我國公立高等教育預算效能之探討

鍾娟兒

摘 要

本研究的主要目的在探討我國公立高等教育預算的效能。首先從預算過程、，預算方式與預算控制三個層面來探討現行公立高等教育預算，此外亦發展了概念性的預算效能效標以爲評量現行制度的基礎。本研究設計是分析的與描述性的，文獻探討與文件分析爲主要的研究方法。

研究發現品質、績效與自主間之平衡、彈性、歧異性、公平、充足、理性與成本改進等八項爲預算效能效標，檢討發現我國公立高等教育預算只符合了績效與自主間的平衡此一效標，其餘的效標實應予以重視且應改進我國須算系統來符應這些效標。

本研究產出包括改進我國公立高等教育預算政策的有關建議。

A Comparison of Three Item Selection Methods in Criterion-Referenced Tests

Hui-Fen Lin

Abstract

In this study, three methods of criterion referenced item selection were compared on the reliability of their mastery/nonmastery classification. These three methods were (a) the agreement approach, (b) the phi coefficient approach, and (c) the random selection approach. The reliability coefficients were compared across three levels of test size and smaple size with a split plot ANOVA. The resulting statistics indicated that the agreement approach was the most appropriate method for selecting criterion-referenced test items at the classroom level, and the phi coefficient approach was most appropriate at the district and/or state levels. In compariosn with the phi and agreement approaches, the tests based on the random selection approach had the lowest reliabilities.

Introduction

In most applications of criterion-referenced testing procedures, test scores are used for making mastery decisions, that is, to determine whether students have mastered an instructional objective and therefore, are able to proceed to the next objective or have not mastered the objective and need further instruction (Hambleton, 1974; Swaminathan, Hambleton & Algina, 1974). In constructing a criterion-referenced test, the primary purpose is to select items that can be maximally sensitive to the differences between test scores corresponding to "masters" and "nonmasters" (Harrris, 1983).

Several statistics may be used to assist in the selection of items for criterion-referenced tests. These statistical tools are used to select items from a large pool of items so that the selected items will have the property of "differential sensitivity"; that is, masters tend to answer the item correctly and nonmasters tend to answer it incorrectly (Mellenbergh & van de Linden, 1982).

A number of item selection methods for criterion-referenced measurement have been proposed in the literature. According to Berk (1980b), each method has its specific theroretical perspective and statistical basis. He suggested that an appropriate item selection method should be "conceptually and computationally simple yet statistically sound" (p. 59). In this article, it contented that three of the more conceptaully and computationally simple criterion-referenced test item selection methods are the phi coefficient, the agreement approach, and the random selection of items from a pool of item. All three methods can be readily computed with a simple hand calculator and each provides results that can be easily understood with a minimum explanation. The first two require only a single administration from which to generate the tests and the random selection does not even require an initial administration in order to be used to select

criterion-referenced test tiems. Each method, therefore, meets Berk's first requirement for a test-statistic but it is unclear which of the three methods select itmes with the greatest degree of statistical soundness.

The Phi Coefficient Approach

According to Hsu (1971), the phi coefficient (ϕ) is computed as follows:

$$\phi = \frac{(n1 * n4) - (n2 * n3)}{\sqrt{(n1 + n2)*(n3 + n4)*(n1 + n3)*(n2 + n4)}}$$

where n1 = the number of masters answering the item correctly.
n2 = the number of nonmasters answering the item correctly.
n3 = the number of masters answering the item incorrectly.
n4 = the number of nonmasters answering the item incorrectly.

As mentioned before phi is easily computed, requires only one test administration, and does not depend on instruction as the index of discrimination. According to Ferguson (1981) , phi is best used with items scored correct/incorrect, and Swezey (1981) suggests that it function best when there are equal numbers of masters and nonmasters. According to Swezey, there appears to be several instances when phi is not appropriate such as when the responses of fewer than eight test-takers are being analyzed. In addition, it is not suitable in situations where every student is declared either a master or nonmaster, or when the item is answered correctly or incorrectly by every test-taker (Hsu, 1971).

The Agreement Approach

The agreement approach P(Xc) was proposed by Harris (1983). To use P(Xc), every test-taker is divided into master and nonmaster based upon a predetermined cut-off score. According to Harris (1983), the P(Xc) is computed as follows:

$$P(Xc) = (n1 + n4)/N$$

where n1 = the number of masters answering the item correctly.

n4 = the number of nonmasters answering the item incorrectly.

N = the total number of test-takers.

This approach, like the phi approach, is easy to compute, requires only one test administration, and does not depend on instructions as the basis of classification. In addition, Harris and Subkoviak (1986) claimed that this approach had similar theoretical characteristics to the latent trait approach.

The Random Selection

In the random selection method (RM), test items are randomly selected from the item pool. Popham (1978), Hambleton (1982), and Hambleton and Gruijter (1983) said that the random selection method was a common and straightforward strategy in constructing criterion-referenced tests. They stated that once criterion-referenced test items were developed consistently with a carefully defined domain of tasks, all items in the item pool should be homogeneous and interchangeable.

Methodogloy

Sampling

In this investigation, the responses of 1,836 students to a 50-item district developed criterion-referenced test were used. The instrument was developed to measure high school level physical science.

Swezey (1981) suggested that at least 50 percent more subjects be needed in the tryout sample than items in the initial item pool. In order to study the effect of sample size, samples of 75, 150, and 300 subjects were used, representing 1-1/2, 3, and 6 times of the number of items in the test being used in the investigation.

Hambleton, Mills, and Simon (1983) stated that test length could affect the reliability of mastery/nonmastery classification. Hambleton

(1984)also claimed "Low probability of misclassification can usually be assured when tests are very long" (p. 144). Popham (1978) suggested that a criterion-referenced test contain between ten to twenty items. Berk (1980b) suggested between five to ten criterion-referenced test items would be needed per objective for classroom decisions, and between ten and twenty items for school system and state level decisions. And Swezey (1981) recommended an item pool of about twice the number of items as required in the final version of a test. Based on these recommendations, three tests were selected: 15 items (between 10 and 20 items), 25 items (half of the 50 item pool), and 35 items (a longer test than the other two).

Ten samples each of 75, 150 and 300 responses were randomly selected from the population of student responses. The total test scores for each individual and the total numbers passing each item were obtained first. Next, student responses were rearranged from the highest to the lowest total scores. Since the uses of the agreement and phi coefficient require a mastery score, a cut-off score of 80% was established. The phi and agreement indices were then computed for each response sample.

After the item discrimination indices were computed, the items were sorted in descending order on the basis of first one index and then the other. Then the items with the 15, 25, and 35 highest phi values were selected as test forms; the same was then done with the agreement indices. A third selection of item forms was conducted by randomly selecting 15, 25, and 35 items for each of the samples. At this point in the investigation, there were 90 different data sets of test responses for three sample sizes composed of either 15, 25, or 35 items which had been selected based on the three methods previously discussed.

Analysis of Reliability of Classification

To compare the effect of these three selection methods on the reliability of mastery/nonmastery classification decisions, coefficients Po and K (Kappa) were calculated for the 90 data sets. Po represents the proportion of consistent decisions (mastery/mastery and

nonmastery/nonmastery) based on two parallel tests, which can be estimated from a single test administration. Coefficient K represents the proportion of consistent classifications corrected for chance, which can also be calculated from a single test administration.

To estimate Po and K coefficients, Huynh's (1976) beta-binomial method was employed because it requires only one test administration and assumes that all items in the item pool are interchangeable and homogeneous in difficulty and content.

Split Plot Analysis of Variance

When the coefficients had been computed for each of the data sets, once for the Po and once for K, a SPF-3.3 factorial analysis of variance design (Kirk, 1982) was used to analyze the effectiveness of the item selection methods at the three levels of test size. Item selection method represented the within factor variance and test size the between factor variance. The alpha level was set at .05 a priori.

Results

The Po analysis showed no significant effects due to the interaction of method-by-sample size or main effects of sample size. Significant main effects due to item selection method were found at the $p<.01$ level across test lengths (see Table 1). Scheffe post hoc comparisons were performed to further examine the mean differences among item selection methods. As shown on Table 2, the agreement approach produced tests with the highest Po values, and the random selection method produced tests with the lowest.

The results of the K analysis indicated that (1) significant effects due to item selection method were found at the $p<.01$ level across test lengths, and (2) significant effects due to the interaction of method-by-sample size and main effects of sample size were found in th 15-item tests (see Table 3).

Because there was a significant interaction effect in the 15-item tests, the tests of simple main effects due to item selection method in a given sample size and due to sample size in a given item selection method were performed, respectively. The results were shown in Table 4 and 5.

Table 1

Results of Split Plot Factorial Analysis of Variance for Po Across Test Lengths

Test Lengths	Sources	df	Sum of Square	Mean Square	F
15	Between Factor	29	0.0192		
	Sample Size	2	0.0008	0.00004	0.73
	Error	27	0.0184	0.00055	
	Within Factor	60	0.2231		
	Method	2	0.2105	0.10526	470.09 **
	Method * Sample Size	4	0.0005	0.00013	0.58
	Error	54	0.0121	0.000224	
25	Between Factor	29	0.0108		
	Sample Size	2	0.0016	0.0008	2.35
	Error	27	0.0092	0.00034	
	Within Factor	60	0.1415		
	Method	2	0.1286	0.06430	315.20 **
	Method * Sample Size	4	0.0019	0.00048	2.35
	Error	54	0.0110	0.000204	
35	Between Factor	29	0.00846		
	Sample Size	2	0.00223	0.000115	0.38
	Error	27	0.00823	0.000305	
	Within Factor	60	0.04929		
	Method	2	0.04650	0.02325	501.08 *
	Method * Sample Size	4	0.00023	0.000057	1.23
	Error	54	0.00251	0.0000464	

** Significant at $p<.01$.

* Significant at $p<.05$.

Table 2

Scheffé Post Hoc Comparisons for Po Mean Differences Among Item Selection Methods Across Three Test Lengths

Test Lengths	Methods	Mean	Mean Differences P(Xc)	ϕ	RM
15	P(Xc)	0.84	-----	0.03 *	0.11 *
	ϕ	0.81		-----	0.08 *
	RM	0.73			----
25	P(Xc)	0.86	-----	0.03 *	0.09 *
	ϕ	0.83		-----	0.06 *
	RM	0.77			----
35	P(Xc)	0.86	-----	0.02 *	0.06 *
	ϕ	0.84		-----	0.04 *
	RM	0.80			----

* Significant at p<.05.

Table 3

Results of Split Plot Factorial Analysis of Variance for K (Kappa) Across Test Lengths

Test Lengths	Sources	df	Sum of Square	Mean Square	F
15	Between Factor	29	0.14936		
	Sample Size	2	0.05550	0.02775	7.97 *
	Error	27	0.09386	0.00348	
	Within Factor	60	0.32173		
	Method	2	0.24590	0.12295	108.61 **
	Method * Sample Size	4	0.01470	0.003675	3.25 *
	Error	54	0.06113	0.001132	
25	Between Factor	29	0.13645		
	Sample Size	2	0.01558	0.00779	1.74
	Error	27	0.12087	0.00448	
	Within Factor	60	0.15454		
	Method	2	0.11683	0.058415	95.14 **
	Method * Sample Size	4	0.00456	0.001140	1.86
	Error	54	0.03315	0.000614	
35	Between Factor	29	0.06815		
	Sample Size	2	0.00113	0.00565	0.23
	Error	27	0.06702	0.002482	
	Within Factor	60	0.072603		
	Method	2	0.05864	0.02932	131.83 **
	Method * Sample Size	4	0.001413	0.000353	1.59
	Error	54	0.01201	0.0002224	

** Significant at p<.01.
* Significant at p<.05.

The results in Table 4 indicated that effects due to item selection methods are significant at the .15 level across sample sizes (according to Kirk (1982), the significant level of the omnibus test is .05 (method effect) + .05 (sample size effect) + .05 (method by sample size)).

Table 4

Analysis of Variance Table With Tests of Simple Main Effects due to Item Selection Methods in the 15-Item Tests

Source	df	Sum of Square	Mean Difference	F
Method at 75-Sample Size	2	0.13994	0.06997	61.81 *
Method at 150-Sample Size	2	0.07325	0.03663	32.35 *
Method at 300-Sample Size	2	0.04742	0.02371	20.95 *
Error	54	0.06113	0.01132	

* Significant at p <.15.

Scheffé post hoc comparisons were performed to further examine the mean differences of K values among item slection methods in a given sample size. The results in Table 5 revealed that in the 15-item tests, the mean of the K values between any two item selection methods were significantly different from each other at 75- and 150-sample sizes; and only the mean of K values in the phi coefficient approach was significantly different from that in the random selection method at the 300-sample size.

Table 5

Scheffé Post Hoc Comparisons for K (Kappa) Mean Differences Among Item Selection Methods Across Three Methods and Sample Sizes in The 15-Item Tests

Sample Size	Methods	Mean	Mean Differencea		
			ϕ	P(Xc)	RM
75	ϕ	0.59	-----	0.06 *	0.16 *
	P(Xc)	0.53		----	0.10 *
	RM	0.43			----
150	ϕ	0.52	-----	0.06 *	0.12 *
	P(Xc)	0.46		-----	0.06 *
	RM	0.40			-----
300	ϕ	0.51	------	0.05	0.09 *
	P(Xc)	0.46		----	0.04
	RM	0.42			----

* Significant at p<.15.

The results in Table 6 show that significant main effects at the agreement and the phi coefficient approaches but not at the randoom selection method.

Table 6

Analysis of Variance Table With Tests of Simple Meain Effects due to Sample Sizes

Source	df	Sum of Square	Mean Square	F
Sample Size at P(Xc)	2	0.03085	0.01542	8.06 **
Sample Size at ϕ	2	0.03642	0.01821	9.52 **
Sample Size at RM	2	0.00294	0.00147	0.77
Error	81	0.15499	0.001914	

** Significant at p<.15.

Scheffé post hoc comparisons were performed to further examine the mean differences of K values among the sample size. The results of these analyses are presented in Table 7 in which the mean differences of K values are significant between the 75-and 150-sample sizes and between 75- and 300-sample sizes but not significant between the 150- and 300-sample sizes.

Table 7

Scheffé Post Hoc Comparisons for K (Kappa) Mean Differences Among Sample Sizes at the Phi Coefficient and the Agreement Approaches

Methods	Sample Size	Mean	Mean Differences 75	150	300
ϕ	75	0.59	-----	0.07 *	0.08 *
	150	0.52		----	0.01 *
	300	0.51			----
p(xc)	75	0.53	-----	0.07 *	0.07 *
	150	0.46		----	-----
	300	0.46			-----

*Significant at p<.15.

The graphic illustrations of Figure 1 and of Figure 2 provide the information about the interaction of method-by-sample size. These two figures illustrate that as sample size increase, the mean difference of K values produced by these three item selection methods tend to decrease.

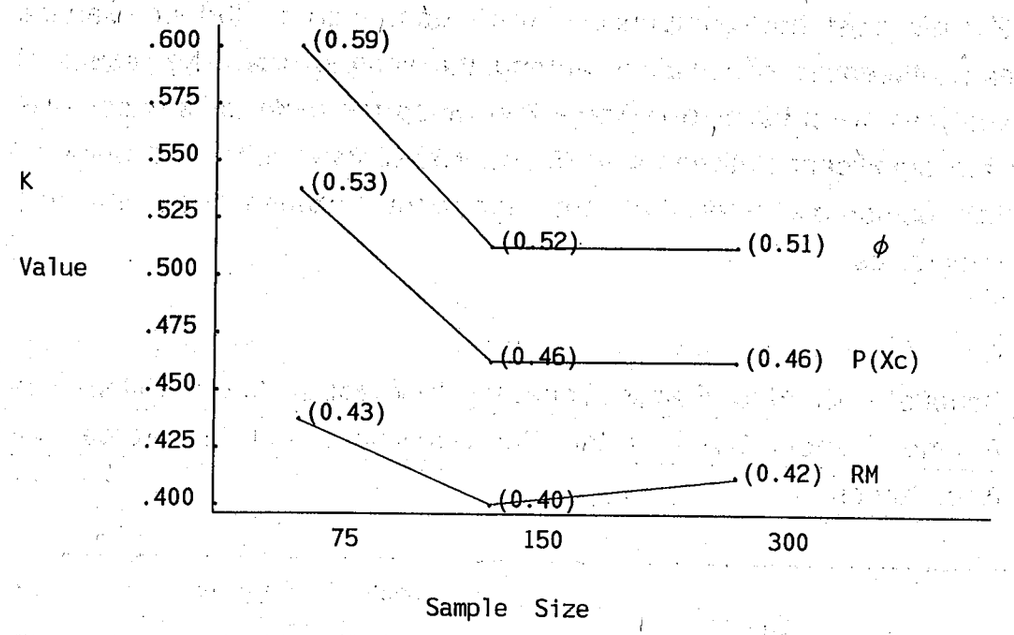

Figure 1. Illustration of the relationship between item selection methods and sample sizes in the 15-item tests.

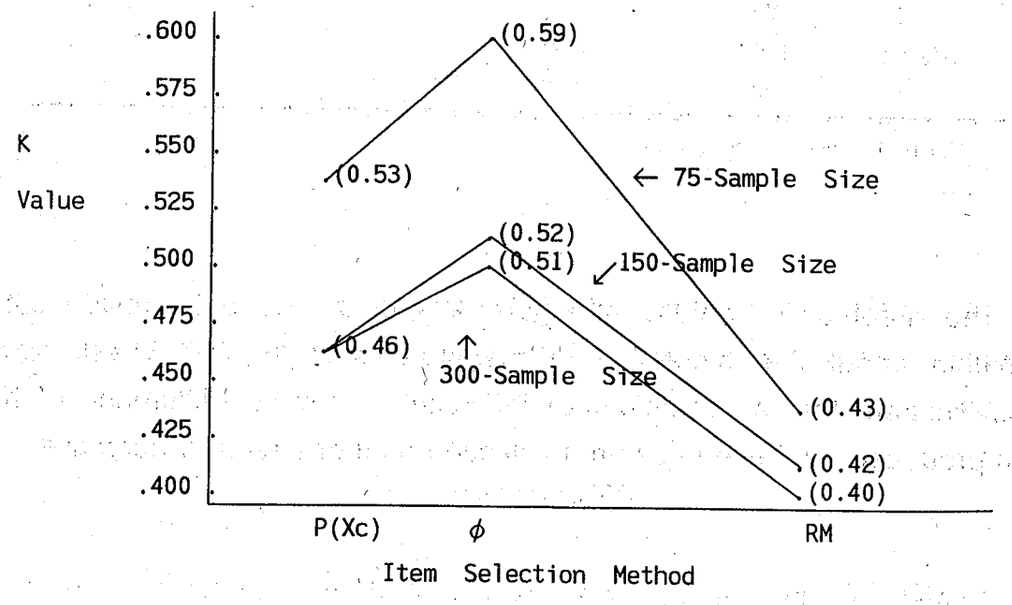

Figure 2. Illustration of the relationship between sample sizes and item selection methods in the 15-item tests.

Because there were significant effects due to item selection methods across test lengths, Scheffé post hoc comparisons were performed to examine the mean difference among item selection methods. As shown on Table 8, the phi coefficient approach produced tests with the highest K values, and the random selection method produced tests with the lowest.

Table 8

Scheffé Post Hoc Comparisons for K Mean Differences Among Item Selection Methods Across Three Test Lengths

Test Lengths	Methods	Mean	Mean Differences ϕ	P(Xc)	RM
15	ϕ	0.54	----	0.06 *	0.12 *
	P(Xc)	0.48		-----	0.06 *
	RM	0.42			----
25	ϕ	0.58	----	0.06 *	0.09 *
	P(Xc)	0.52		-----	0.03 *
	RM	0.49			-----
35	ϕ	0.61	----	0.03 *	0.06 *
	P(Xc)	0.58		----	0.03 *
	RM	0.58			----

* Significant at p<.05.

In addition, the results in Tables 2 and 8 reveal that as test length increased, both Po and K values also increased in these item selection methods.

Discussion

Based on the results of the present study, several issues are discussed below:

The Random Selection Method

Popham (1978), Hambleton (1982) and Hambleton and Gruijter (1983) suggested the use of the random selection method to construct a test since all test items in the criterion-referenced test were developed on the basis of a carefully defined domain of tasks, and all items were homogeneous and interchangeable. The results of the present study show that the Po and K values of the tests based on the random approach were lower than those of the other two item selection methods. Similiar findings were found in other studies (Crehan, 1974; Haladyna & Roid, 1983; and Smith, 1978). The results may be interpreted that the random item selection method is not as efficient as the other two, but more likely it is because items in the item pool are not homogeneous enough nor as carefully selected as is necessary for successful use of random selection. Several other investigations have shown similiar findings (Saupe, 1966; Hsu, 1971; Smith, 1978; Haladyna & Roid, 1981; and van de Linden, 1981).

Po and K Results

In this study, the results of the Po and K analysis were different, in the former, the agreement approach produced the highest coefficients and in the later, the phi approach produced the highest coefficients. An examination of the literature showed that such a differential result had been noted previously (Millman, 1974; Berk, 1984; and Subkoviak, 1980 & 1984). The two estimates of reliability of the mastery/nonmastery decision have been shown to reflect different aspects of measurement. According to Berk (1984), both Po and K are sensitive to the selected cut-off scores. However, for Po, lower values correspond to the cut-off scores near the mean score. K does the reverse, high values are associated with cut-off scores near the middle of the score range and lower values at the outer range. The formula for phi contains the middle scores in n2 and n3 but the agreement approach does not, therefore, the differential Po and K results are probably due to the differences in soultions with respect to the two formulas.

Effects of Sample Sizes

In the present study, samples of 75, 150 and 300 subjects were used, representing 1 1/2, 3 and 6 times of the initial 50 items. Significant effects due to sample size were not found, however, except in the 15-item tests of the K analysis. The fact that sample size had relatively little effect supports research that has shown both Po and K are sensitive to selected cut-off scores, test length, and score variability, but not sample size (Berk, 1984; Subkoviak, 1984).

Conclusion

Based on the findings of this study, the random selection method produced tests with the lowest reliability of mastery/nonmastery classifications, when compared with the agreement and the phi coefficient approaches. Researchers like Hsu (1971), Smith (1978), and van de Linden (1981) stated that empirical item analysis is absolutely necessary for constructing criterion-referenced test items even though items were selected from a well-defined domain of tasks. The results of this study support this statement.

The results of Po and K analyses were inconsistent in recommending one selecton method over another. These results support the statement provided by Berk (1984) and Subkoviak (1980, 1984); that is, Po analysis shows a converse relationship with the K analysis. The results of the Po and K depend on whether an absolute or relative cut-off score is set such as when the tests are for classroom levels or district and state levels (Berk, 1984). For example, when the tests are administered at classroom levels and absolute cut-off scores are used, then Po is appropriate. In such a case, the agreement approach appears to be preferable because it has yielded significantely higher Po values in this study. In other instances associated with relative cut-off scores the phi approach appears to be preferable because of the K results.

Recommendation for Further Research

Both Po and K analyses are sensitive to the selected cut-off score, test length and score variability (Berk, 1984; Subkovbiak, 1984). The present study only dealt with one cut-off score (80%) and one score variability. As a result, the effects due to cut-off scores and/or score variabilities for recommending one method over another are unknown. It is, therefore, recommended that in future studies, more than one cut-off score and different score variabilities be included, thereby allowing the effect of these variables to be better understood.

References

Berk, R. A. (1980a). Item analysis. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art (pp. 49-79). Baltimore, MD: The Johns Hopkins University Press.

Berk, R. A. (1980b). Practical guidelines for determining the length of objective-based criterion-referenced tests. Educational Technology, 20(11), 36-41.

Berk, R. A. (1984). Conducting the items analysis. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 97-143). Baltimore, MD: The Jonhs Hopkins University Press.

Crehan, K. D. (1974). Item analysis for teacher-made mastery tests. Journal of Educational Measurement, 11 (4), 255-262.

Ferguson, G. A. (1981). Statistical analysis in psychology and education (5th edition). New York: McGraw-Hill Book Company.

Haladyna, T. M., & Roid, G. (1981). The role of instructional sensitivity in the empirical review of criterion-referenced items. Journal of Educational Measurement, 18 (1), 39-53.

Haladyna, T. M., & Roid, G. (1983). A comparison of two approaches

tocriterion-referenced test construction. Journal of Educational Measurement, 20(3), 271-282.

Hambleton, R. K. (1974). Testing and decision-making procedures for selected individualized instructional program. Review of Educational Research, 44 (3) , 371-400.

Hambleton, R. K. (1982). Advances in criterion-referenced testing technology. In C. Reynolds, & T. Gutkin (Eds.), Handbook of school psychology (pp. 351-379). New York: John Wiley & Sons.

Hambleton, R. K. (1984). Determining test length. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 144-168). Baltimore, MD: The Johns Hopkins University Press.

Hambleton, R. K., & de Guijter, D. N. M. (1983). Application of item response models to criterion-referenced test item selection. Journal of Educational Measurement, 20 (4), 355-367.

Hambleton, R. K., Mills, C. N., & Simon, R. (1983). Determining the lengths for criterion-referenced tests. Journal of Educational Measurement, 20(1), 27-38.

Harris, D. J. (1983). Item selection for mastery tests: A comparison of three procedures (Doctoral dissertation, University of Wisconson-Madison, 1983). Dissertation Abstracts International , 44, 2741A.

Harris, D. J., & Subkoviak, M. J. (1986). Item analysis: A short-cut statistic for mastery tests. Educational and Psychological Measurement, 46(3), 495-507.

Hsu, T. M. (1971, February). Empirical data on criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New York, NY. (ERIC Document Reproduction Service No. ED 050 139).

Huynh, H. (1976). On the reliability of decision in domain- referenced testing. Journal of Educational Measurement , 13(4), 253-364.

Kirk, R. E. (1982). Experimental design. Belmont, CA: Brooks/Cole.

Mellenbergh, G. J., & van de Linden, W. J. (1982). Selecting items

forcriterion-rerferenced tests. In B. H. Choppin, & T. N. Postlethwaite(Eds.), Evaluation in education: International review series (pp. 177-190). Elmsford, NY: Pergamon Pres Ltd.

Millman, J. (1974). Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current application (pp. 311-397). Berkeley, CA: MuCutchan.

Popham, W. J. (1978). Criterion-referenced measurement . Englewood Cliffs, NJ: Prentice-Hall Inc.

Saupe, J. L. (1966). Selecting items to measure change. Journal of Educational Measurement, 3(3), 223-228.

Smith, D. U. (1978, March). The effects of various item selection methods on the classification accuracy and classification of criterion-referenced instruments. Paper presented at the annual meeting of the American Education Research Association, Tornoto, Ontario, Canada. (ERIC Document Reproduction Service No. ED 159 222).

Subkoviak, M. J. (1980). Decision-consistency approaches. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art (pp. 129-185). Baltimore, MD: The Johns Hopkins Unversity Press.

Subkoviak, M. J. (1984). Estimating the reliability of mastery-nonmastery classifications. In R. A. Berk (Ed.), A guide to criterion-referenced test construction (pp. 267-291). Baltimore, MD: The Johns Hopkins University Press.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 11 (4), 263-267.

Swezey, R. W. (1981). Individual performance assessment : An approach to criterion-referenced test development . Reston, VA: Reston Publishing Co., Inc.

van de Linden, W. J. (1981). A latent trait look at pretest-posttest validation of criterion-referenced test items. Reviewof Educational Research, 51(3), 379-402.

自然與應用科學

Natural & Applied Science