Behavior Research Methods, Instruments, & Computers
1986, 18 (1), 59-60

A computer program for administering and scoring confidence tests

ROBERT M. RIPPEY
University of Connecticut Health Center
Farmington, Connecticut

In confidence-test procedures, subjects are asked to respond to all the options of multiple-choice items with weights. They may assign the entire weight to a single option or they may distribute their confidence. With Shuford's truncated logarithmic scoring function (Shuford, Albert, & Massengill, 1966), incentives are provided for honest reporting of confidence, and guessing is discouraged. By utilizing adjustments based on least squares fitting of the subject response ($\bar{r}$) to the performance-based likelihood vector ($\bar{p}$), a score of realism can be computed. The realism score indicates whether or not the subject was appropriately certain or cautious (Brown & Shuford, 1973). Adjustment of the knowledge score for inappropriate realism improves reliability and validity (Rippey & Voytovich, 1983). In addition, feedback from the realism score can lead to improvements in realism or suggest deficits in basic forecasting skills and knowledge (Yates, 1982). The logarithmic function has identified occasional bizarre behavior of respondents, suggestive of a pathology of reasoning (Rippey & Voytovich, 1985). The logarithmic method of scoring is used rather widely in England in a course in risk analysis offered by the Open University (1980). Because of its amenability to the scoring of traditional multiple-choice test items, the logarithmic function has been applied to studies of cognitive achievement. Other scoring functions have also been used in other studies, especially in connection with forecasting (Blattenberger & Lad, 1985; Yates, 1982). The Brier score, for example, has been widely studied because of its partitionability using components of variance methods.

Confidence testing initially suffered from problems of administration and scoring (Ebel, 1968). Many of the problems were solved by the use of mainframe computer scoring (Rippey & Donato, 1978). A PLATO version of confidence testing was developed by the Rand Corporation (Landa, 1976), expanded by Rippey and Smith (1979), and improved by Anderson (1982). However, access to the system is much more immediate with microcomputers; the microcomputer program described here is a complete system for preparing tests and scoring keys, and for administering and scoring the tests, either individually or in batch mode. The individual testing procedures are most appropriate for educational use, whereas the batch mode has been used for research purposes in the areas of clinical reasoning and school learning.

The program is user friendly and menu driven. It utilizes either two- or three-option test items. A manual provides details of the underlying theory and a tutorial that leads the user through the seven program components: (1) scoring key preparation, (2) question preparation, (3) test administration with screen display of items, (4) test administration without item display, (5) test scoring and analysis (brief and extended formats), (6) batch response file writer, and (7) batch scoring system.

The subject is asked to select a payoff, which assigns probability in increments of 0.1 to each option of a question. Special scoring functions having the reproducing property have been shown to maximize a student's score if, and only if, the probabilities assigned are equal to the conditional likelihoods of being correct when each probability is used. These conditional likelihoods are computed from the student's performance on the test. If a student assigned a probability of 0.3 ten times to selected options on a test and was correct twice, the conditional likelihood of being correct, given an assigned probability of 0.3, would be 0.2. In essence, only a perfect probability assessor can expect the highest possible score.

A Knowledge Improvement (KI) score and a Realism Improvement (RI) score are determined by construction of a regression line relating assigned probability (in increments of 0.1) to actual proportion of correct probability assignments. If the assessment of realism was perfect, responses to which a likelihood of 0.5 was assigned should be correct 50% of the time. Responses for which 0.9 was assigned should be correct 90% of the time.

The scoring function is $S = 50 \operatorname{Log} 3p_k + 76$, where p_k equals the probability assigned to the correct answer and $.01 \leq p \leq .99$. Although this function is not strictly reproducing, it has the reproducing property for values of p between .027 and .973 (Shuford et al., 1966). Truncation to between .01 and .99 results in a negligible deviation. Three scores result. The first, called Overall Improvement (OI), consists of the differences between the maximum possible score and the actual score. OI can be separated into two parts—an RI score (points a student would make by improving realism) and a KI score (overall improvement score adjusted for errors in realism). If a student is accurate in assigning probabilities, the KI and the OI scores will be the same and the RI score will be zero, indicating that the student is accurate in assessing the likelihood that a given option is correct.

The three scores are obtained as follows. On a 25-item three-option test, the user has, at the conclusion of the test, 75 assigned probabilities ranging from zero to one in increments of 0.1. Of these probabilities, 25 are associated with right answers and 50 are associated with wrong answers. From these data, one can compute the

The author's mailing address is: Department of Research in Health Education, University of Connecticut Health Center, Farmington, CT 06032.

proportion of the time each of the 11 probability values was assigned to the correct answer. For example, if a subject assigned a probability of 0.6 to a correct answer 10 times and was correct 4 times, one could then record 10 superimposed points, each having coordinates (r,p) of 0.6 and 0.4. From the complete set of 75 points, a line of best fit can be obtained by a regression of p into r constrained to the point $(\frac{1}{3},\frac{1}{3})$. If the slope of the regression line is less than one, one may conclude that the subject has overvalued information. If it is greater than one, it was undervalued. There is one regression for each student for each test.

Because the reproducing function maximizes item scores when the probability assigned to the correct answer equals the corresponding conditional probability for each item, substituting the regressed estimate of $r,\hat{r}$, for p will increase each item score. The test is thus rescored using a regression estimate of conditional probability in place of the originally assigned probabilities. Prior to rescoring, the regressed estimates of p are renormalized to make certain that they sum to one. The increase in the RI score is due to the reduction of errors of realism in assessing knowledge. The OI score is equal to the initially obtained score subtracted from the maximum possible (99 × number of items). The score improvement that is possible, based on knowledge alone with realism errors removed, then equals OI−RI. This score is called KI.

Input. Upon presentation of an item, the subject responds with an integer in the range 0-66, representing one of all possible distributions of probabilities over three options in increments of 0.1. The payoffs are displayed initially, but not the probabilities. The probability display is deferred until later, because it has been shown experimentally and theoretically that decision making in one's self-interest is jointly dependent upon a knowledge of payoff and subjective belief. The sequence of response may be entered directly from the keyboard or from a floppy disk file.

Output. Output is either printed or screen displayed. It consists of: (1) assigned and selected probabilities; (2) item scores; (3) a total score; (4) a realism score, which indicates how many points could have been gained by more realistic assessment of confidence; (5) a score of knowledge adjusted for discrepancies in realism; (6) a statement identifying a tendency of overvaluing, undervaluing, or correctly valuing one's confidence; and (7) a Pearson correlation coefficient relating the probability assigned to each option to the conditional likelihood of assigning that probability to the correct option.

Computer and Language. This program is written in Applesoft BASIC for any of the Apple II computer series with 64K memory. Either one or two drives may be specified. The program uses certain enhancements contained in Beagle Brothers Pronto-DOS operating system. these are required and included on the disk.

Restrictions. The maximum number of test items is 50, the maximum number of test-item options is three, and the maximum number of students for batch processing is 200.

Availability. Unprotected copies of the system on a 5¼-in. floppy disk, the 60-page manual, and sample questions can be obtained from the author for $10 (to cover reproduction and handling). Persons sending a large (10×13 in.) stamped, self-addressed envelope, and a punched, double-sided, 5¼-in. blank disk may obtain the program and the manual for $5.

REFERENCES

ANDERSON, R. (1982). *Confidence testing on the PLATO system*. Savoy, IL: University of Illinois, Aviation Research Laboratory. (ERIC Document Reproduction Service No. ED 200 164)

BLATTENBERGER, G., & LAD, F. (1985). Separating the Brier score into calibration and refinement components: A graphical exposition. *The American Statistician*, **39**, 26-32.

BROWN, T. A., & SHUFORD, E. H. (1973). *Quantifying uncertainty into numerical probabilities for reporting of intelligence*. Santa Monica, CA: The Rand Corp.

EBEL, R. (1968). Review of "valid confidence testing demonstration." *Journal of Educational Measurement*, **5**, 353-354.

LANDA, S. (1976). CAAPM: Computer aided admissible probability measurement of PLATO IV. Santa Monica, CA: The Rand Corp.

OPEN UNIVERSITY. (1980). *Risk: A second level university course, blocks 1-6*. Milton Keynes, England: Author.

RIPPEY, R. M., & DONATO, J. (1978). Interactive confidence test scoring and interpretation. *Journal of Educational Measurement*, **7**, 165-170.

RIPPEY, R. M., & SMITH, S. (1979). Improving the reliability and validity of confidence scored tests by adjusting for realism. *Evaluation & The Health Professions*, **1**, 100-109.

RIPPEY, R. M., & VOYTOVICH, A. E. (1983). Linking knowledge, realism and diagnostic reasoning by computer-assisted confidence testing. *Journal of Computer-Based Instruction*, **9**, 88-97.

RIPPEY, R. M., & VOYTOVICH, A. E. (1985). Anomalous responses on confidence-scored tests. *Evaluation & The Health Professions*, **8**, 109-120.

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

YATES, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior & Human Performance*, **30**, 132-156.

(Revision accepted for publication February 9, 1986.)