Bulletin of the Psychonomic Society
1987, 25 (4), 280-282

Scoring rules and probability testing

PIETER KOELE, ROBERT DE BOO, and PAUL VERSCHURE
University of Amsterdam, Amsterdam, The Netherlands

When using multiple-choice items in educational testing, the examinee is usually forced to choose exactly one of the answer possibilities as the supposedly correct answer. An alternative response mode is to let examinees assign probabilities to each of the answers of the item, corresponding to their belief in the correctness of each answer. In this study student attitudes toward this response mode were investigated, and scoring rules for assessing item scores were compared. It appeared that students had no clear preference for either using subjective probabilities or making forced choices, and that from a psychometrical point of view scoring rules based only on the probability assigned to the correct answer have to be preferred to scoring rules using all probabilities assigned to each of the answers.

In any form of psychological or educational testing, a distinction can be made between response mode and scoring rule. The response mode is the way in which the subject is required to respond to the stimulus material. The scoring rule prescribes how numerical values are assigned to the responses. When multiple-choice examinations are used to measure a specific ability, the common response mode is the forced choice; the examinee is requested to mark precisely one of the answer possibilities as the (supposedly) correct one. The usual scoring rule assigns an item score of one to the examinee if the correct answer has been marked and an item score of zero if not. The test score is then taken to be the sum of the item scores.

The combination of this forced-choice response mode and dichotomous scoring rule has the advantage of a quick and efficient computerized processing of large numbers of examinations. On the other hand, there are two disadvantages. The forced-choice response mode introduces error variance into the test scores because of guessing, and the dichotomous scoring rule does not discriminate between different levels of partial knowledge. Both of these factors are a potential threat to the reliability of the test.

One of the possible ways of overcoming these disadvantages is to use personal probabilities as a response mode to multiple-choice items. In its purest form, this response mode requires assigning a personal probability to each of the answers of a test item, according to the degree to which the answer is believed to be the correct one. Per item, these probabilities should add to unity. De Finetti (1965) gave an extensive overview of the different varieties of the response mode based on personal probabilities.

The personal probability response mode requires a specific scoring rule, a function assigning a real number to an item based on the distribution of the personal probabilities over the answers of the item. Van Naerssen (1961), de Finetti (1965), and Shuford, Albert, and Massengill (1966) developed a number of those scoring rules. Shuford et al. also introduced the reproducing property as an important aspect of a scoring rule. A scoring rule is reproducing if an examinee cannot maximize his/her score by assigning other probabilities to the answers than the probabilities corresponding with his/her true beliefs. Or more loosely speaking, reproducing scoring rules exclude the possibility that an examinee earns a higher test score than he/she deserves, because clever response strategies do not exist.

Stanley and Wang (1970) and Echternacht (1972) presented critical summaries of the empirical findings in this area of so-called *probability testing*. Their conclusion was that, in general, most studies report small increases in reliability when comparing probability tests with conventional forced-choice tests. However, these small increases hardly seem to warrant the extra amount of time and labor it takes to process probability tests. Hakstian and Kansup (1975; Kansup & Hakstian, 1975), in comparing several scoring rules for probability tests with a conventional forced-choice test, arrived at the same conclusion.

The present study was primarily undertaken to investigate the use of the so-called Brier scoring rule. This scoring rule is well known in the context of research on calibrating subjective probabilities (Lichtenstein, Fischhoff, & Phillips, 1982). The Brier rule is a reproducing scoring rule, and it uses the probabilities assigned to each of the k answers of the test item. According to this rule, the item score S_B is given by

$$S_B = \sum_{i=1}^{k} (p_i - c_i)^2, \qquad (1)$$

where p_i is the subjective probability assigned to answer i of the item, and c_i is the true probability of answer i; c_i is equal to unity for the correct answer and equal to 0 for all other answers. As a consequence, S_B has the maximum value of 2 if a subjective probability of unity is assigned to a wrong answer, and a minimum value of 0 if

Address correspondence to: Pieter Koele, Faculty of Psychology, University of Amsterdam, Weesperplein 8, 1018 XA Amsterdam, The Netherlands.

a subjective probability of unity has been assigned to the correct answer. In order to have a more convenient scale, a simple linear transformation is carried out:

$$S'_B = 1 - S_B/2. \quad (2)$$

After some algebra, we find

$$S'_B = p_c + (1 - \sum_{i=1}^{k} p_i^2)/2, \quad (3)$$

where the p_is are defined as above and p_c is the probability assigned to the correct answer. S'_B has a minimum of 0 when p_c is 0, and a maximum of unity when p_c is unity. Equation 3 shows that the Brier rule can be seen as the sum of two components. The first term is the probability assigned to the correct answer, and the second term is inversely related to the variance of the assigned probabilities. For a fixed value of p_c, the value of the item score S'_B increases as the variance of the assigned probabilites decreases. To investigate the influence of this second term on the reliability of the test, we compare the Brier rule with the following scoring rule:

$$S_L = p_c. \quad (4)$$

This scoring rule consists of the first component of the Brier rule, and it is a simple linear function of the probability assigned to the correct answer. It does not possess the reproducing property. An examinee can maximize his/her expected item score by assigning a probability of 1 to the answer he/she feels most confident about, no matter what his/her "true" probability might be.

The second aim of this study was to investigate examinee opinions about probability testing. Responding with subjective probabilities is cognitively more demanding than marking an answer possibility, and it is not certain that examinees appreciate the alleged advantages of probability testing.

To compare a conventional forced-choice test with a probability test, we had a large sample of subjects take the same multiple-choice test twice, once as a forced-choice test and once as a probability test. Content of the test was a general-knowledge topic, and the test result had no consequences for the subjects. After completion of the test, the subjects were requested to respond to a short questionnaire on their opinions about the use of subjective probabilities in educational testing.

METHOD

Subjects

Subjects were 125 first-year psychology students of the University of Amsterdam, participating in the experiment to fulfill part of a course requirement.

Procedure

Subjects were randomly assigned to one of two groups. Both groups received the same multiple-choice test twice, once as a forced-choice test and once as a probability test. To control for order effects, the order in which the two versions were administered differed in the two groups. In the instruction to the probability test, the use of personal probabilities was explained, and some illustrations were given. Subjects were requested to express their personal probabilities as percentages (i.e., natural numbers between 0 and 100, limits inclusive). Per item, these numbers should add to 100. In front of each answer of an item, a block was provided for writing down the personal probability. The forced-choice version was administered in the conventional way.

After completion of the second version of the test, subjects were asked to fill out a short questionnaire about their attitudes toward probability testing.

Materials

The test consisted of 50 four-choice items. The items asked for the meaning of rarely used or obsolete Dutch words of Germanic, Greek, or Latin origin.

The questionnaire consisted of five short questions in which the subjects could express their opinions about various aspects of probability testing.

Analyses

Test scores were taken to be the sum of the item scores. Cronbach's alpha was calculated for test scores on the forced-choice version of the test using the dichotomous scoring rule, and for test scores on the probability version using the Brier and the linear scoring rules. In addition, intercorrelations among the three test scores were calculated. Responses on the questionnaire were classified and tabulated.

RESULTS

It appears that the order in which the two versions of the test were administered had no effect on any descriptive statistic. The two groups are therefore combined, and the following results are based on one sample of 125 subjects.

The values of Cronbach's alpha for the scores based on the dichotomous, the Brier, and the linear scoring rules were .68, .73, and .78, respectively. The Pearson correlation between dichotomous and Brier scoring rules was .80; between dichotomous and linear scoring rules was .88, and between Brier and linear scoring rules was .80.

The test proved to be rather difficult. On the forced-choice version, the mean number of correct answers was 24.82 out of 50 (SD = 5.64). The mean probability assigned to the correct answer was .52 (SD = .10).

On the first question of the questionnaire, subjects had to indicate whether they preferred probability testing or forced-choice testing. Probability testing was preferred by 38% of the subjects, forced-choice testing by 34%, and 28% of the subjects were indifferent. The second question asked whether probability testing was more demanding than making forced choices. An affirmative answer was given by 52%, a negative answer by 44%, and 4% of the subjects were indifferent. The third question asked whether probability testing was fairer than making forced choices. Sixty-nine percent of the subjects answered yes, 25% answered no, and 6% were indifferent. The fourth question inquired whether subjects would like to take probability tests in real educational settings. Forty-eight percent of the subjects answered yes, 43% answered no, and 9% were indifferent. On the last question, subjects could list as many advantages and disadvantages of probability testing as they could perceive. Frequently mentioned advantages were that it is fairer because it gives a better representation of partial knowledge (37% of the

subjects) and that there is less probability of receiving a 0-item score (28%). Frequently mentioned disadvantages were that there is more hesitation in responding (18%), it is arithmetically cumbersome (18%), a person cannot be lucky (12%), and it takes more time (6%).

DISCUSSION

The results of the questionnaire do not indicate an overwhelming student enthusiasm for probability testing. Although its major advantage, giving credit for partial knowledge, is frequently recognized, the extra amount of time and trouble it takes to respond with subjective probabilities is evidently a drawback. Of course, the answers to the questionnaire may very well reflect a rather general indifference to any form of educational testing. It is obvious that before probability testing is introduced, an extensive campaign explaining the ins and outs of subjective probabilities and scoring rules has to be carried out.

The general finding that probability tests are somewhat more reliable than their forced-choice equivalents has been replicated in this study. It is striking that the greatest gain in reliability is reached using the nonreproducing, linear scoring rule. This finding is in line with results of Michael (1968) and Rippey (1970). It means that the extra information the Brier scoring rule supplies about the way in which probabilities are assigned to all answers of the item introduces variance into the test scores not related to the ability level of the examinee. This argument can be illustrated as follows.

Consider two examinees, E and F, both assigning a probability of .70 to the correct answer of a four-choice item. Examinee E furthermore assigns a probability of .30 to a wrong answer and 0 probabilities to the two remaining answers. Examinee F assigns probabilities of .10 to each of the wrong answers. Examinee E receives a Brier item score $S'_B = .91$, Examinee F has $S'_B = .94$. One might argue that this is fair, because E is more confident about a wrong answer than F, thus exhibiting less knowledge. On the other hand, E has excluded two wrong answers, whereas F gives all wrong answers a nonzero probability; from this point of view, F displays less knowledge than E. Evidently, given the probability assigned to the correct answer, the information about the probabilities assigned to the wrong answers is equivocal. What matters psychometrically is the subjective probability of the correct answer; the rest is error.

REFERENCES

DE FINETTI, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical & Statistical Psychology*, **13**, 87-123.

ECHTERNACHT, G. J. (1972). The use of confidence testing in objective tests. *Review of Educational Research*, **42**, 217-236.

HAKSTIAN, A. R., & KANSUP, W. (1985). A comparison of several methods of assessing partial knowledge in multiple-choice tests: II. Testing procedures. *Journal of Educational Measurement*, **12**, 231-239.

KANSUP, W., & HAKSTIAN, A. R. (1975). A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring procedures. *Journal of Educational Measurement*, **12**, 219-230.

LICHTENSTEIN, S., FISCHHOFF, B., & PHILLIPS, L. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tverski (Eds.), *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

MICHAEL, J. C. (1968). The reliability of a multiple choice examination under various test-taking instructions. *Journal of Educational Measurement*, **5**, 307-314.

RIPPEY, R. M. (1970). A comparison of five different scoring functions for confidence tests. *Journal of Educational Measurement*, **7**, 165-170.

SHUFORD, E. H., ALBERT, A., & MASSENGILL, H. E. (1966). Admissible probability measurement procedures. *Psychometrika*, **31**, 125-145.

STANLEY, J. C., & WANG, M. D. (1970). Weighting test items and test item options, an overview of the analytic and empirical literature. *Educational & Psychological Measurement*, **30**, 21-35.

VAN NAERSSEN, R. F. (1961). A scale for the measurement of subjective probability. *Acta Psychologica*, **19**, 159-166.

(Manuscript received for publication February 16, 1987.)