ED 286 912 TM 870 544

AUTHOR Cahan, Sorel; Cohen, Nora

TITLE Improving the Validity of Classification Decisions on the Basis of Competency Test Performance by Means of Self Correction.

PUB DATE 30 Mar 87

NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).

PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Classification; *Cutting Scores; Foreign Countries; *Mastery Tests; Primary Education; Testing Problems; *Test Reliability; *Test Validity

IDENTIFIERS *Answer Changing (Tests); Israel; *Self Correction; Test Retest Reliability; Type I Errors; Type II Errors

ABSTRACT

Two types of classification error are possible in competency tests: erroneous classification of an individual as a "master" of the subject (Type II error), and erroneous classification of a master as a "nonmaster" of the subject (Type I). If steps are taken to minimize Type II errors, an artificially high number of true masters will be classified as nonmasters. The remedy for this problem is to empirically correct for incorrect answers caused by irrelevant factors (fatigue, lack of motivation, etc.). This is done by self-correction: administering the test, checking it, and resubmitting the incorrectly answered questions to the student. The examinee is allowed to correct those incorrect answers. The self-correction technique was used in testing some 4,000 third-grade students in 80 elementary schools in Israel. Forty-one open-ended mathematics questions were answered and corrected. Seventy percent of students with incorrect answers provided correct answers when retested. Based on this finding, the probability of misclassification of true masters as nonmasters (Type I error) was calculated to be 40%. Further analysis showed self-correction to increase the test's internal consistency and its replicability. Retesting two weeks later provided nearly identical results. (MGD)

IMPROVING THE VALIDITY OF CLASSIFICATION DECISIONS ON THE BASIS OF COMPETENCY TEST PERFORMANCE BY MEANS OF SELF CORRECTION

Sorel Cahan

The Hebrew University of Jerusalem

Nora Cohen

Chief Scientist's Office

The Israeli Ministry of Education

March 30, 1987

Paper presented at the 1987 Annual Meeting of the American Educational Research Association, Washington, D.C.

Address: Sorel Cahan, School of Education, The Hebrew University, Mount
Scopus, Jerusalem 91905 ISRAEL

Performance on competency tests is usually the basis for the dichotomous classification of individuals as "masters" and "nonmasters". Associated with this classification procedure are two types of "Platonic" errors: (a) Type I error: the misidentification of a true master as a nonmaster and (b) Type II error: the misidentification of a true nonmaster as a master. In order for the classification procedure to be valid, the probabilities of these errors should be minimized. However, these probabilities are not equally manipulable by the tester. Thus, while testers can lower the probability of a correct answer being given to an item when the student doesn't know this item (e.g., by using open questions, which minimize the posibility of guessing, and by secluding the student from the outside world), their ability to manipulate the factors which may cause a student who knows the answer to an item to give a mistaken one (due to lack of motivation, fatigue, etc.) is much more restricted.

Therefore, unlike the psychometric error in normative testing, the Platonic error in competency testing needs not be unbiased. In fact, if the appropriate steps for the minimization of Type II errors are taken, the errors will be mainly of Type I, i.e., misidentification of true masters as nonmasters. As a result, the proportion of nonmasters identified by the testing procedure will be artifactually high.

One way of overcoming this difficulty is to "correct" for measurement error by lowering the standard of competence. The rationale underlying this solution is best understood in the context of a state-model conceptualization of mastery, where the true-score standard is set at 100 percent. After a consideration of measurement errors, observed-score standards are often set at values less than 100 percent, for example 80% or 70%, thus allowing for 20% or 30% incorrect answers, supposedly unrepresentative of the student's true knowledge. Glass referred to this approach as "counting backwards from 100%".

Even though it has been frequently adopted by various competency testing programs, counting backwards from the true score standard has two serious shortcomings. First, in order for testers to be able to determine the percentage of "acceptable" incorrect answers, they have to know the conditional probability of an incorrect answer being given to an item when the student knows the particular item. Second, they must assume that this probability is invariant across both students and items.

The unpleasant reality is that both conditions are not fulfilled. Testers do not know the conditional probability of students failing an item when they know it. Moreover, they have no good reason to assume that this probability does not vary among items, for a particular student, and among students, with respect to a particular item. On the contrary, there are very good reasons to assume the existence of substantial variations, both within and between students. Thus, the determination of a particular percentage of acceptable incorrect answers is necessarily arbitrary and devoid of any empirical justification. Most probably this percentage will be too large for some students, thus leading to their misidentification as masters, and too small for other students, thus leading to their misidentification as nonmasters.

Moreover, the 20% or 30% acceptable incorrect answers are not likely to consist of a random sample of test items. Most probably the failed items will be the most difficult ones. As a result, testers will never know whether failure on these items was caused by irrelevant factors (as they assume) or by the student's lack of knowkedge.

We suggest a different approach to the measurement error problem in competency testing, namely the empirical correction for unrepresentative incorrect answers. It consists of administering the test, checking it, and administering it again. This time students are told to answer only the items they failed on the first administration of the test.

According to the argument advocated in this paper, the self corrected test-behavior would be more representative of the student's true knowledge. The rationale underlying this argument is as follows: By using open ended questions and by secluding the examinee, the tester can significantly minimize the probability of Type II error. Consequently, successful performance would be a safe enough basis for valid inferences concerning the existence of the relevant knowledge. Failure on the test, on the other hand, may be multiply caused and, therefore, has a much weaker diagnostic significance. Specifically, it needs not reflect lack of the relevant knowledge. Rather, it may be due to lack of motivation, fatigue, etc. Some of these causes may still be operating on an additional administration of the failed items. However, other causes may be administration specific. Therefore, assuming that the probability of Type II error on the second trial is still low, self correction of all or part of the previously failed items will necessarily reflect existing knowledge. Hence, the self corrected test score would be a more valid basis for classification decisions concerning the examinee's true mastery level. Note that unlike the artificial 70% or 80% standard, the proportion of self corrected items is not constant for all students.

In this paper we present the results of an empirical test of the self correction paradigm, based on a sample of about 4,000 3rd grade students in a representative sample of 80 elementary schools in Israel. The students were administered a 41-item minimum competency math test, developed according to the specifications of the Ministry of Education. In order to eliminate guessing, all the items were presented in an open-ended form. Efforts were made also to minimize the impact of other potential sources of Type-II error.

The tests were corrected on the same day and for each student the unattempted or failed items were marked on a new and identical test form, together with the student's name. On the following day, students were presented with the new test forms and instructed to solve only the marked items.

The effect of self correction was dramatic. It is best illustrated by the fact that 70% of the students given this oportunity (1/3 of the entire sample) achieved the maximal score (see Table 1). Assuming a state model of competence and the associated 100% standard, these students would have been classified as nonmasters on the basis of their uncorrected test behavior. According to our argument, most of them are true masters. Therefore, their classification as nonmasters on the basis of the uncorrected test behavior would have been mistaken.

If this argument is correct, then the probability of misclassification of true masters as nonmasters (i.e., the probability of Type-I errors) can be estimated on the basis of the crosstabulation of the mastery/nonmastery decisions before and after self correction. This is presented in Table 1.

Table 1 about here

The probability of Type-I error is estimated to be 0.40. That is, 40% of the true masters would have been misclassified as nonmasters on the basis of their uncorrected test behavior, resulting in an artifactually high percentage of nonmasters: 49% instead of 15%. Thus, the self correction procedure allows for the estimation of the probability of Type-I errors as well as for their empirical correction. Obviously, the validity of both estimate and correction depends on the validity of the basic argument.

Unfortunately, due to the Platonic nature of the true scores involved, this is not an empirical issue. However, some of the predictions which follow from this argument can be empirically tested. In the remainder I shall report empirical results relevant to two such predictions.

The first one concerns the internal consistency of the test. If the self corrected answers are more valid measures of the underlying knowledge than the uncorrected ones, than the internal consistency of the test should be increased by self correction. The reliability analyses performed on the corrected and uncorrected test responses support this prediction. The results of these analyses are presentd in Table 2.

Table 2 about here

The second prediction concerns the greater replicability of the mastery/nonmastery decisions based on the self corrected test behavior. In order to test this prediction, the same test was readministered to the entire sample two weeks after the first administration, and the self correction procedure outlined above was repeated. Table 3 presents the crosstabulation of the mastery/nonmastery decisions on the two administrations, on the basis of the uncorrected (A) and corrected (B) answers.

Table 3 about here

The percentages of nonmasters on the second administration, both before and after self correction, were almost identical to those found on the first administration. This facilitates the comparison between the uncorrected and the corrected test behavior in terms of the replicability of the mastery/nonmastery decisions.

The results of this comparison point to a considerable increase in the overall agreement between the mastery/nonmastery decisions on the two administrations following self correction: from 66% (Table 3A) to 86% (Table 3B). This increase exceeds the one expected only on the basis of the more extreme marginal distribution of the mastery decisions following self correction. This is reflected in an increase in the corresponding Kappa values from .32 before self correction to .39 following it. We interpret this increase as evidence for a net effect of the self correction procedure on overall agreement.

The empirical results concerning the effect of self correction on the internal consistency of the test and on the replicability of the mastery/nonmastery decisions support our agreement concerning the higher validity of the self corrected test behavior. In principle, the effect of this procedure is similar to the effect of the "counting backward from 100%" approach. Indeed, the establishment of an observed score standard lower than the true score one also increases the validity of the mastery/nonmastery decisions. However, two advantages of the self correction procedure are worth mentioning:

1) First, it does not involve the application of a single "correction factor" to all examinees.

2) Second, it does not involve arbitrary decisions. Therefore, it is not likely to vary between testers.

The gain associated with the application of the self correction procedure will be higher, the higher the probability of Type-I errors and the higher the proportion of true masters in the population. Therefore, other things being equal, it is likely to be higher the lower the grade level and the easier the domain to be mastered.

In any case, the validity of the self correction procedure depends on the effective minimization of Type-II errors. Prevention of guessing stands out as a necessary requirement in this respect. Therefore, effective application of the self correction procedure seems to imply open ended, instead of the usual multiple choice items.

A final remark should refer to the relevance of the self correction procedure to continuum conceptualizations of mastery. However, this is a complex issue, which has not yet been fully explored. Your comments and suggestions concerning it will be particularly welcome.

Table 1: The percentage of mastery and nonmastery decisions before and after self correction, assuming a 100% standard.

		Decision after self correction		
		Nonmaster	Master	
Decision before self correction	Nonmaster	14.6	33.8	48.4
	Master	-	51.6	51.6
		14.6	85.4	100.0

Table 2: The internal consistency of the test before and after self correction.

Statistic	Before self correction	After self correction
Mean inter item correlations	.22	.28
Cronbach's Alpha	.92	.93
Standardized Alpha	.92	.94

Table 3: The stability of the mastery/nonmastery decisions before and after self correction.

A. Before self correction

		Decision on the 2nd administration		
		Nonmaster	Master	
Decision on the 1st administration	Nonmaster	30.6*	17.8	48.4
	Master	16.3	35.3*	51.6
		46.9	53.1	100.0

* Indicates agreement.

B. After self correction

		Decision on the 2nd administration		
		Nonmaster	Master	
Decision on the 1st administration	Nonmaster	6.3*	8.3	14.6
	Master	5.7	79.7*	85.4
		12.0	88.0	100.0

* Indicates agreement.