AN APPROXIMATION OF THE K OUT OF N RELIABILITY OF A TEST, AND A SCORING PROCEDURE FOR DETERMINING WHICH ITEMS AN EXAMINEE KNOWS

Rand R. Wilcox

CSE Report No. 212
1983

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Table of Contents

1. Introduction

Consider an n-item multiple-choice test, and suppose that every examinee can be described as either knowing or not knowing the correct response. In some situations, particularly with respect to some instructional program, the goal of a test might be to determine how many of the n items an examinee actually knows; in terms of diagnosis, it may even be desirable to determine which specific items an examinee knows or does not know. Under a conventional scoring procedure, about the only scoring rule available is one where it is decided that an examinee knows if and only if a correct response is given. Obviously guessing will affect the accuracy of this rule. If it is assumed that examinees who know will always give the correct response, and if most examinees really do know the correct response, then of course guessing has little impact on the accuracy of the test or the effectiveness of the distractors in terms of the typical examinee. However, if ζ is the proportion of examinees who know the answer to an item, then as ζ decreases, the importance of having effective distractors increases in order to avoid incorrect decisions about whether an examinee knows.

Guessing can seriously affect various other measurement problems as well (e.g., Weitzman, 1970; van den Brink & Koele, 1980; Wilcox, 1980, 1982c; Ashler, 1979). For example, when estimating the biserial correlation coefficient, guessing can substantially affect the results (Ashler, 1979). Ashler gives a method of correcting the estimate for the effects of guessing, but it requires a procedure for determining which items an examinee really knows. The conventional rule is to decide an examinee knows if and only if the correct response is given, but this can be unsatisfactory.

Suppose, for example, ζ=5, and the probability of a correct response, given that the examinee does not know, is 1/3. Then 1/6 of the examinees would be misclassified. The extreme case is where none of the examinees know, in which case 1/4 would be incorrectly judged as knowing the correct response.

As another example, suppose an investigator wants to determine whether the proportion of examinees who know an item is relatively large. In order to ensure a reasonably high probability of a correct decision about this proportion, it follows from Wilcox (1980) that it might be necessary to sample ten, perhaps even forty times as many examinees as would be required if guessing did not exist.

For a specific examinee taking a test, let x_i=1 if a correct decision is made about whether the answer to the ith item is known; otherwise x_i=0. For an examinee randomly sampled from the population of potential examinees, let

$$\rho_k = Pr(\sum x_i \geq k).$$

This is just the probability of making at least k correct decisions among the n items for a randomly sampled examinee; ρ_k is called the k out of n reliability of a test.

Suppose every item has t alternatives. One approach to designing a reasonably accurate test is to assume random guessing, and then choose t so that ρ_k is reasonably close to one. If x_i is independent of x_j for all i≠j, then ρ_k is easily calculated on a computer. Unfortunately, there are three serious problems with this approach. First, there is considerable empirical evidence that guessing is seldom at random (Coombs et al., 1956; Bliss, 1980; Cross & Frary, 1977; Wilcox, 1982a, 1982b). Second, even if guessing is at random, some situations will require more alternatives than is practical in order for ρ_k to be close to one (Wilcox, 1982c). Finally, there is no particular reason for assuming x_i independent of x_j, $i \neq j$, or to believe that such an assumption will give a good approximation of ρ_k. If $\mathrm{cov}(x_i, x_j) \neq 0$, bounds on ρ_k are available (Wilcox, 1982c, 1982d) but point estimates do not exist.

One goal in this paper is to suggest an approximation of ρ_k that can be estimated with an answer-until-correct test. Another and perhaps more important goal is to describe a scoring procedure that might be used when the estimate of ρ_k is judged to be too small under a conventional scoring rule. The new rule is based on a recently proposed latent structure model for test items. Included are some results on how to test whether this model is consistent with observed test scores.

2. An Approximation of ρ_k

Let $y=(y_1,\ldots,y_n)$ be any vector of length n where $y_i=0$ or 1, and let f(y) be the probability density function of y. Bahadur (1961) shows that f(y) can be written as

$$f(y) = f_1(y)h(y)$$

where

$$f_1(y) = \prod_{i=1}^{n} \alpha_i^{y_i} (1-\alpha_i)^{1-y_i}$$

$$\alpha_i = \Pr(y_i=1)$$

$$h(y)=1 + \sum_{i<j} r_{ij} z_i z_j + \sum_{i<j<m} r_{ijm} z_i z_j z_m + \cdots + r_{12\ldots n} z_1 \cdots z_n$$

$$z_i = (y_i - \alpha_i)/[\alpha_i(1-\alpha_i)]^{\frac{1}{2}}$$

$$r_{ij} = E(z_i z_j)$$

$$r_{ijm} = E(z_i z_j z_m)$$

$$\vdots$$

$$r_{12\ldots n} = E(z_1 z_2 \ldots z_n)$$

An mth order Bahadur approximation of f is one where the first m summations are used in the expression for h. Several authors have used a second order approximation when investigating problems in discrete discriminate analysis (e.g., Dillon & Goldstein, 1978; Gilbert, 1968; Moore, 1973). In this case f(y) is approximated with

$$g(y)=f_1(y)\ [1 + \sum_{i<j} r_{ij} z_i z_j] \tag{2.1}$$

Other approximations have been proposed, but as will become evident, (2.1) is particularly convenient for the situation at hand.

Occasionally (2.1) will not be a probability function. In particular, it may be that $g(y)<0$ for some vectors y. In this paper, whenever this occurred, g(y) was assumed to be zero, but the g(y) values were not rescaled so that they sum to one.

Bahadur (1961) discusses how to assess the goodness of fit of the approximation. Here, however, interest is in approximating ρ_k. Note that for a random vector y, ρ_k can be written as

$$\sum_{y:S\geq k} f(y) \tag{2.2}$$

where $S=\sum y_i$ and the summation in (2.2) is over all vectors y such that $S\geq k$.

Of course, when approximating ρ_k, f(y) would be replaced by f(x) where the vector x indicates items for which a correct decision is made about whether an examinee knows. To gain some insight into how well g(y) approximates ρ_k, assuming α_i and r_{ij} are known, we set n=5, k=4 and randomly chose values for the $2^5=32$ probability cells. Next, ρ_k was evaluated with (2.2), and then it was approximated with $\tilde{\rho}_k$ where $\tilde{\rho}_k$ is given by (2.2) with f(y) replaced by g(y). This process was repeated 100 times yielding a wide range of values for ρ_k. The values for ρ_k and $\tilde{\rho}_k$ were rounded to the second decimal place after which it was found that 85% of the time, $|\rho_k-\tilde{\rho}_k|\le .02$. For 5% of the approximations it was found that $|\rho_k-\tilde{\rho}_k|\ge .05$. For $|\rho_k-\tilde{\rho}_k|\le .05$ it was also found that $\tilde{\rho}_k<\rho_k$. The poorest approximation was for a probability function where $\rho_k = .365$ and $\tilde{\rho}_k=.232$. Although hardly conclusive, these results suggest that $\tilde{\rho}_k$ is generally useful when approximating ρ_k, at least when n is small. For n large the test can be broken into subtests containing five items or less, and Bonferroni's inequality (e.g., Tong, 1980) can be applied. For example, suppose n=10. If for the first five items $\tilde{\rho}_4=.95$, and for the remaining five items $\tilde{\rho}_4=.98$, then for the entire test it is estimated that

$$\rho_8\ge 1-(1-.98)-(1-.95)=.93. \tag{2.3}$$

Estimating $\tilde{\rho}_k$

There remains the problem of estimating $\tilde{\rho}_k$. What is needed is an estimate of the parameter r_{ij} in the expression for g(y). An estimate is available using a slight extension of the model in Wilcox (1982d) which can be briefly summarized as follows. Assume that examinees take the test according to an answer-until-correct scoring procedure. That is, they choose a response, and if it is wrong they choose another. This process continues until the correct response is selected. Administering such tests is easily accomplished with especially designed answer sheets that are available commercially.

Consider a specific item and let P_i be the probability that a randomly selected examinee gets the item correct on the ith attempt, i=1,...,5 where t is the number of alternatives. Let ζ_i be the proportion of examinees who can eliminate i distractors (i=0,...,t-1). It is assumed that for examinees who do not know, there is at least one effective distractor in which case ζ_{t-1} is the proportion of examinees who know. It is also assumed that once examinees eliminate as many distractors as they can, they guess at random from among those alternatives that remain. It follows that

$$P_i = \sum_{j=0}^{t-i} \zeta_j/(t-j) \qquad (i=1,\ldots,t) \tag{2.4}$$

and the model implies that

$$P_1 \geq P_2 \geq \cdots \geq P_t \tag{2.5}$$

which can be tested (Robertson, 1978). For empirical results in support of this model, see Wilcox (1982a, 1982b, 1983). In the few instances where (2.5) seems to be unreasonable, a misinformation model appears to explain the observed test scores. When (2.5) is assumed, the pool-adjacent violators algorithm (Barlow et al., 1972) yields a maximum likelihood estimate of the P_i's. These estimates in turn yield an estimate of the ζ_i's.

For any pair of items, let P_{ij} be the probability of a correct response on the ith attempt of the first and the jth attempt of the second, respectively.

And let ζ_{ij} be the probability that a randomly chosen examinee can eliminate i distractors from the first, and j distractors from the second. Then $\zeta_{t-1,t-1}$ is the proportion of examinees who know both. It is assumed that an examinee's guessing rate is independent over the items not known, and so

$$p_{km} = \sum_{i=0}^{t-k} \sum_{j=0}^{t-m} \zeta_{ij}/ [(t-i)(t-j)] \quad . \tag{2.6}$$

If the second item has t´ alternatives, t≠t´, simply replace t with t´ in the second summation. Testing certain implications of (2.6) is discussed below.

For the i<u>th</u> item on the test, let τ_i=E(x_i) be the probability of a correct decision about whether the examinee knows when a conventional scoring procedure is used. Thus, τ_i plays the role of α_i when approximating ρ_k. For an answer-until-correct test, a conventional rule means to decide an examinee knows if and only if the correct response is given on the first attempt. In this case (Wilcox, 1982a)

$$\begin{aligned}\tau_i &= \zeta_{t-1}+1-p_1\\ &= 1-p_2 \ .\end{aligned}$$

Thus, if for the i<u>th</u> item, c_j of N examinees get the correct response on the j<u>th</u> attempt under an answer-until-correct scoring procedure, then

$$\hat{\tau}_i = 1-c_2/N$$

is an estimate of τ_i. If the c_j's are inconsistent with (2.5), apply the pool-adjacent-violators algorithm (Barlow et al., 1972, pp. 13-16), as was previously mentioned.

In a similar manner, let $\tau_{ij}=Pr(x_i=1,x_j=1)$, i.e., τ_{ij} is the probability of a correct decision for both items i and j. For the conventional decision rule under an answer-until-correct model, it can be seen that

$$\tau_{ij} = \sum_{k=1}^{t} \sum_{m=1}^{t} q_{km}$$

where

$$q_{11} = \zeta_{t-1,t-1}$$

$$q_{i1} = \sum_{k=0}^{t-i} \zeta_{k,t-1}/(t-k) \qquad (i=2,\ldots,t)$$

$$q_{1j} = \sum_{k=0}^{t-j} \zeta_{t-1,k}/(t-k) \qquad (j=2,\ldots,t)$$

$$q_{ij} = P_{ij} \qquad (i>1 \text{ and } j>1)$$

(Wilcox, 1982d). Thus, r_{ij}, z_i and z_j in equation (2.1) are easily determined. In particular,

$$r_{ij} = \frac{\tau_{ij}-\tau_i\tau_j}{[\tau_i\tau_j(1-\tau_i)(1-\tau_j)]^{\frac{1}{2}}}$$

where τ_i plays the role of α_i in the definition of z_i. But as noted in Wilcox (1982d), the ζ_{ij}'s in equation (2.6) are easily estimated, and these estimates yield an estimate of τ_{ij} which in turn gives an estimate of r_{ij}. Hence, $\tilde{\rho}_k$ can be estimated with equation (2.1) which gives an approximation of ρ_k.

Testing Certain Implications of the Model

For any pair of items, equation (2.6) implies that

$$p_{11} \geq p_{12} \geq \cdots \geq p_{1t'} \geq p_{2t'} \geq \cdots \geq p_{tt'} \qquad (2.7a)$$

$$p_{11} \geq p_{21} \geq \cdots \geq p_{t1} \geq p_{t2} \geq \cdots \geq p_{tt'} \qquad (2.7b)$$

$$p_{i1} \geq p_{i2} \geq \cdots \geq p_{it'} \quad (i=2,\ldots,t-1) \qquad (2.7c)$$

and

$$p_{1j} \geq p_{2j} \geq \cdots \geq p_{tj} \quad (j=2,\ldots,t'-1) \qquad (2.7d)$$

whereas before, t and t′ are the number of alternatives for the first and second items, respectively. A few other inequalities are implied if the ζ_{ij}'s are assumed to be probabilities, but these have not been derived.

Experience with real data suggests that when observed scores are consistent with (2.5), the inequalities in (2.7) will also hold. If some of the observed proportions are inconsistent with (2.7), maximum likelihood estimates can be obtained when the model is assumed to be true by applying the minimax order algorithm in Barlow et al. (1972).

Robertson (1978) includes some asymptotic results on testing (2.7). At the moment, however, his proposed procedure cannot be applied because certain constants (the $P_q(\ell,k)$'s in Robertson's notation) are not available. An alternative approach is to perform a separate test of the inequalities in (2.7d), one corresponding to every j, j=2,...,t′-1, then perform a test of (2.7c), one for every i=2,...,t-1, then test (2.7b) and finally (2.7a). The total number of tests is m-t+t′-2. If the critical value for every test is set at α/m, then from the Bonferroni inequality (e.g., Tong, 1980), the probability of a Type I error among the m tests is at most α.

Consider, for example, the inequalities in (2.7d) for j=2. That is, the goal is to test

$$H_0: \ p_{12} \geq p_{22} \geq p_{32} \geq \cdots \geq p_{t2} \tag{2.8}$$

Let λ be the likelihood ratio for testing (2.8) where the alternative hypothesis is no restriction on the proportions. From Robertson (1978, Theorem 2), the asymptotic null distribution of $T=-2 \underline{\ln} \lambda$ is

$$\Pr(T>T_0) = \sum_{\ell=1}^{k-1} P(\ell,k)\Pr(\chi^2_{k-\ell} \geq T_0) \tag{2.9}$$

where $P(\ell,k)$ is the probability that the maximum likelihood estimate of $p_{12},\ldots,p_{t2}$ subject to (2.8) will have ℓ distinct values among the k parameters being estimated, and $\chi^2_{k-\ell}$ is a chi-square random variable with $k-\ell$ degrees of freedom. For (2.8), k=t. (As previously mentioned, the pool-adjacent-violators algorithm yields maximum likelihood estimates when (2.8) is assumed.) The constants $P(\ell, k)$ can be read from Table A.5 in Barlow et al. (1972).

Thus, in order for the m tests to have a critical level of at most α, choose T_0 so that (2.9) equals α/m, and reject H_0 if $T>T_0$. This process is repeated for the other inequalities to be tested, but note that k (the number of parameters being tested) will have a different value for (2.7a) and (2.7b).

To facilitate this procedure, critical values are reported in Table 1 for $t=2(1)5$, $\alpha=.1$, $.05$, $.01$; and some appropriately chosen values for m. (Additional values for m were not used because for $t\leq 5$, these are the only values of m that will occur.)

As an illustration, suppose $t-t'=3$. Then there are $m=4$ sets of inequalities to be tested. If $\alpha=.05$, then from (2.7a) there are $k=5$ parameters, and so $T_0=10.81$. For (2.7b) again $k=5$ and $T_0=10.81$. For (2.7c) there is only one set of inequalities which corresponds to $i=2$, $t=k=3$, and $T_0=7.24$. The same is true for (2.7d).

3. A Scoring Procedure for Tests

Consider a specific item on an n-item test. In contrast to most of the existing scoring procedures, the goal here is to minimize the expected number of examinees for whom an incorrect decision is made about whether they know the answer to the item. It is interesting to note that when items are scored right/wrong, this criterion can rule out the conventional rule where it is decided an examinee knows if and only if the correct response is given. The extreme case is where $\zeta_{t-1}=0$, i.e., none of the examinees know, in which case the optimal rule is to decide that an examinee does not know regardless of the response given. If $\beta=\Pr$ (correct | examinee does not know), it can be seen that if an item is scored right/wrong, and if $\beta>\zeta_{t-1}/(1-\zeta_{t-2})$ the optimal rule is to always decide that examinees do not know. If $\beta<\zeta_{t-1}/(1-\zeta_{t-1})$, use the conventional rule. From Copas (1974), this approach (in terms of parameters) is admissible.

These parameters can be estimated which yields an estimate of the optimal decision rule (e.g., Macready & Dayton, 1977). The goal here is to derive a decision rule based on an answer-until-correct scoring procedure. The advantage of this new approach is that it is not necessary to assume all n items are equivalent as was done in Macready and Dayton. (Two items are said to be equivalent if every examinee knows both or neither one.) The results in Macready and Dayton (1977) could be extended to the case of hierarchically related items by applying results in Dayton and Macready (1976), but here the goal is to derive a rule where no particular relationship is assumed among the items. However, the situation considered by Macready and Dayton (1977) has the advantage of allowing Pr(incorrect response | examinee knows) > 0, while here this probability is assumed to be zero.

Consider the ith item on a test taken by a specific examinee, and let $w_i=1$ if it is decided the examinee knows; otherwise $w_i=0$. Consider the jth item on the test $i\neq j$ for the purpose of assisting in the decision about whether w_i should be 1 or 0. (The optimal choice for the second item will become evident.) It is assumed that items are administered according to an answer-until-correct scoring procedure. For a specific examinee, let v_i be the number of attempts needed to choose the correct response to the ith item. The decision rule to be considered is

$$w_i(v_j) = \begin{matrix} 1, \text{ if } v_i<v_{0i} \text{ and } v_j\leq v_{0j} \\ 0, \text{ if otherwise} \end{matrix} \qquad (3.1)$$

where v_{0i} (=1 or 2) and v_{0j} ($1\leq v_{0j}\leq t'$) are constants to be determined. Note that when $v_{0i}=2$ and $v_{0j}=1$, the rule is similar to the one in Macready and Dayton (1977). Also note that $v_{0j}=t'$ corresponds to the conventional decision rule where the information about the jth item plays no role in determining whether the examinee knows the ith. It is evident, therefore, that in terms of parameters, (3.1) always improves upon the conventional approach. The improvement actually achieved will of course vary. If ζ_{t-1} is close to one for every item, ρ_k will also be close to one under a conventional scoring rule, in which case there is little motivation for using (3.1). However, when ρ_k is unacceptably small, (3.1) can increase ρ_k by a substantial amount.

One problem is choosing the constants v_{0i} and v_{0j}. A solution is as follows. For a randomly sampled examinee responding to the ith and jth items, let p_{km1} be the probability of choosing the correct response on the kth attempt of the ith item, the mth attempt of the jth item, and making a correct decision under the rule (3.1). The probability of a correct decision for a randomly sampled examinee is

$$p_c=\sum_{k=1}^{t}\sum_{m=1}^{t'} p_{km1}$$

which is a function of v_{0i} and v_{0j}. Thus, the obvious choice for v_{0i} and v_{0j} is the one that maximizes p_c.

Let

$$q_{1j}=\sum_{k=0}^{t'-j} \zeta_{t-1,k}/(t'-k) \qquad (j=1,\ldots,t') \qquad (3.2)$$

and

$$Q = \sum_{i=2}^{t} \sum_{j=1}^{t} P_{ij} \quad .$$

For $\quad v_{0i}=2$ and any v_{0j}

$$P_c = Q + \sum_{k=1}^{v_{0j}} q_{1k} + \sum_{k=v_{0j}+1}^{t'} (p_{1k}-q_{1k}) \ . \tag{3.3}$$

When $v_{0j}=t'$, the second sum in (3.3) is taken to be zero. As for $v_{0i}=1$,

$$p_c = Q + \sum_{k=1}^{t'} (p_{1k}-q_{1k}). \tag{3.4}$$

Thus, to determine the optimal choice for v_{0i} and v_{0j} in (3.1), simply evaluate p_c for every possible choice of v_{0i} and v_{0j}, and then set v_{0i} and v_{0j} equal to the values that maximize p_c. Of course, when making a decision about the ith item, this process can be repeated over the n-1 other items on the test. The item that maximizes p_c is the one that should be used when determining whether an examinee knows the ith item.

An Illustration

As a simple illustration, the optimal rule is estimated for two items used in Wilcox (1982a). The observed frequencies are shown in Table 2. Note that the observed frequencies already satisfy (2.7a)-(2.7d). For the first item the estimate of τ, the probability of correctly determining whether a randomly sampled examinee knows, is $\hat{\tau}_1=(236-71)/236=.699$. For the second items it is $\hat{\tau}_2=.78$.

Suppose the second item is used to help determine whether an examinee knows the first. Let $v_{01}=2$ and $v_{02}=1$. Thus, a correct response must be (the joint probability of making a correct decision about the ith and jth item) must be known. (See Section 2.) But when $v_{0j}<t$, τ_{ij} may depend on two other items, say items k and m. That is, information on the kth item and mth item will be used to determine whether the examinee knows the ith and jth items respectively. Hence, (2.6) is no longer adequate for determining ρ_k.

One solution might be to extend (2.6) to include four items. In theory the parameters could be estimated under the resulting inequalities by applying the minimax order algorithm. However, writing an appropriate computer program that is valid for $t\leq 5$ will be a relatively involved task.

Another and perhaps more practical approach might be to restrict the decision rule so that if the response to the jth item is used in the decision about whether an examinee knows the ith item, then the response to the ith will be used in deciding about the jth. An advantage of this approach is that it simplifies the process of choosing a decision rule by reducing the number of pairs of items that are considered. A second advantage is that an approximation of ρ_k can be made using the results in Section 2. A disadvantage is that by restricting the class of decision rules, the potential increase in ρ_k (over what it is under a conventional scoring rule) is reduced. Perhaps this is not a serious problem; at the moment it is impossible to say.

An approach to choosing a scoring rule might be as follows: First estimate ρ_k under a conventional scoring rule. If it is judged to be too small, choose a decision rule from among the rules described in the preceding paragraph and then estimate ρ_k in the manner indicated below. If ρ_k is still too small, choose a decision rule from among the broader class of rules described in the preceding subsection. In this case, however, an approximation of ρ_k is no longer available for the reasons just given.

Suppose that if the j<u>th</u> item is chosen to aid in the decision about the i<u>th</u>, then the i<u>th</u> item is used in the decision rule for the j<u>th</u>. What is needed in order to approximate ρ_k is an expression for the joint probability of making a correct decision for both items. Accordingly, consider any two items, and let $u_1(k,m)=1$ if it is decided that an examinee knows the first item if the correct response is given on the k<u>th</u> attempt of the first item, and the m<u>th</u> attempt of the second; otherwise $u_1(km,)=0$. Similarly, $u_2(k,m)=1$ if it is decided that an examinee knows the second item if the correct response is given on the k<u>th</u> attempt of the first and the m<u>th</u> attempt of the second; otherwise $u_2(k,m)=0$. Let

$$s_{1i}(k,n) = \begin{cases} 1, \text{ if } u_1(k,m)=1 \text{ and } i=t-1, \text{ or if} \\ \quad u_1(k,m) = 0 \text{ and } i<t-1 \\ 0, \text{ otherwise} \end{cases}$$

and

$$s_{2j}(km) = \begin{cases} 1, \text{ if } u_2(k,m)=1 \text{ and } j=t'-1, \text{ or if} \\ \quad u_2(k,m)=0 \text{ and } j<t'-1 \\ 0, \text{ otherwise.} \end{cases}$$

Recall that the probability of getting the correct response on the k<u>th</u> attempt of the first item and the m<u>th</u> attempt of the second is given by (2.6). From this expression it can be seen that the joint probability of k attempts on the first item, m attempts on the second, and a correct decision on both items is

$$\gamma_{km} = \sum_{i=0}^{t-k} \sum_{j=0}^{t'-m} s_{1i}(k,m)s_{2j}(k,m)\zeta_{ij}/[(t-i)(t'-j)]$$

Thus, for a randomly sampled examinee, the joint probability of a correct decision for both items, say items i and j, is

$$\tau_{ij} = \sum_{k=1}^{t} \sum_{m=1}^{t'} \gamma_{km}$$

The joint probability of a correct decision about the first item, k attempts on the first and m attempts on the second is

$$\psi_{km} = \sum_{i=0}^{t} \sum_{j=0}^{t'} s_{1i}(k,m)\ \zeta_{ij}/[(t-i)(t'-j)]$$

The corresponding probability for the second item is

$$\eta_{km} = \sum_{i=0}^{t} \sum_{j=0}^{t'} s_{2j}(k,m)\zeta_{ij}/[(t-i)(t'-j)]$$

Thus, τ_i, the probability of a correct decision about the i<u>th</u> item on a test (using the j<u>th</u> item in (3.1)) for a randomly sampled examinee, is

$$\tau_i = \sum_{k=1}^{t} \sum_{m=1}^{t'} \psi_{km}$$

Similarly, for the second item, item j,

$$\tau_j = \sum_{k=1}^{t} \sum_{m=1}^{t'} \eta_{km}$$

Hence, ρ_k can be approximated as described in Section 2.

Concluding Remarks

Virtually all of the results on the proposed scoring rule have been in terms of parameters. These parameters are not known, but they are easily estimated. The question arises as to the sampling effects on estimating the approximation of ρ_k, and on estimating the optimal decision rule for determining whether an examinee knows the correct response. In some instances, a large number of examinees will be available, and so very accurate estimates of the parameters can be obtained. This is the case for certain testing firms where literally thousands of examinees take the same test. When the number of examinees is small, however, sampling fluctuations need to be taken into account; this problem is currently being investigated.

Another important feature of the proposed scoring rule is that the decision about whether an examinee knows an item is a function of the responses given by the other examinees. If the goal is to minimize the number of examinees for whom an incorrect decision is made, there is no problem. However, in some instances, this feature might be objectionable. Suppose, for example, an examinee takes a test to determine whether a high school diploma will be received. It is possible for an examinee to fail because of how other examinees perform on the test even though the examinee in question deserves to pass. If this type of error is highly objectionable, perhaps the proposed scoring rule should be used only in diagnostic situations where the goal is to determine how many items an examinee actually knows, or which specific items are not known.

A technical point that should be mentioned is that a few of the $\hat{\zeta}_{ij}$'s were slightly negative in which case $\hat{\zeta}_{ij}$ was set equal to zero. As a result, the $\hat{\zeta}_{ij}$'s sum to .997 rather than one as they should. The problem is that equations (2.7a)-(2.7d) are necessary but not sufficient conditions for the model to hold. For example, these inequalities do not guarantee that ζ_{12} will be positive.

Despite these difficulties, there will be situations where correcting for guessing can be important. Some examples were given at the beginning of the paper. Even if a conventional scoring procedure is to be used in operational versions of a test, it might be important to first estimate the effects of guessing using an answer-until-correct scoring procedure.

Many scoring rules have been proposed that are based on various criteria. If a particular criterion is deemed important, of course the corresponding scoring rule should be considered. The point is that most of these rules are not based on the goal of determining how many items an examinee knows, or which specific skills an examinee has failed to learn. Moreover, typical rules usually ignore guessing or assume guessing is at random. Thus, the results reported here might be useful in certain situations.

TABLE 1

Critical Values T_0 for the Bonferroni Test of Equations (2.7)

k	m	α: .1	.05	.01
3	4	5.90	7.24	10.38
3	5	6.33	7.67	10.81
3	6	6.68	8.03	11.17
4	3	7.03	8.49	11.86
4	4	7.64	9.10	12.46
4	5	8.11	9.56	12.92
4	6	8.49	9.95	13.30
5	4	9.25	10.81	14.36
5	5	9.75	11.31	14.85
5	6	10.16	11.71	15.25
5	7	10.51	12.05	15.58
5	8	10.81	12.35	15.87
6	5	11.32	12.96	16.67
7	6	13.29	15.00	18.85
8	7	15.18	16.95	20.93
9	8	17.02	18.84	22.94

TABLE 2

Observed Frequencies for Two Items Administered Under
An Answer-Until-Correct Scoring Procedure

		Number of Attempts for the Second Item				
		1	2	3	4	
Number of	1	81	21	10	3	115
Attempts	2	44	18	6	3	71
for the	3	20	7	5	1	33
First Item	4	10	6	1	0	17
		155	52	22	7	236

References

Ashler, D. Biserial estimators in the presence of guessing. Journal of Educational Statistics, 1979, 4, 325-355.

Bahadur, R. R. A representation of the joint distribution of responses to n dichotomous items. In H. Solomon (Ed.), Studies in item analysis and prediction. Stanford: Stanford University Press, 1981.

Barlow, R., Bartholomew, D., Bremner, J., & Brunk, H. Statistical inference under order restrictions. New York: Wiley, 1972.

Bliss, L. B. A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational Measurement, 1980, 17, 147-153.

Coombs, C. H., Milholland, J. E., & Womer, F. B. The assessment of partial information. Educational and Psychological Measurement, 1956, 16, 13-37.

Copas, J. B. On symmetric compound decision rules for dichotomies. Annals of Statistics, 1974, 2, 199-204.

Cross, L. H., & Frary, R. B. An empirical test of Lord's theoretical results regarding formula-scoring of multiple-choice tests. Journal of Educational Measurement, 1977, 14, 313-321.

Dayton, C. M., & Macready, G. B. A probabilistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

Dillon, W. R., & Goldstein, M. On the performance of some multinomial classification rules. Journal of the American Statistical Association, 1978, 73, 305-313.

Gilbert, E. S. On discrimination using qualitative variables. Journal of the American Statistical Association, 1968, 63, 1399-1412.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Moore II, D. H. Evaluation of the five discrimination procedures for binary variables. Journal of the American Statistical Association, 1973, 68, 399-404.

Robertson, T. Testing for and against an order restriction on multinomial parameters. Journal of the American Statistical Association, 1978, 73, 197-202.

Tong, Y. L. Probability inequalities in multivariate distributions. New York: Academic Press, 1980.

van den Brink, W. P., & Koele, P. Item sampling, guessing and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1980, 33, 104-108.

Weitzman, R. A. Ideal multiple-choice items. Journal of the American Statistical Association, 1970, 65, 71-89.

Wilcox, R. R. Determining the length of a criterion-referenced test. Applied Psychological Measurement, 1980, 4, 425-446.

Wilcox, R. R. Some empirical and theoretical results on an answer-until-correct scoring procedure. British Journal of Mathematical and Statistical Psychology, 1982, 35, 57-70. (a)

Wilcox, R. R. Some new results on an answer-until-correct scoring procedure. Journal of Educational Measurement, 1982, 19, 67-74. (b)

Wilcox, R. R. Using results on k out of n system reliability to study and characterize tests. Educational and Psychological Measurement, 1982, 42, 153-165. (c)

Wilcox, R. R. Bounds on the k out of n reliability of a test, and an exact test for hierarchically related items. Applied Psychological Measurement, 1982, 6, 327-336. (a)

Wilcox, R. R. How do examinees behave when taking multiple-choice tests? Applied Psychological Measurement, 1983, 7, 239-240.