proportion of error in estimating kappa would exceed γ_κ. The R-value for n = 20 would overestimate m, and the observed proportion of error would then be less than γ_κ.

The relationships between R and test difficulty or cutoff scores are more complex. No simple trends can be observed in the tables. In many testing situations, the cutoff score typically ranges from 60% to 80% correct. For cutoff scores falling between the values in the tables, find R for both bracketing values and use the larger. Again, consider the situation in the numerical example above.

Suppose t' : cutoff score was 13 (65% correct). From Tables 1 and 2, the values of R corresponding to c = 60% and 70% are .365 and .362, respectively. The larger of these (corresponding to c = 60%) should provide a reasonable value for R.

4. CONCLUSIONS

In this paper, an approximation method has been presented for determining the minimum sample size necessary to achieve a specified degree of precision in estimating raw agreement (p) and kappa (κ) indices of reliability for mastery tests. The method uses the quantity R which can be calculated for known test score distributions. Tables of R have been constructed for test score distributions typically found in mastery testing, for a variety of test lengths and cutoff scores. In addition, suggestions have been made for obtaining reasonable estimates of R for situations not directly covered by the tables.

Of course, precision is only one of the factors that must be considered in any study. Feasibility, cost, and classroom management considerations also play important roles. However, knowledge of necessary sample sizes should facilitate and simplify the planning of reliability studies. The tables presented here should be particularly useful for tests involving the basic skills, and perhaps other tests of similar construction.

BIBLIOGRAPHY

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement 13, 253-264.

Huynh, H. (1978). Computation and inference for two reliability indices in mastery testing based on the beta-binomial model. Research Memorandum 78-1, Publication Series in Mastery Testing. University of South Carolina College of Education.

Huynh, H. (1979). Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. Journal of Educational Statistics 4, 231-246.

Huynh, H. (1980). Adequacy of the asymptotic error in estimating reliability for mastery tests based on the beta-binomial model. Research Memorandum 80-2, Publication Series in Mastery Testing, University of South Carolina College of Education.

Huynh, H. & Saunders, J. C. (1979). Accuracy of two procedures for estimating reliability of mastery tests. Research Memorandum 79-1, Publication Series in Mastery Testing. University of South Carolina College of Education.

Subkoviak, M. J. (1978). Empirical investigation of procedures for estimating reliability of mastery tests. Journal of Educational Measurement 15, 111-116.

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare. Huynh Huynh, Principal Investigator. Points of view or opinions stated do not neccessarily reflect NIE positions or policy and no official endorsement should be inferred. The comments of Anthony J. Nitko and Elizabeth M. Haran are gratefully acknowledged.