

BOUNDS ON THE K OUT OF N RELIABILITY
OF A TEST, AND AN EXACT TEST FOR
HIERARCHICALLY RELATED ITEMS

Rand R. Wilcox

CSE Report No. 211
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

	<u>Page</u>
Introduction	1
An Extension of an Answer-Until-Correct Scoring Procedure	4
Bounds On p_k	8
An Illustration	9
Estimating τ_{ij} When There is Misinformation	11
Testing Whether Items are Equivalent or Hierarchically Related	14
A Concluding Remark	15
Table 1	16
References	17

Introduction

Consider a randomly sampled examinee responding to a multiple-choice test item. In mental test theory there are, of course, many procedures that might be used to analyze this item. One approach might be as follows. Suppose a conventional scoring procedure is used where it is decided that an examinee knows the correct response if the correct alternative is chosen, and that otherwise the examinee does not know. If it were possible to estimate the probability, τ , of correctly determining an examinee's latent state (whether he/she knows the correct response) based on the above decision rule, this would give an indication of how well the distractors are performing for the typical examinee. The obvious problem is that under normal circumstances, there is no way of estimating this probability unless additional assumptions are made. One approach is to assume that examinees guess at random among the alternatives when they do not know the answer. If this knowledge or random guessing model holds, τ is easily estimated. However, empirical investigations (Bliss, 1980; Cross & Frary, 1977) suggest that this assumption will frequently be violated, and some related empirical results (Wilcox, 1982a, 1982b) indicate that such a model can be entirely unsatisfactory for other reasons as well.

Another approach is to use a latent structure model, and many such models have been proposed for measuring achievement (e.g., Brownless & Keats, 1958; Marks & Noll, 1967; Knapp, 1977; Dayton & Macready, 1976, 1980; Macready & Dayton, 1977; Wilcox, 1977a, 1977b, 1981a; Bergan et al., 1980). The choice of a model depends on what one is willing to assume in

a particular situation. These models make it possible to estimate errors at the item level such as

$$\beta = \text{Pr}(\text{randomly selected examinee gives the correct response} | \text{examinee does not know}) \quad [1]$$

which in turn yields an estimate of τ . An illustration is given in a later section. (For a review of latent structure models vis-a-vis criterion-referenced tests, see Macready and Dayton, 1980.) For some recent general comments on using latent structure models to measure achievement, see Molenaar (1981) and Wilcox (1981b).

Assume for a moment that for each item on an n-item test, an estimate of τ can be made. Let $x_i = 1$ if a randomly selected examinee's latent state is correctly determined for the ith item; otherwise $x_i = 0$. Then $E(x_i) = \tau_i$ ($i = 1, \dots, n$) is the probability of a correct decision on the ith item where the expectation is taken over the population of examinees.

Within the framework just described, how should an n-item test be characterized? Observing that $\sum x_i$ is the number of correct decisions among the n items, an obvious approach is to use

$$\mu = E(\sum x_i) = \sum \tau_i \quad [2]$$

where the expectation is over some particular population of examinees. The parameter μ is just the expected number of correct decisions among the n items for a typical examinee.

Knowing μ might not be important for certain types of tests, but surely it is important for some achievement tests. However, even if μ is known exactly, it would be helpful to have some additional related

information about Σx_i . For instance, a test constructor would have a better idea of how the test performs if $\text{VAR}(\Sigma x_i)$ could be determined. The problem is that $\text{VAR}(\Sigma x_i)$ depends on $\text{COV}(x_i, x_j)$, but this last quantity is not known and at present there is no way of estimating it. An alternative approach is to use the k out of n reliability of the test (Wilcox, 1982a) which is given by

$$\rho_k = \text{Pr}(\Sigma x_i \geq k). \quad [3]$$

In other words, if the goal of a test is to determine which of n items an examinee knows, and if a conventional scoring procedure is used, ρ_k is the probability of making at least k correct decisions for the typical examinee.

Suppose, for example, $n = 10$ and μ is estimated to be 7. Thus, the expected number of correct decisions is 7, but there is no information about the likelihood that at least 7 correct decisions will be made. If ρ_k were known, a test constructor would have some additional and useful information for judging the accuracy of the test. ρ_k might also be used as follows. Suppose it is desired to have $\rho_8 \geq .9$. If μ is estimated to be 9.1, this is encouraging, but it is not clear what implications this has in terms of making at least 8 correct decisions for the typical examinee.

If x_i is independent of x_j , $i \neq j$, an exact expression for ρ_k is available via the compound binomial distribution. Perhaps there are situations where this independence might be assumed, but it is evident that this independence will not always hold. If it can be assumed that

$\text{COV}(x_i, x_j) \geq 0$, bounds on ρ_k are available (Wilcox, 1983a). Recently Sathe, Pradhan, and Shah (1980) derived bounds on ρ_k that make no assumption about $\text{COV}(x_i, x_j)$. The main point of this paper is that these bounds can be estimated using an extension of an answer-until-correct (AUC) scoring procedure proposed by Wilcox (1981a).

An Extension of an Answer-Until-Correct (AUC) Scoring Procedure

As just indicated, an extension of results in Wilcox (1981a) is needed in order to apply the bounds derived by Sathe et al. (1980). First, however, it is helpful to briefly review the procedure and basic assumptions in Wilcox (1981a).

Consider a specific test item having t alternatives from which to choose, one of which is the correct response. Assume examinees respond according to an AUC scoring procedure. This means that examinees choose an alternative, and they are told immediately whether the correct response has been identified. If they are incorrect another response is chosen, and this process continues until they are successful. Special forms are generally available for administering AUC tests which make these tests easy to use in the classroom.

Let ζ_{t-1} be the proportion of examinees who know the correct response, and let ζ_i ($i=0, \dots, t-2$) be the proportion of examinees who can eliminate i distractors given that they do not know the correct response. Wilcox (1981a) assumes that examinees eliminate as many distractors as they can, and then choose at random from among those that remain. If p_i is the

probability of choosing the correct response on the ith attempt, then

$$p_i = \sum_{j=0}^{t-i} \zeta_j / (t-j) \quad (i=1, \dots, t). \quad [4]$$

Note that the model assumes that at least one effective distractor is being used. Put another way, no distinction is made between examinees who know the answer and examinees who can eliminate all of the distractors. Also, the model assumes $\text{Pr}(\text{incorrect response} | \text{examinee knows}) = 0$. In certain special cases this assumption can be avoided (e.g., Macready & Dayton, 1977), and the results reported here are easily extended to this case (cf. Molenaar, 1981; Wilcox, 1981b).

Assuming the model holds,

$$\zeta_{t-1} = p_1 - p_2 \quad [5]$$

and

$$\tau = \zeta_{t-1} + 1 - p_1 = 1 - p_2. \quad [6]$$

If in a random sample of N examinees, y_i examinees are correct on their ith attempt, $\hat{p}_i = y_i/N$ is an unbiased estimate of p_i which yields an estimate of ζ_{t-1} and τ .

Although empirical studies suggest that this model will frequently be reasonable (Wilcox, 1982a, 1982b), there are instances where this will not be the case. For example, some items might require a misinformation model, and an appropriate modification of the AUC scoring procedure has been proposed (Wilcox, 1982a, 1982b). The results outlined here are readily extended to this case, and a brief outline of how this can be done is given below.

Consider any two items on an n -item test, say items i and j . Apply-

ing results in Sathe et al. requires an estimate of $\tau_{ij} = \Pr(x_i=1, x_j=1)$, i.e., the joint probability of making a correct decision for both items i and j . The remainder of this section outlines how this might be done.

It is assumed that an examinee's guessing rate is independent over the items that he/she does not know. This means, for example, that if an examinee can eliminate all but 2 alternatives on item i , and all but 3 alternatives on item j , the probability of choosing the correct response on the first attempt of both items is $(1/2)(1/3) = 1/6$.

For the two items under consideration, let p_{km} ($k, m=1, \dots, t$) be the probability that a randomly selected examinee chooses the correct response on the k th attempt of the first item, and the correct response on the m th attempt of the second. If ζ_{gh} is the proportion of examinees who can eliminate g distractors from the first item and h distractors from the second ($g, h = 1, \dots, t-1$), then

$$p_{km} = \sum_{i=0}^{t-k} \sum_{j=0}^{t-m} \zeta_{ij} / (t-i)(t-j) . \quad [7]$$

The last expression can be used to express $\zeta_{t-1, t-1}$ in terms of the p_{km} 's which can be used to estimate $\zeta_{t-1, t-1}$. Note that if the first item has t' alternatives, $t' \neq t$, simply replace $t-k$ with $t'-k$ in equation 7.

To clarify matters, consider the special case $t = 3$. Equation 7 says that

$$p_{11} = \zeta_{22} + \zeta_{21}/2 + \zeta_{20}/3 + \zeta_{12}/2 + \zeta_{11}/4 + \zeta_{10}/6 + \zeta_{02}/3 \quad [8]$$

$$+ \zeta_{01}/6 + \zeta_{00}/9$$

$$p_{12} = \zeta_{21}/2 + \zeta_{20}/3 + \zeta_{11}/4 + \zeta_{10}/6 + \zeta_{01}/6 + \zeta_{00}/9 \quad [9]$$

$$p_{13} = \zeta_{20}/3 + \zeta_{10}/6 + \zeta_{00}/9 \quad [10]$$

$$p_{21} = \zeta_{12}/2 + \zeta_{02}/3 + \zeta_{11}/4 + \zeta_{01}/6 + \zeta_{10}/6 + \zeta_{00}/9 \quad [11]$$

$$p_{22} = \zeta_{11}/4 + \zeta_{10}/6 + \zeta_{01}/6 + \zeta_{00}/9 \quad [12]$$

$$p_{23} = \zeta_{10}/6 + \zeta_{00}/9 \quad [13]$$

$$p_{31} = \zeta_{02}/3 + \zeta_{01}/6 + \zeta_{00}/9 \quad [14]$$

$$p_{32} = \zeta_{01}/6 + \zeta_{00}/9 \quad [15]$$

$$p_{33} = \zeta_{00}/9 \quad [16]$$

Thus, starting with equation 16

$$\zeta_{00} = 9p_{33} \quad [17]$$

$$\zeta_{01} = 6(p_{32} - p_{33}) \quad [18]$$

and eventually ζ_{22} can be expressed in terms of the p_{km} 's. Replacing the p_{km} 's with their usual unbiased estimate yields an estimate of ζ_{22} say $\hat{\zeta}_{22}$. But it can be seen that for the two items under consideration (items i and j),

$$\tau_{ij} = \zeta_{22} + 1 - p_{11} \quad [19]$$

Replacing ζ_{22} and p_{11} with $\hat{\zeta}_{22}$ and \hat{p}_{11} yields an estimate of $\tau_{ij} = \Pr(x_i=1, x_j=1)$, say $\hat{\tau}_{ij}$. For arbitrary t , τ_{ij} is given by equation 19 with ζ_{22} replaced with $\zeta_{t-1,t-1}$. Note however, that the model implies that certain inequalities among the p_{km} 's must hold. For example, $p_{31} \geq p_{32} \geq p_{33}$. Estimating the p_{km} 's assuming these inequalities are true requires an application of the pool-adjacent violators algorithm (Barlow et al., 1972). Testing these inequalities can be accomplished by applying results in Robertson (1978).

Bounds on p_k

This section describes how the results in the previous section can be used to estimate bounds on p_k . First, however, results in Sathe et al. (1980) are summarized. Recall that $\mu = \sum \tau_i$ and let

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \tau_{ij} \quad [20]$$

$$U_k = \mu - k \quad [21]$$

and

$$V_k = (2S - k(k - 1))/2 . \quad [22]$$

Then,

$$\rho_k \geq \frac{2V_{k-1} - (k - 2)U_{k-1}}{n(n - k + 1)} . \quad [23]$$

If $2V_{k-1} < (n + k - 2)U_{k-1}$, then

$$\rho_k \geq \frac{2((k^* - 1)U_{k-1} - V_{k+1})}{(k^* - k)(k^* - k + 1)} \quad [24]$$

where $k^* + k - 3$ is the largest integer in $2V_{k-1}/U_{k-1}$. Two upper bounds on ρ_k are also given. The first is

$$\rho_k \leq 1 + ((n + k - 1)U_k - 2V_k)/kn \quad [25]$$

and the second is that if $2V_k < (k - 1)U_k$,

$$\rho_k \leq 1 - 2 \frac{(k^* - 1)U_k - V_k}{(k - k^*)(k - k^* + 1)} \quad [26]$$

where $k^* + k - 1$ is the largest integer in $2V_k/U_k$.

An Illustration

To illustrate how ρ_k might be applied and interpreted, observations on seven items were analyzed according to the procedure outlined above. Each item had two distractors, and they were found to be consistent with the assumptions of the answer-until-correct scoring model. (See Wilcox, 1981a.) Table 1 shows the observed frequencies for the first two items. The question to be answered is if these seven items are taken to be the whole test, do they give reasonably accurate information about what the typical examinee knows?

As previously mentioned, the model described above implies that various inequalities among the p_{ij} 's must hold. These inequalities were tested at the .25 level of significance with the procedure in Robertson (1978). In every case the observed responses were consistent with the model.

Generally, when estimating ζ_{22} there is no need to estimate all of the ζ 's in equations 8-16. For the situation at hand, ζ_{22} can be estimated as follows. First compute

$$\hat{\zeta}_{02}/3 = \hat{p}_{31} - \hat{p}_{32} \quad [27]$$

for the data in Table 1, this is .107. Next compute

$$\hat{\zeta}_{12}/2 = \hat{p}_{21} - \hat{p}_{22} - \hat{\zeta}_{02}/3 \quad [28]$$

which is .074. Then

$$\hat{\zeta}_{22} = \hat{p}_{11} - \hat{p}_{12} - \hat{\zeta}_{12}/2 - \hat{\zeta}_{02}/3 \quad [29]$$

which is equal to .225. Substituting these values into equation 19, the

estimate of τ_{12} is $\hat{\tau}_{12} = .75$. Applying equation 6 to all seven items, it is seen that $\mu = 5.434$. In other words, it is estimated that the expected number of correct decisions is 5.434.

Next consider ρ_5 . The value of S was estimated to be 16.929. From equations 20 - 26, this implies that

$$.42 \leq \rho_5 \leq .74 . \quad [30]$$

This analysis suggests that these seven items, taken as a whole, are not very accurate since there is at least a 26 percent chance of making an incorrect decision on three or more items. How should the test be modified? Another important question is to what extent can it be improved? One approach to improving the test is to increase the number of distractors, and another approach is to try to modify or replace the distractors that are being used. The latter approach will be considered first.

The initial step in trying to decide whether to replace or modify the existing distractors is to determine the extent to which they can be improved. This can be done with the Δ measure in Wilcox (1981, eq. 20). This measure is just the difference between the maximum possible value of τ and the estimated value given that $\tau_2 = \hat{\tau}_2$. Another related measure is the entropy function (see Wilcox, 1981a). This measures the effectiveness of the distractors among the examinees who do not know the correct response by indicating the extent to which p_2, \dots, p_t are unequal. The closer they are to being equal, the more effective are the distractors, i.e., guessing is closer to being random. It has been pointed out (Wilcox, 1981a) that Δ might be objectionable as a measure of the extent to which p_2, \dots, p_t

are equal, but for present purposes it would seem to be of interest because increasing ρ_k depends on the extent to which τ can be increased for each item.

Referring to Wilcox (1981a), a little algebra shows that for the case $t = 3$,

$$\Delta = (p_2 - p_3)/2 . \quad [31]$$

For item 1 in Table 1, $\Delta = .024$, and for item 2 it is .034 (Δ is assumed to be positive; so if $p_2 < p_3$, apply the pool-adjacent violator algorithm in which case Δ is estimated to be zero.)

If the number of alternatives for item 1 is increased to $t = 5$, and if guessing is at random, then the value of τ would be .893 which represents an increase of .126 over the value of τ using the existing distractors. Thus, it would seem that one approach to improving item 1 is to find two more distractors that are about as effective as the two being used. In practice, of course, this might be very difficult to do.

Estimating τ_{ij} When There Is Misinformation

Among the 30 items analyzed by Wilcox (1982a), the observed test scores suggest that two of the items do not conform well to the AUC scoring model described in a previous section. Thus, the proposed estimate of τ_{ij} is inappropriate. This section outlines how this problem might be solved when a misinformation model appears to be more appropriate for some of the items on the test.

Consider a test item with t alternatives, and let τ_t be the propor-

tion of examinees who eliminate the correct response from consideration on their first attempt of the item. (An AUC scoring procedure is being assumed.) Once an examinee eliminates all of the distractors that are consistent with his/her misinformation, it is assumed that the examinee chooses the correct response on the next attempt. This assumption is made here because it seems to give a good approximation to how examinees were behaving on the items used in Wilcox (1982a). It is also assumed that if an examinee does not know and does not have misinformation, then he/she guesses at random among the t alternatives. Finally, for examinees with misinformation, assume that they believe the correct response is one of c alternatives that are in actuality incorrect. Thus, examinees with misinformation will require at least $c + 1$ attempts before getting the item correct. As an illustration, consider $t = 5$ and $c = 3$. Then,

$$p_1 = \zeta_{t-1} + \zeta_{t+1}/5 \quad [32]$$

$$p_2 = \zeta_{t+1}/5 \quad [33]$$

$$p_3 = \zeta_{t+1}/5 \quad [34]$$

$$p_4 = \zeta_t + \zeta_{t+1}/5 \quad [35]$$

$$p_5 = \zeta_{t+1}/5 \quad [36]$$

where ζ_{t+1} is the proportion of examinees who do not know and who do not have misinformation.

Various modifications of the model are, of course, possible and presumably this model (with some appropriately chosen c value) will give a good fit to the observed test scores. For illustrative purposes, equations

32 - 36 are assumed. The point of this section is that it is now possible to again estimate τ_{ij} where the misinformation model is assumed to hold for one or both of the items in any item pair. Note that for a single item where equations 32 - 36 hold,

$$\tau = \zeta_{t-1} + \zeta_{t+1}/t. \quad [37]$$

To estimate τ_{ij} , the joint probability of making a correct decision on a pair of items where, say, the first item is represented by a misinformation model, equation 7 must be rederived. Accordingly, let t' be the number of alternatives on the first item, and t is the number of alternatives on the second. The misinformation model assumes that on the first attempt of the item, examinees belong to one of three mutually exclusive categories, namely, they know the answer and choose it, they have misinformation and eliminate the correct response, or they do not know and guess at random. Thus, using previously established notation, equation 8 becomes,

$$p_{11} = \zeta_{42} + \zeta_{41}/2t' + \zeta_{40}/3t' + \zeta_{02}/t' + \zeta_{01}/2t' + \zeta_{00}/3t' \quad [38]$$

where, in this illustration, $t' = 5$. There is no ζ_{i3} term ($i = 0, 1, 2$) because the misinformation model assumes that if examinees do not know, they cannot eliminate any of the distractors. More generally,

$$p_{11} = \zeta_{t'-1,t-1} + \sum_{j=0}^{t-1} \zeta_{t'-1,j}/(t-j)t' + \sum_{j=0}^{t-1} \zeta_{0j}/(t-j)t'. \quad [39]$$

Also,

$$p_{k1} = p_{11} - \zeta_{42} \quad (k = 2, \dots, t') \quad [40]$$

$$p_{12} = \zeta_{41}/2t' + \zeta_{40} \quad [41]$$

$$p_{1m} = \sum_{j=0}^m \zeta_{4j}/(t-j)t' \quad (m = 0, \dots, t-2). \quad [42]$$

The remaining p_{ij} values can be determined in a similar manner. For the two items being used here

$$p_{2m} = \sum_{j=0}^m \zeta_{0j} (t - j)t^j \quad (m = 2, \dots, t) \quad [43]$$

and $p_{3m} = p_{2m}$.

The expressions for p_{4m} and p_{5m} involve the proportion of examinees who have misinformation on the first item. The necessary equations can be derived as was illustrated above. This in turn yields an estimate of τ which can be used to estimate the bounds on p_k .

Testing Whether Items are Equivalent or Hierarchically Related

The model described in this paper might also be useful when empirically checking the assumptions of other latent structure models. For example, Macready and Dayton (1977) and Wilcox (1977) propose models where it is assumed that pairs of equivalent items are available. Two items are defined to be equivalent if examinees either know both or neither one. When equivalent items are available, the proportion of examinees who know both can be estimated (assuming local independence). Macready and Dayton checked their model with a chi-square goodness-of-fit test, but this requires at least three items that are equivalent to one another. (When there are only two items, there are no degrees of freedom left.)

For illustrative purposes, assume $t=3$, and consider equations 8-16. If two items are equivalent, then

$$\zeta_{21} = \zeta_{20} = \zeta_{12} = \zeta_{02} = 0 \quad [44]$$

$$p_{12} = p_{21} = p_{22} \quad [45]$$

$$p_{13} = p_{23} \quad [46]$$

and

$$p_{31} = p_{23} \quad [47]$$

For $N \leq 50$, an exact test of these last three qualities can be made using the critical values in Katti (1973) and Smith et al. (1979). (Note that the conditional distribution of multinomial random variables is multinomial.) For larger N , the usual chi-square test can be used. From Smith et al. (1979), a slight adjustment to the usual chi-square test appears to be useful. Finally, if one of these items is assumed to be hierarchically related to the other, again certain equalities must hold among equations 8-16, and this can again be tested (cf. While & Clark, 1973; Dayton & Macready, 1976).

A Concluding Remark

It should be stressed that ρ_k is of interest after it has been decided which items are to be included on a test. ρ_k is not intended to measure validity -- it is designed to measure the overall effectiveness of the distractors that are being used. Put another way, ρ_k is not meant to be the one and only index for characterizing a test -- it is intended to be one of several indices that might be used. The reason for raising this issue is that a test constructor can ensure that ρ_k is large by using easy items. This is an improper procedure that misses the point of how ρ_k is to be used.

Table 1

Number of Examinees Requiring i Attempts on Item 1
and j Attempts on Item 2

		Number of Attempts on Item 2			Total
		1	2	3	
Number of Attempts on Item 1	1	179	26	14	219
	2	76	8	4	88
	3	53	13	4	70
	Total	308	47	22	377

References

- Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. Statistical inference under order restrictions. New York: Wiley, 1972.
- Bergan, J. R., Cancelli, A. A., & Luiten, J. W. Mastery assessment with latent class and quasi-independence models representing homogeneous item domains. Journal of Educational Statistics, 1980, 5, 65-81.
- 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.
- Brownless, V. T., & Keats, J. A. A retest method of studying partial knowledge and other factors influencing item response. Psychometrika, 1958, 23, 67-73.
- 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.
- Dayton, C. M., & Macready, G. B. A scaling model with response errors and intrinsically unscalable respondents. Psychometrika, 1980, 45, 343-356.
- Katti, S. K. Exact distribution for the chi-square test in the one way table. Communications in Statistics, 1973, 2, 435-447.

- Knapp, T. R. The reliability of a dichotomous test-item: A 'correlationless' approach. Journal of Educational Measurement, 1977, 14, 237-252.
- Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.
- Macready, G. B., & Dayton, C. M. The nature and use of state mastery models. Applied Psychological Measurement, 1980, 4, 493-516.
- Marks, E., & Noll, G. A. Procedures and criteria for evaluating reading and listening comprehension tests. Educational and Psychological Measurement, 1967, 27, 335-348.
- Molenaar, I. On Wilcox's latent structure model for guessing. British Journal of Mathematical and Statistical Psychology, 1981, 34, 224-228.
- Robertson, T. Testing for and against an order restriction on multinomial parameters. Journal of the American Statistical Association, 1978, 73, 197-202.
- Sathe, Y. S., Pradhan, M., & Shah, S. P. Inequalities for the probability of the occurrence of at least m out of n events. Journal of Applied Probability, 1980, 17, 1127-1132.
- Smith, P. J., Rae, D. S., Manderscheid, R. W., & Silbergeld, S. Exact and approximate distributions of the chi-square statistic for equiprobability. Communications in Statistics -- Simulation and Computation, 1979, B8, 131-149.
- White, R. T., & Clark, R. M. A test of inclusion which allows for errors of measurement. Psychometrika, 1973, 38, 77-86.

- Wilcox, R. R. New methods for studying stability. In C. W. Harris, A. Pearlman, & R. Wilcox, Achievement test items: Methods of Study. CSE Monograph No. 6, Los Angeles: Center for the Study of Evaluation, University of California, 1977. (a)
- Wilcox, R. R. New methods for studying equivalence. In C. W. Harris, A. Pearlman, & R. Wilcox, Achievement test items: Methods of study. CSE Monograph No. 6, Los Angeles: Center for the Study of Evaluation, University of California, 1977. (b)
- Wilcox, R. R. Solving measurement problems with an answer-until-correct scoring procedure. Applied Psychological Measurement, 1981, 5, 399-414. (a)
- Wilcox, R. R. Recent advances in measuring achievement: A response to Molenaar. British Journal of Mathematical and Statistical Psychology, 1981, 34, 229-237. (b)
- Wilcox, R. R. Some empirical and theoretical results on an answer-until-correct scoring procedure. British Journal of Mathematical and Statistical Psychology, 1982, 3(5), 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, 1983, 42, 153-165. (a)
- Wilcox, R. R. Determining the length of multiple-choice criterion-tests when an answer-until-correct scoring procedure is used. Educational and Psychological Measurement, 1983, 42, 789-797. (b)