DOCUMENT RESUME

ED 270 482 TM 860 365

AUTHOR Hambleton, Ronald K.
TITLE Determining Optimal Test Lengths with a Fixed Total Testing Time.
PUB DATE 21 Mar 86
NOTE 17p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Algorithms; *Criterion Referenced Tests; Elementary Secondary Education; Psychometrics; Scores; Statistical Analysis; Statistical Studies; *Test Construction; Test Format; Testing Problems; *Test Length; Test Reliability; Test Validity; *Timed Tests

ABSTRACT

The problem of determining optimal test lengths with fixed total testing time has proved to be a difficult one for criterion-referenced test developers. An algorithm is needed which can be used by test developers to allocate available testing time to maximize the validity of their total criterion-referenced tests or testing programs. To be maximally useful, the algorithm should allow test developers to specify: (1) the various informational needs they have in relation to the objectives measured by a test or the uses they have for several tests in their testing program; and (2) the relative importance that they attach to their informational needs. This paper describes such an algorithm for determining the number of items to measure each objective in a criterion-referenced test when testing time is fixed and when the objectives, in general, vary in their levels of importance, reliability, and validity. Four examples in the paper highlight possible applications of the procedure. While the offered solution is simple to apply, it is only applicable when basic psychometric data are available on the scores at the objective level and when the objective scores are being used to make descriptive statements about examinee performance. (PN)

Determining Optimal Test Lengths with a Fixed Total Testing Time

Ronald K. Hambleton
University of Massachusetts at Amherst

Abstract

The problem of determining optimal test lengths with fixed total testing time has proved to be a difficult one for criterion-referenced test developers. A solution is offered in this paper in which test lengths are determined by considering the importance, reliability, and validity of scores measuring objectives in a criterion-referenced test with a constraint set on the total desired test length.

TESTLEN.1

3/21/86

Determining Optimal Test Lengths with a Fixed Total Testing Time

Ronald K. Hambleton
University of Massachusetts at Amherst

A problem which arises nearly every time a criterion-referenced test is constructed is an instance of the bandwidth-fidelity dilemma (Cronbach and Gleser, 1965). The dilemma is the following: When the total test length is fixed, a test developer must decide whether it is more useful to measure a relatively small number of objectives precisely or a larger number of objectives less precisely. For our purposes here, available testing time is defined in terms of the total number of test items which can be administered within the available testing time.

A good example of the dilemma arises in the United States Army's Skills Qualification Testing Program (SQT) (Department of the Army, 1986). The SQTs are used to assess the competencies of soldiers in relation to the jobs they do. Test results are very important to a soldier's Army career: A soldier's test performance influences promotion decisions. Each Army job specialty is defined by a set of tasks (which correspond to objectives). As the maximum two- to

three-hour time limit for SQTs is not usually sufficient time to allow test developers to assess all the tasks describing a soldier's job, some type of sampling must be done. The following form of the problem introduced earlier arises: Is it better to sample a large number of the soldiers' tasks and use only a few test items per task, or to sample a smaller number of tasks and use a larger number of test items per task. The first strategy provides more comprehensive job domain coverage but the accuracy of mastery-non-mastery decisions at the task level may be low. The second strategy affords the reverse outcome: job domain coverage is less comprehensive; but the accuracy of mastery-non-mastery decisions at the task level is higher.

The dilemma also arises in many school testing programs. For example, there will be placement tests, pre- and post-unit tests, curriculum embedded tests, and year-end achievement tests (Hambleton, 1974). As there is rarely sufficient testing time available to assess examinee domain score performance in relation to the objectives of interest with high levels of reliability and validity for all test uses, the major portion of the available testing time must be assigned to the most important objectives and/or uses. When the domain scores serve important uses (for example, determining examinee performance levels on objectives which are prerequisites to later objectives in a school curriculum) high test score validity is extremely important. Longer tests to measure the most important objectives are needed. On the other hand, with the less important objectives and/or with the

less important test score uses (e.g., diagnoses on objectives for which minimal instructional time is needed), shorter tests seem justified.

An algorithm is needed which can be used by test developers to allocate available testing time to maximize the validity of their total criterion-referenced tests or testing programs. To be maximally useful, the algorithm should allow test developers to specify (a) the various informational needs they have in relation to the objectives measured by a test or the uses they have for several tests in their testing program (e.g., measurement of all objectives once for diagnoses in the school year, and measurement of the most important objectives on a final examination) and (b) the relative importance that they attach to their informational needs.

Woodbury and Novick (1968) and Jackson and Novick (1970) provided the definitive answer for allocating a fixed amount of testing time to subtests of norm-referenced tests. However, the work of Novick and his colleagues is not applicable to the problems at hand because (a) there is no mechanism for a priori weighting of objectives or tests to reflect their relative levels of importance to test developers, and, more importantly, (b) the function Novick, Jackson, and Woodbury chose to maximize, the multiple correlation of a test battery with a fixed criterion, is not appropriate for the test length problems that arise with criterion-referenced tests. The effectiveness of a criterion-referenced test or testing program cannot usually be determined by correlating a composite criterion-referenced test score derived from a

number of objectives or tests, with a single criterion measure. It seems more appropriate to validate each objective or test against an appropriately chosen criterion measure. The effectiveness of a criterion-referenced test or testing program can then be assessed (for example) by summing the squared validity coefficients of individual objectives or tests scaled by weights reflecting their levels of relative importance. Nevertheless, the general approach of Novick and his colleagues proved to be useful in the current research.

Hambleton (1984) reviewed five promising methods in the psychometric literature for determining criterion-referenced test lengths. None of these methods, however, introduces testing time as a constraint, though the time constraint is a common one in test development work. The two major purposes of the research described in this paper were: (a) to prepare an algorithm for determining the number of items to measure each objective in a criterion-referenced test when testing time is fixed and when the objectives, in general, vary in their levels of importance, reliability, and validity, and (b) to present the results from several applications of the procedure.

The choice of loss function in this research applies to only one of two popular uses of criterion-referenced test scores: assessment of examinee levels of performance in relation to well-defined domains of content measuring the objectives (or competencies) of interest. The more popular use of criterion-referenced test scores which involves making mastery and non-mastery decisions is not directly addressed in the present research.

The current research is to be described in terms of the test lengths to measure objectives included in a single test because this test length determination problem is very familiar to criterion-referenced test developers (Hambleton, 1984). However, the paradigm can be adapted easily to fit the situation where the objectives become tests themselves designed to accomplish different purposes such as diagnosis, unit mastery, and year-end assessment.

Procedures for Determining Test Lengths

Assumptions and Considerations

Several assumptions underlying the research study were:

1. A large pool of valid test items is available to measure each objective.
2. The amount of testing time which is fixed can be specified in the form of the total number of test items that can be administered.
3. Test items require approximately equal amounts of administration time. Thus, the research results can be most safely applied to tests using the same item formats and possessing a limited range of item difficulty levels.
4. A criterion-referenced test measures several objectives.
5. Examinee performance data on each objective is used to make domain score estimates.
6. The correlation between domain scores estimates obtained with each objective and an appropriately chosen criterion measure (the test validity coefficient) can be used as an evaluative measure of the test effectiveness.

Within the context of the assumptions above, test length (i.e., the number of test items) is the major factor influencing the reliability and validity of domain score estimates.

A procedure is needed to help test developers determine the number of test items to measure each objective. The procedure must provide a way to recognize several important considerations in an optimal solution: (a) some objectives are more important than others, (b) some objectives are more difficult to measure than others, (c) criterion-referenced tests are used to accomplish different purposes which vary in their levels of importance, and (d) testing time should be kept low (in most situations).

Steps to be Taken

To begin with, the test developer must identify the n objectives to measure (D_i, $i=1, 2, \ldots, n$), and the weight (W_i, $i=1, 2, \ldots, n$) to reflect the relative importance of the score on each objective. For example, one could suppose that with 50 objectives, 10 are broader and more important than the remaining 40. With respect to D_i, $i=1, 2, \ldots, 50$, modest levels of score reliability and validity may be sufficient for the 40 narrow instructional objectives, whereas considerably higher levels may be needed for the 10 broader objectives. With respect to the relative importance of the informational needs, a test designed to assess domain scores which are used to influence instructional decisions which may last for a week or two (e.g., assignment of some remedial work) are certainly more important than informational needs which might affect students for a single day. If the former informational need is D_1 and the latter D_2, one might set

$\underline{W}_1$ = .67 and $\underline{W}_2$ =.33 to reflect the relative judged importance of the two informational needs. In the example just cited with 50 objectives, it may be reasonable to set $\underline{D}_i$ = .01, $\underline{i}$ = $\underline{1}$, $\underline{2}$, ..., $\underline{40}$; and $\underline{D}_i$ = .06, $\underline{i}$=$\underline{41}$, $\underline{42}$, ..., $\underline{50}$. This assignment is equivalent to attaching 40% of the weight to the specific instructional objectives and the remainder of the weight (60%) to the broader outcomes of instruction.

It is well known that

$$\rho^2(\underline{Z}_i, \underline{Y}_i) = \frac{\underline{k}_i \rho^2(\underline{X}_i, \underline{Y}_i)}{1+(\underline{k}_i-1)\rho(\underline{X}_i, \underline{X}_i')}, \qquad [1]$$

where $\rho(\underline{Z}_i, \underline{Y}_i)$ is the correlation between a lengthened criterion-referenced test $\underline{Z}_i$ and a criterion measure $\underline{Y}_i$, where $\rho(\underline{X}_i, \underline{Y}_i)$ and $\rho(\underline{X}_i, \underline{X}_i')$ are the corresponding validity and reliability indices of the unit test length $\underline{X}_i$, and where $\underline{k}_i$ is the factor by which the test is lengthened (Lord and Novick, 1968). To allow that $\underline{k}_i$ can be interpreted as test length (or the number of items measuring the objective), in the work to follow, only single-item reliability and validity coefficients will be used in Equation [1]. Now $\rho^2(\underline{Z}_i, \underline{Y}_i)$ is a measure of the predictive efficiency of $\underline{Z}_i$ for predicting $\underline{Y}_i$ (the criterion test for the ith objective). Its value depends, among other things, on $\underline{k}_i$. A desirable goal is to maximize the predictive efficiency associated with the set of objective scores in the test. That is, an expression such as

$$\sum_{i=1}^{n} \rho^2(\underline{Z}_i, \underline{Y}_i) \qquad [2]$$

seems desirable to maximize, or even better

$$\sum_{i=1}^{n} \underline{W}_i \rho^2(\underline{Z}_i, \underline{Y}_i). \qquad [3]$$

In equation [3] a weight to reflect the relative importance of each objective (informational need) is introduced

$$(\sum_{i=1}^{n} \underline{W}_i = 1).$$

The squared validity coefficient for a set of scores measuring an objective is commonly used as an indicator of that test's usefulness. In our case, where there are several objective scores and each score is weighted a priori for its importance, the sum of weighted and squared validity coefficients seemed to serve as an appropriate criterion function for representing the predictive efficiency of the objective scores in the criterion-referenced test.

There is a need for one constraint, the total amount of testing time (operationalized as the total number of test items which can be administered in the testing time available, denoted $\underline{T}$). This constraint is necessary because seldom will a test developer have complete freedom to determine the length of a criterion-referenced test. For example, one could suppose that an instructor is assigned 40 hours of time with a group of students (e.g., a one-semester course). If the instructor is prepared to expend 5% of the available time on assessment and if it is assumed that typical multiple-choice test items require about three-quarters of a minute to complete, then $\underline{T} = 160$.

The problem then of allocating testing time becomes one of determining $\underline{k}_i$, $\underline{i}=\underline{1}$, ..., $\underline{n}$, with the constraint, $\sum_{i=1}^{n} \underline{k}_i = \underline{T}$, so as to maximize the function, $\sum_{i=1}^{n} \underline{W}_i \rho^2(\underline{Z}_i, \underline{Y}_i)$.

The problem becomes one of differentiating the expression

$$\underline{S} = \sum_{i=1}^{n} \underline{W}_i \rho^2(\underline{Z}_i, \underline{Y}_i) - \lambda\left(\sum_{i=1}^{n} \underline{k}_i - \underline{T}\right) \qquad [4]$$

with respect to $\underline{k}_i$, $\underline{i}=\underline{1}$, $\underline{2}$, ..., $\underline{n}$, and with respect to the Lagrange multiplier λ. The partial derivatives can be written as

$$\frac{\partial \underline{S}}{\partial \underline{k}_i} = \frac{\underline{W}i\ \rho^2(\underline{X}i,\underline{Y}i)[1-\rho(\underline{X}i,\underline{X}i')]}{[1+(\underline{k}_i-1)\rho(\underline{X}_i, \underline{X}_i')]^2} - \lambda \qquad [5]$$

for $\underline{i}=\underline{1}$, $\underline{2}$, ... $\underline{n}$,

and

$$\frac{\partial S}{\partial \lambda} = \sum_{i=1}^{n} \underline{k}_i - \underline{T}, \qquad [6]$$

Setting the partial derivatives to zero and solving, one obtains values for $\underline{k}_i$, $\underline{i}=\underline{1}$, $\underline{2}$, ..., $\underline{n}$.

After use of some algebra, the solution can be shown to be

$$\underline{k}_i = \frac{\lambda^{-\frac{1}{2}}\{\underline{W}_i\rho^2(\underline{X}_i,\underline{Y}_i)[1-\rho(\underline{X}_i,\underline{X}_i')]\}^{\frac{1}{2}}-1}{\rho(\underline{X}_i, \underline{X}_i')} + 1 \qquad [7]$$

for $\underline{i}=\underline{1}$, $\underline{2}$, ..., $\underline{n}$,

and λ can be obtained by solving,

$$\sum_{i=1}^{n} \underline{k}_i = \underline{T}.$$

To apply the procedure, it is necessary to have reliability and validity estimates for the assessment of each objective $\underline{i}$ 'adjusted to "single-item" estimates), $\underline{i}=\underline{1}$, ..., $\underline{n}$; $\underline{W}_i$, $\underline{i}=\underline{1}$, $\underline{2}$, ..., $\underline{n}$ ($\Sigma\underline{W}_i=1$); and $\underline{T}$.

The criterion variable for validating each objective score is often an independent measure of performance on the objective such as instructor judgment, or a longer set of test items measuring the objective.

Insert Table 1 About Here

Application of the Procedure

Table 1 presents the results for four special applications of the algorithm to help in understanding how the algorithm works. In the first application, the effect of variable weights on tests of equal reliability and validity can be seen. The more important tests are lengthened to obtain an optimal allocation of test items. Though the three tests were initially of the same length, the optimal solution resulted in one test being nearly 6 times longer than the shortest test. The second application highlights the role of test reliabilities in optimal solutions. Other factors equal (i.e., test length, validity, and relative importance), in an optimal solution designed to enhance test validity, tests with higher reliabilities are shortened, whereas tests with lower reliabilities are lengthened. This result from classical test theory is well-known (Lord and Novick, 1968).

The third application illustrates the effect of different validities on an optimal solution, other factors being equal. The more

valid tests are the ones that are lengthened in an optimal solution. Finally, in many practical test development efforts, factors such as relative importance, reliability and validity vary from one test to the next. The fourth application demonstrates this situation by showing a dramatic shift from equal test lengths when an optimal solution is obtained. The five tests were initially of equal length (20 items/objective) but in the optimal solution test lengths vary from 13 to 31 items.

Conclusion

The solution offered in this paper for allocating criterion-referenced testing time is simple to apply. It is, however, only applicable when some basic psychometric data are available on the scores at the objective level and whenever the choice of function to maximize used in this paper is reasonable. The function seems reasonable when the objective scores are being used to make descriptive statements about examinee performance. A typical descriptive statement is, "The student has demonstrated mastery of about 75% of the content spanned by the objective."

The four examples in the paper highlight possible applications of the procedure. One troublesome problem in any application is the necessity of having reliability and validity information for each pool of items matched to an objective. On the other hand, such psychometric data are important, whether or not they are to be used in an algorithm to allocate testing time.

TESTLEN.1

In our subsequent research, several questions will be addressed:

1. Will the algorithm lead to negative test lengths ($\underline{k}_i$, $\underline{i}=\underline{1}$, $\underline{2}$, ..., $\underline{n}$)?

2. Are there other more reasonable measures of testing program efficiency than

$$\sum_{i=1}^{n} \underline{W}_i \rho^2(\underline{Z}_i, \underline{X}_i) \quad ?$$

3. By what process can the relative weights be set? What factors should be considered?

4. How can additional constraints on the solution be incorporated? (For example, it may be desirable to insure that domain validity coefficients do not fall below some minimal value.)

5. Can analytic solutions be found for the case when mastery/non-mastery decisions are the primary purpose for testing?

In this paper, the algorithm was applied to a criterion-referenced test which measures several objectives. However, the same algorithm for allocating testing time can be applied nearly as easily to the more complex and common situation where there exist many criterion-referenced tests in use (each measuring several objectives) and these tests (and objectives within the tests) vary in their levels of importance, reliability, and validity.

References

Cronbach, L.J., and Gleser, G.C. (1965). Psychological tests and personnel decisions. Urbana: University of Illinois, 1965.

Department of the Army (1986). Skill qualification test and common task test development policy and procedures. (TRADOC Reg. 351-2). Fort Monroe, VA: U.S. Army Training and Doctrine Command.

Hambleton, R.K. (1974). Testing and decision-making procedures for selected individualized instructional programs. Review of Educational Research, 44, 371-400.

Hambleton, R.K. (1984). Determining test lengths. In R. Berk (Ed.), Guidelines for Preparing Criterion-Referenced Tests. Baltimore, MD: The Johns Hopkins Press.

Jackson, P.H., and Novick, M.R. (1970). Maximizing the validity of a unit-weight composite as a function of relative component length with a fixed total testing time. Psychometrika, 35, 333-347.

Lord, F.M., and Novick, M.R. (1968). Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley.

Woodbury, M.A., and Novick, M.R. (1968). Maximizing the validity of a test battery as a function of relative test lengths for a fixed total testing time. Journal of Mathematical Psychology, 5, 242-259.

TESTLEN.1

Footnote

[1] The author is grateful to Daniel Eignor and Linda Murray for preparing the computer program which was used to obtain several of the results in Table 1.

Table 1

Allocation of Testing Time to Maximize Predictive Efficiency

Example	Total Testing Time	Objective	Relative Weight	Single Item		Equal Length			Optimal Length		
				r[1]	v[2]	n	r	v	n	r	v
1	60	1	.60	.07	.14	20	.60	.40	34	.72	.44
		2	.30	.07	.14	20	.60	.40	20	.60	.40
		3	.10	.07	.14	20	.60	.40	6	.31	.29
2	80	1	.25	.31	.23	20	.90	.40	10	.81	.38
		2	.25	.13	.17	20	.75	.40	16	.71	.39
		3	.25	.07	.14	20	.60	.40	23	.63	.41
		4	.25	.04	.12	20	.45	.40	31	.56	.45
3	80	1	.25	.07	.24	20	.60	.70	42	.76	.79
		2	.25	.07	.17	20	.60	.50	26	.66	.52
		3	.25	.07	.10	20	.60	.30	10	.43	.26
		4	.25	.07	.10	20	.60	.30	2	.43	.26
4	100	1	.50	.05	.08	20	.50	.25	31	.61	.27
		2	.25	.05	.08	20	.50	.25	16	.44	.24
		3	.15	.10	.19	20	.70	.50	22	.72	.51
		4	.05	.17	.41	20	.80	.90	18	.78	.89
		5	.05	.17	.32	20	.80	.70	13	.72	.66

[1] r ≡ reliability estimate

[2] v ≡ validity estimate