

Estimating the Validity of a Multiple-Choice Test Item Having k Correct Alternatives

Rand R. Wilcox

University of Southern California and University of California, Los Angeles

In various situations, a multiple-choice test item may have more than one correct alternative, and the goal is to determine how many correct alternatives an examinee actually knows. For a randomly sampled examinee, the validity of an item is defined as the prob-

ability of deciding that the examinee knows i correct alternatives, when in fact exactly i correct alternatives are known. This article describes how latent class models can be used to estimate this probability.

Consider a test item having N alternatives, K of which are correct. Most multiple-choice test items have only $K = 1$ correct alternative, but in some situations (e.g., Duncan & Milton, 1978; Gibbons, Olkin, & Sobel, 1979; Schmittlein, 1984), $K > 1$. For example, a person might be asked to choose the K characteristics that most accurately describe a particular product or service, or the alternatives might contain synonyms or equivalent forms of the correct response.

A natural measure of validity of a test item and a particular scoring procedure is the probability of determining how many of the correct alternatives are actually known by a randomly sampled examinee. A strong true-score model was proposed by Schmittlein (1984) for solving this problem, and an alternative scoring procedure was proposed by Gibbons et al. (1979), but there is now considerable evidence that their results may not supply a satisfactory approach to assessing validity. For example, suppose a randomly sampled individual knows i of the K correct responses. Schmittlein assumed that this individual chooses these i responses, and then guesses at random from among the remaining $N - i$ alternatives. It is this knowledge or random guessing assumption that causes difficulty when assessing validity.

There are two aspects to the problem. First, there is considerable evidence that people do not guess at random (e.g., Bliss, 1980; Coombs, Milholland, & Womer, 1956; Cross & Frary, 1977; Wilcox, 1982b, 1982c). Since all models are just approximations of reality, these empirical studies do not necessarily imply that Schmittlein's model is inadequate. Second, various theoretical results indicate that if a random guessing assumption is made, but guessing is not at random, inferences made from a random guessing model can be highly inadequate (e.g., Ashler, 1979; Weitzman, 1970; Wilcox, 1980, 1982c). The problem considered by Wilcox (1983a) is directly relevant to the validity issue examined by Schmittlein

(1984) for the special case $K = 1$, and again, it can be seen that violating the random guessing assumption can seriously affect the resulting estimate of validity. Similar problems can be found in the scoring procedure proposed by Gibbons et al. (1979). Duncan and Milton (1978) addressed the issue of partial information, but the model proposed here deals with partial information in a more explicit fashion.

A New Latent Class Model

For the reasons given above, an alternative approach to assessing validity may be needed. Below is described a model that yields an estimate of validity; the new model contains the model in Wilcox (1983a) as a special case. As is indicated below, there is empirical evidence that the new set of assumptions provide a better approximation of how people respond to test items.

Here, it is assumed that if an individual knows i of the correct alternatives, then these alternatives are chosen. In contrast to Schmittlein (1984), it is also assumed that a particular person may be able to eliminate j incorrect alternatives through partial information, and that once these j alternatives are eliminated, the individual guesses at random from among the $N - i - j$ alternatives remaining. Consistent with Schmittlein, it is assumed that an individual's responses are independent of one another. In contrast to Schmittlein (p. 26), it is not assumed that the correct alternatives are equally difficult.

Two additional assumptions are needed in order to ensure that the proposed latent class model is identifiable. The first is that $N \geq 2K + 1$. The second is that people choose alternatives until they obtain K correct responses. In Schmittlein's (1984) model, it is assumed that individuals choose only K alternatives.

Empirical evidence in support of the above assumptions can be found in Wilcox (1982b, 1982c, 1983b) for the special case $K = 1$. It is noted that Coombs et al. (1956) appear to be the first to have provided empirical evidence that individuals have partial information when responding to test items, and in most cases it appears to be partial information that invalidates the knowledge or random guessing assumptions adopted by Schmittlein (1984). A method of empirically checking the new model is indicated below.

For a randomly sampled individual, let P_{ij} be the probability of obtaining i correct responses among the first K alternatives chosen, and choosing the K th correct response on the j th attempt *after* the first K attempts. For example, if $K = 2$ and $N = 5$, then P_{12} is the probability that a randomly sampled individual had one correct response among the first $K = 2$ attempts, and that the second correct response is chosen on the second attempt after the first two choices have been made. In other words, the individual's response pattern is either (1,0,0,1) or (0,1,0,1), where a 1(0) means a correct (incorrect) alternative was chosen.

For some population of people, let τ_{ij} be the probability that a randomly sampled person knows i of the correct alternatives and can eliminate j incorrect responses through partial information. Then, based on the assumptions made above,

$$P_{I,N-K-J} = \sum_{i=0}^I \sum_{j=0}^J \tau_{ij} \lambda_{ij}(K, N - K - J) / \eta_{ij} \quad (1)$$

where η_{ij} is the total number of response patterns that are possible for a person who knows i items and can eliminate j incorrect alternatives, and $\lambda_{ij}(K, N - K - J)$ is the number of response patterns that this same person can give, where there are I correct responses on the first K attempts, and the K th correct alternative is chosen on the $(N - K - J)$ th attempt *after* the first K attempts. For $I = 0, \dots, K - 1$, the possible values of J are $0, \dots, N - 2K + I$, whereas for $I = K$, the only possible value for J is zero. No distinction is made between individuals who know all K correct alternatives and those who can eliminate

all $N - K$ incorrect choices. Thus, the parameter τ_{k0} is the proportion of people who know all of the K correct choices. The ratio $\lambda_{ij}(K, N - K - J)/\eta_{ij}$ is the conditional probability of a response pattern where there are I correct alternatives among the first K attempts and where the K th correct alternative is chosen on the $(N - K - J)$ th attempt after the first K attempts, given that the individual knows i correct alternatives and can eliminate j incorrect choices. The constants $\lambda_{ij}(K, N - K - J)$ and η_{ij} can be derived in the same way as are the coefficients in the binomial and negative-binomial distributions, and these values are

$$\eta_{ij} = \binom{N - j - i}{K - i} \quad (2)$$

and

$$\lambda_{ij}(K, N - K - J) = \binom{K - i}{I - i} \binom{N - K - J - 1}{K - I - 1} \quad (3)$$

Thus, $\tau_{ij}\lambda_{ij}(K, N - K - J)/\eta_{ij}$ is the joint probability of knowing i correct alternatives, being able to eliminate j incorrect choices, choosing I correct alternatives among the first K attempts, and choosing the K th correct alternative on the $(N - K - J)$ th attempt after the first K attempts. This yields Equation 1.

The important point is that the τ_{ij} s can be determined if the $P_{I, N - K - J}$ s are known. For instance,

$$P_{0, N - K} = \tau_{00}\lambda_{00}/\eta_{00} \quad (4)$$

and since λ_{00} and η_{00} are known, τ_{00} is also known. Once τ_{00} has been determined, Equation 1 can be solved for τ_{01} since

$$P_{0, N - K - 1} = \tau_{00}\lambda_{00}/\eta_{00} + \tau_{01}\lambda_{01}/\eta_{01} \quad (5)$$

The remaining τ_{ij} values can be determined in a similar manner. Applying a result in Zehna (1966), it follows that if $\hat{P}_{I, N - K - J}$ is a maximum likelihood estimate of $P_{I, N - K - J}$, and if the $P_{I, N - K - J}$ s in Equation 1 are replaced with $\hat{P}_{I, N - K - J}$, the resulting value for τ_{ij} , say $\hat{\tau}_{ij}$, is a maximum likelihood estimate of τ_{ij} .

To help clarify matters, consider the special case $N = 5$ and $K = 2$. Then,

$$P_{03} = 2\tau_{00}/10 \quad (6)$$

$$P_{02} = \tau_{00}/10 + \tau_{01}/6 \quad (7)$$

$$P_{13} = 2\tau_{00}/10 + \tau_{10}/4 \quad (8)$$

$$P_{12} = 2\tau_{00}/10 + 2\tau_{01}/6 + \tau_{10}/4 + \tau_{11}/3 \quad (9)$$

$$P_{11} = 2\tau_{00}/10 + 2\tau_{01}/6 + 2\tau_{02}/3 + \tau_{10}/4 + \tau_{11}/3 \quad (10)$$

$$P_{20} = \tau_{00}/10 + \tau_{01}/6 + \tau_{02}/3 + \tau_{10}/4 + \tau_{11}/3 + \tau_{20} \quad (11)$$

If \hat{P}_{03} is the usual maximum likelihood estimate of P_{03} , then $\hat{\tau}_{00} = 5\hat{P}_{03}$ is a maximum likelihood estimate of τ_{00} , and maximum likelihood estimates of the remaining τ_{ij} values can also be determined. Note, however, that if the model is assumed to hold, various inequalities among the $P_{I, N - K - J}$ s must be true. For instance,

$$\frac{1}{2}P_{03} \leq P_{02} \leq P_{12} \leq P_{11} \leq P_{20} \quad (12)$$

General results on obtaining maximum likelihood estimates of the $P_{I, N - K - J}$ s assuming these inequalities are true, can be found in Barlow, Bartholomew, Bremner, & Brunk (1972). For recent results on this problem, see Dykstra and Robertson (1982). Observe that testing these inequalities provides a check on the model, and Robertson (1978) as well as Robertson and Wright (1981) described how this can be done.

Once the τ_{ij} s have been estimated, it is a simple matter to estimate a validity coefficient that is comparable to the measure of validity used by Schmittlein (1984). Suppose, for instance, it is decided that an individual knows that a correct alternative is indeed correct if and only if it is chosen prior to the selection of any incorrect choice. As already indicated, the measure of validity to be used in the probability that, for a randomly sampled individual, the number of correct alternatives actually known will be correctly determined. That is, the measure of validity is

$$\omega = \sum_i Pr(\text{knows } i \text{ alternatives and chooses exactly } i \text{ correct alternatives during the first } K \text{ choices}) \quad (13)$$

Let

$$\xi_{I,N-K-J} = \sum_{i=0}^I \sum_{j=0}^J A_{I=i} \tau_{ij} \lambda_{ij}(K, N - K - J) / \eta_{ij} \quad (14)$$

where $A_{I=i}$ is the usual indicator function defined by

$$A_{I=i} = \begin{cases} 1, & \text{if } I = i \\ 0, & \text{otherwise} \end{cases} \quad (15)$$

Then,

$$\omega = \sum \sum \xi_{I,N-K-J} \quad (16)$$

For the special case $K = 2$ and $N = 5$,

$$\begin{aligned} \omega &= P_{03} + P_{02} + \tau_{11}/3 + \tau_{11}/3 + \tau_{20} \\ &= 3\tau_{00}/10 + \tau_{01}/6 + 2\tau_{11}/3 + \tau_{20} \end{aligned} \quad (17)$$

An Illustration

As a simple illustration, suppose examinees are asked:

What is $2^{-1/2}$ equal to?

- a) 2, b) $1/\sqrt{2}$, c) $1/2$, d) $\sqrt{2}/2$, e) $-1/2$.

Answers *b* and *d* are both correct, $K = 2$ and $N = 5$. Suppose this item is administered to a random sample of examinees yielding $\hat{P}_{03} = .01$, $\hat{P}_{02} = .04$, $\hat{P}_{13} = .08$, $\hat{P}_{12} = .20$, $\hat{P}_{11} = .32$, and $\hat{P}_{20} = .35$. From Equations 6 through 11, maximum likelihood estimates of the τ s are

$$\begin{aligned} \hat{\tau}_{00} &= 5(.01) = .05, \\ \hat{\tau}_{01} &= (.04 - .005)6 = .21, \\ \hat{\tau}_{10} &= (.08 - .01)4 = .28, \\ \hat{\tau}_{11} &= (.20 - .01 - .07 - .07)3 = .15, \\ \hat{\tau}_{02} &= (.32 - .01 - .07 - .07 - .05) \left(\frac{3}{2}\right) = .18, \text{ and} \\ \hat{\tau}_{20} &= (.35 - .005 - .035 - .06 - .07 - .06) = .12. \end{aligned}$$

Thus, the estimated proportion of examinees who know both correct alternatives is $\tau_{20} = .12$. While all of the $\hat{\tau}$ s are positive, estimating the τ s with the \hat{P} s can yield negative values. This problem can be avoided by using results in Dykstra and Robertson (1982). Finally, $\hat{\omega} = 3(.05)/10 + .035 + .10 + .12 = .27$ is the estimate of the probability of correctly determining how many alternatives are known by a randomly sampled examinee.

Conclusions

The advantage of the model proposed here over the models used by Schmittlein (1984) and Gibbons et al. (1979) is that partial information has been taken into account; this is important because empirical studies indicate that partial information is common, and because ignoring the existence of partial information can seriously affect the conclusions drawn about the knowledge level of individuals. Although all indications are that the present model provides a more realistic approximation of how people behave when taking test items, it should be noted that in some cases an even more general model might be needed. In particular, there are instances when it might be necessary to take into account misinformation (e.g., Birenbaum & Tatsuoka, 1982; Wilcox, 1982c). Individuals are said to have misinformation if they eliminate a correct choice in the belief that it is incorrect. Unfortunately, there is at the moment no reasonable way of incorporating both partial information and misinformation into the model described above. Despite this difficulty, the new model proposed here should have practical value since experience with the case $K = 1$ indicates that the model provides a reasonable approximation of how examinees respond to most multiple-choice questions. For a possible approach to misinformation, see Duncan and Milton (1978).

Another advantage of the model proposed here is that it can be used to empirically check the model used by Schmittlein (1984). In particular, Schmittlein's model assumed $\tau_{ij} = 0$ for $j > 0$, and if this is true, certain equalities among the $P_{I,N-K-j}$ s must hold. For instance, for the case $K = 2$ and $N = 5$ it is seen that $\frac{1}{2}P_{03} = P_{02}$ and $P_{13} = P_{12} = P_{11}$, and these equalities can be tested in the usual way. Small sample critical values for testing the equality of multinomial cell probabilities have been reported by Katti (1973) and Smith, Rae, Manderscheid, and Silberg (1979). For recent results on approximating the null distribution of the usual chi-square test for equal cell probabilities, see Wilcox (1982a).

As a final note, Schmittlein's model (1984) assumed that the correct alternatives are equally difficult, and there is some evidence that this assumption can yield good results when estimating test-retest reliability (Subkoviak, 1978). However, in terms of assessing validity, it is not known what effect this assumption has. The model proposed here has the advantage of avoiding this assumption.

References

- Ashler, D. (1979). Biserial estimators in the presence of guessing. *Journal of Educational Statistics*, 4, 325-355.
- Barlow, R., Bartholomew, D., Bremner, J., & Brunk, H. (1972). *Statistical inference under order restrictions*. New York: Wiley.
- Birenbaum, M., & Tatsuoka, K. (1982). On the dimensionality of achievement test data. *Journal of Educational Measurement*, 19, 259-266.
- Bliss, L. B. (1980). A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. *Journal of Educational Measurement*, 17, 147-153.
- Coombs, C. H., Milholland, J. E., & Womer, F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement*, 16, 13-37.
- Cross, L. H., & Frary, R. B. (1977). An empirical test of Lord's theoretical results regarding formula-scoring of multiple-choice tests. *Journal of Educational Measurement*, 14, 313-321.
- Duncan, G. T., & Milton, E. O. (1978). Multiple answer multiple-choice test items: Responding and scoring through Bayes and minimax strategies. *Psychometrika*, 43, 43-57.
- Dykstra, R. L., & Robertson, T. (1982). An algorithm for isotonic regression for two or more independent variables. *Annals of Statistics*, 10, 708-716.
- Gibbons, J. D., Olkin, I., & Sobel, M. (1979). A subset selection technique for scoring items on a multiple-choice test. *Psychometrika*, 44, 259-270.
- Katti, S. K. (1973). Exact distribution for the chi-square test in the one-way table. *Communications in Statistics*, 2, 435-457.
- Robertson, T. (1978). Testing for and against an order restriction on multinomial parameters. *Journal of the American Statistical Association*, 73, 197-202.
- Robertson, T., & Wright, F. I. (1981). Likelihood ratio tests for and against a stochastic ordering between multinomial populations. *Annals of Statistics*, 9, 1248-1257.

- Schmittlein, D. C. (1984). Assessing validity and test-retest reliability for "pick k of n " data. *Marketing Science*, 3, 23-40.
- Smith, P. J., Rae, D. S., Manderscheid, R., & Silberg, S. (1979). Exact and approximate distributions of the chi-square statistics for equiprobability. *Communications in Statistics—Simulation and Computation*, B8, 131-149.
- Subkoviak, M. (1978). Empirical investigation of procedures for estimating reliability for mastery tests. *Journal of Educational Measurement*, 15, 111-116.
- Weitzman, R. A. (1970). Ideal multiple-choice items. *Journal of the American Statistical Association*, 65, 71-89.
- Wilcox, R. R. (1980). Determining the length of a criterion-referenced test. *Applied Psychological Measurement*, 4, 425-446.
- Wilcox, R. R. (1982a). A comment on approximating the χ^2 distribution in the equiprobable case. *Communications in Statistics—Simulation and Computation*, 11, 619-623.
- Wilcox, R. R. (1982b). Some empirical and theoretical results on an answer-until-correct scoring procedure. *British Journal of Mathematical and Statistical Psychology*, 35, 57-70.
- Wilcox, R. R. (1982c). Some new results on an answer-until-correct scoring procedure. *Journal of Educational Measurement*, 19, 67-74.
- Wilcox, R. R. (1983a). An approximation of the k out of n reliability of a test and a scoring procedure for determining which items an examinee knows. *Psychometrika*, 48, 211-222.
- Wilcox, R. R. (1983b). How do examinees behave when taking multiple-choice tests? *Applied Psychological Measurement*, 7, 239-240.
- Zehna, P. W. (1966). Invariance of maximum likelihood estimation. *Annals of Mathematical Statistics*, 37, 744.

Author's Address

Send requests for reprints or further information to Rand R. Wilcox, Department of Psychology, University of Southern California, Los Angeles CA 90089, U.S.A.