The influence of examinee erring on expected item scores in achievement testing

PIETER KOELE and WULFERT P. VAN DEN BRINK
University of Amsterdam, Amsterdam, The Netherlands

In achievement testing, examinees usually have to mark off the selected answers to multiple-choice questions on an answer sheet. With the use of the item-sampling model of criterion-referenced measurement, the influence of erring in marking off on expected item scores is investigated. It is found that, in general, this type of erring lowers expected item scores and that the magnitude of this decrease is substantial enough to deserve some attention.

When taking multiple-choice tests for achievement assessment, the examinee usually has to mark off the chosen answers to the items on a separate answer sheet, thus enabling quick and efficient digestion by computer of large numbers of exams. This marking off is a potential source of error. That is to say, there is a distinct possibility that the examinee errs when marking off the chosen answer on the answer sheet. It is the purpose of this article to investigate the consequences of this type of examinee erring, especially the influence on expected item scores. In the state models of criterion-referenced measurement (Macready & Dayton, 1977, 1978), there appears a parameter that is related to examinee erring. However, state models have a limited practical value, and the available estimates of the relevant parameter are not based on multiple-choice exams. For that reason, in the present study, we shall use the item sampling model of criterion-referenced measurement, corrected for guessing (see Van den Brink & Koele, 1980; Van den Brink, 1982). An illustration, based on empirical multiple-choice test scores, will conclude our discussion.

A MODEL FOR ERRING

In the item-sampling model, a test is conceived as a random sample of n items from a large item universe. The ability τ_i is defined as the expected proportion of items in the universe that a fixed examinee i knows because she or he has mastered the material. On the remaining items, the examinee has to guess. The expected proportion of the nonmastered items in the universe that the examinee answers correctly because of guessing is denoted by γ_i. For the fixed examinee i, the probability of making an error in marking off is ϵ_i. We assume erring to be a random process; that is to say, if an examinee errs, the probability of marking off any answer other than the intended one is distributed uniformly over the remaining answers.

Thus, if an item has a answer possibilities and if the examinee has originally selected a wrong answer, erring may result either in marking off the correct answer with probability $1/(a-1)$ or in marking off a wrong answer with probability $(a-2)/(a-1)$.

For examinee i, the probability π_i of selecting the correct answer on any item in the test, either by knowing or by guessing, is given by

$$\pi_i = \tau_i + (1-\tau_i)\,\gamma_i . \tag{1}$$

Taking erring into account, let us now consider the process that takes place when the examinee answers an item. Two stages must be distinguished. In the first stage, an answer is selected, either by knowing or by guessing. In the second stage, the selected answer is marked off on the answer sheet, and the examinee may or may not err in doing so. If the items are binary scored, there are two simple events in this chance process that have as outcome an item score 1: The examinee selects the correct answer and does not err in marking off, or the examinee selects a wrong answer, errs in marking off, and marks off the correct answer. The expected item score equals the sum of the two corresponding probabilities:

$$E[I_i] = \pi_i\,(1-\epsilon_i) + (1-\pi_i)\,\epsilon_i\frac{1}{a-1}. \tag{2}$$

Substitution of Equation 1 into Equation 2 yields an expression that directly gives expected item score as a function of mastering, guessing, and erring:

$$E[I_i] = \tau_i\,(1-\epsilon_i) + (1-\tau_i)\left[\gamma_i - \frac{\epsilon_i(a\gamma_i-1)}{a-1}\right]. \tag{3}$$

DISCUSSION

When no erring occurs, that is, when $\epsilon_i = 0$, the expected item score equals π_i. It is easily deduced that erring lowers expected item scores, provided that π_i is larger

than $1/a$. Only when π_i is smaller than $1/a$ (a rather exceptional situation, we daresay), erring is "beneficial" for the examinee.

As far as the magnitude of ϵ is concerned, no direct empirical data are available. Van den Brink and Koele (1984) experimentally manipulated one possible source of erring in marking off by comparing item scores on two versions of 30 test items. In one version of the test, the three answer possibilities of the items were indicated by $a.$, $b.$, and $c.$ and in the other version, by 1., 2., and 3. On the answer sheet, 1., 2., and 3. were used for both versions. In all other respects, the two versions of the test were identical. A group of 114 examinees was randomly divided over the two versions. The results showed that the average item score on the 30 $a.$ $b.$ $c.$ items was .03 lower than the average item score on the 1. 2. 3. items, a difference that was significant at the $\alpha = .05$ level [t(29) = 1.88]. Of course, this finding only reflects the effect of a manipulation on erring, but it illustrates that examinees do make mistakes in filling in answer sheets. There are more indications for this: Students frequently err when marking off their registration number, and they sometimes mark off on the sheet answer possibilities that do not exist on the test.

All this supports the idea that there is a rather substantial effect of erring, and we think that ϵ may well vary somewhere between .01 and .05. This means that erring has a distinct influence on test scores. For instance, consider a test with three-choice items and an examinee with $\tau_i = .50$, $\gamma_i = .40$, and $\epsilon_i = .05$. The expected item score without correction for erring (Equation 1) equals .70; corrected for erring, Equation 3 yields an expected item score of .67. Although this is a small difference, it should be realized that, for a test consisting of 30 items, this difference corresponds to a difference in expected test scores of almost one point. Such a difference may decide between passing or failing of an examinee.

In general, erring has a negative influence on expected item scores. The magnitude of this influence is probably not very large, yet may be substantial enough to justify some attention.

REFERENCES

MACREADY, G. B., & DAYTON, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, **2**, 99-120.

MACREADY, G. B., & DAYTON, C. M. (1980). The nature and use of state mastery models. *Applied Psychological Measurement*, **4**, 493-516.

VAN DEN BRINK, W. P. (1982). Binomial test models for domain referenced testing. *Evaluation in Education*, **5**, 165, 176.

VAN DEN BRINK, W. P., & KOELE, P. (1980). Item sampling, guessing and decision making in achievement testing. *British Journal of Mathematical and Statistical Psychology*, **33**, 104-108.

VAN DEN BRINK, W. P., & KOELE, P. (1984). *Het vergis-effect*. Universiteit van Amsterdam, Psychologisch Laboratorium.