AN APPROACH TO MEASURING THE ACHIEVEMENT OR PROFICIENCY OF AN EXAMINEE

Rand R. Wilcox

CSE Report No. 148

October, 1980

Test Design Project
Center for the Study of Evaluation
Graduate School of Education, UCLA
Los Angeles, California 90024

Table of Contents

PREFACE

A part of our goal at CSE has been to develop new and improved psychometric techniques to study, develop and characterize achievement tests and achievement test items. Recently our efforts have been focused on certain errors that occur when using criterion-referenced tests. In particular, we have investigated problems related to estimating and controlling the false-positive and false-negative error rates associated with a test and a population of examinees. In other words, we are concerned about passing those examinees who should pass, and retaining those examinees who need remedial work. This paper deals with one aspect of that problem.

ABSTRACT

Throughout the United States, various school systems are developing what is referred to here as proficiency tests. These tests are conceptualized as representing a variety of skills with one or more items per skill. One purpose of the test might be to determine whether a student will receive a high school diploma. This paper discusses how certain recent technical advances might be extended to examine these tests. In contrast to existing analyses, errors at the item level are included. It is shown that inclusion of these errors implies that a substantially longer test might be needed. One approach to this problem is described and directions for future research are also suggested.

INTRODUCTION

Throughout the United States efforts are being made to develop tests to measure the proficiency of students attending the local schools. In some cases, these tests are used to determine whether a student will be awarded a high school diploma; in other instances, they might be used to decide whether an examinee should be advanced to the next grade level. In some instances these tests are conceptualized and constructed as follows: a group of teachers, parents, content experts and other interested parties work together to identify those skills that are believed to be a basic part of a student's education. For example, interest might focus on competency in mathematics, in which case, the skills might include addition, subtraction, computing percentages, etc. Corresponding to each skill, test items are constructed for the purpose of determining whether an examinee has acquired the skill in question. Here it is assumed that these test items have been examined for any ambiguities or misrepresentations and that appropriate corrections have been taken when necessary.

Because of the large number of skills that have been identified, it is impractical to test an examinee on every one. Accordingly, a random sample of skills is used to make inferences about the proportion of skills that an examinee has acquired. The test administered to an examinee consists of items that represent the skills. Decisions concerning proficiency are made according to some predetermined passing score. For example, a requirement for receiving a high school diploma might include taking a mathematics test and successfully answering 70% of the items or

demonstrating mastery of 70% of the skills. Note that these two decisions are not necessarily equivalent. As a simple illustration, imagine a test of 10 skills with 3 items per skill for a total of 30 items. Further suppose that a mastery decision is made for a particular skill if the examinee responds correctly to two out of the three corresponding items. In other words, an allowance is being made for the possibility that an examinee has acquired the skill but gives an incorrect response because of some distraction or carelessness, for instance. In this case, it is possible (but perhaps unlikely) that an examinee will get less than 70% of the skills.

The purpose of this paper is to demonstrate how certain recent technical advances can be extended and applied to the type of test described above. Emphasis is given to the problem of determining how many skills to include on a test. As will become evident, the analysis has implications about how many items to use per skill. In the case of multiple-choice test items, there are also possible implications about the number and quality of the distractors that are being used.

Before continuing, it is of interest to observe that the situation considered here is similar to a common conceptualization of a mastery test. A mastery test is frequently regarded as consisting of items randomly sampled from some larger item pool (e.g., Wilcox, 1977; Harris, 1974; Novick and Lewis, 1974; Huynh, 1976). The item domain might exist de facto or it might be a convenient conceptualization. Based on this "item sampling" view, the binomial error model (Lord and Novick, 1968, Chapter 23) is then used to describe the observed responses of the

examinees. In particular, the probability function of x, the observed (number correct) score of an examinee, is given by

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} \qquad (1)$$

where p is referred to as the examinee's percent correct true score. The goal of the test is to determine whether p is above or below a known constant p_0. The main difference between mastery tests and the present situation is that here we take the view that skills, not items, are being sampled and that there might be more than one item per skill. Moreover, the analysis given here includes errors at the item level while for the binomial error model, these errors are ignored. For the case in which only one skill is being examined in terms of a population of examinees, the reader is referred to Macready and Dayton (1977).

Let ζ be the proportion of skills that an examinee knows. Consistent with the approach to mastery tests, it is assumed that the goal of a proficiency test is to determine whether ζ is above or below a known constant, ζ_0. Before describing the main results on solving this problem, we give a more precise description of the framework within which we propose to work.

Some Definitions

Consider a specific, randomly selected skill and let k be the number of items used to determine mastery of this skill. For each of these k items it is assumed that an examinee who has mastered the skill might give an incorrect response because of a momentary distraction, carelessness, or some other reason. Let $\alpha_i (i=1,\ldots,k)$ be the probability of this event for the

ith item. In a similar manner, let β_i be the probability of not knowing and guessing the correct response to the ith item. Note that α_i and β_i are both conditional probabilities. Finally, a mastery decision is made for the skill if y, the number correct out of the k items associated with the skill, is greater than or equal to a specified passing score y_0.

It should be mentioned that the framework described above is similar to a number of models proposed by various authors to describe tests (e.g., Wilcox, 1979b; Macready and Dayton, 1977; Brownless and Keats, 1958; Marks and Noll, 1967; Knapp, 1977). Macready and Dayton (1977, p. 100) imply that their model is appropriate when mastery of a skill is an all-or-none process. However, as noted by Wilcox (1979b), this does not mean that an all-or-none view of learning is required in order to use their model.

Macready and Dayton (1977) use a more general family of decision rules for determining mastery of a particular skill. Their decision rule is defined in terms of a particular skill and a population of examinees while here, at least for the moment, the emphasis is on making a decision for a specific examinee in terms of a particular randomly selected skill. It is readily seen, therefore, that their decision rule does not apply to the present situation.

Finally, let the vector $\underline{y}=(y_1,\ldots,y_k)$ be a sequence of 1's and 0's designating a particular response pattern of corrects and incorrects on the k items where a 1 means a correct, and a 0 an incorrect response.

Based on the above definitions, and for the assumption of local independence (Lord and Novick, 1968, section 16.3), it follows that the probability of a mastery decision for the skill is

$$\begin{aligned} &\Pr(y \geq y_0 \mid \text{mastery of the skill}) \\ &\quad = \xi_1 \text{ (say)} \\ &\quad = \sum_{\underline{y}:y\geq y_0} \prod_{i=1}^{k} (1-\alpha_i)^{y_i} \alpha_i^{1-y_i}. \end{aligned} \tag{2}$$

where the summation is over all vectors $\underline{y}$ such that $y \geq y_0$. In addition,

$$\begin{aligned} &\Pr(y \geq y_0 \mid \text{nonmastery of the skill}) \\ &\quad = \xi_2 \text{ (say)} \\ &\quad = \sum_{\underline{y}:y\geq y_0} \prod_{i=1}^{k} \beta_i^{y_i} (1-\beta_i)^{1-y_i}. \end{aligned} \tag{3}$$

If, as in Macready and Dayton's model II, it is assumed that $\alpha_i=\alpha$ and $\beta_i=\beta$ for $i=1,\ldots,k$, then equations 2 and 3 take on the more familiar form of the binomial probability function, namely,

$$\xi_1 = \sum_{y=y_0}^{k} \binom{k}{y}(1-\alpha)^y \alpha^{k-y} \tag{4}$$

and

$$\xi_2 = \sum_{y=y_0}^{k} \binom{k}{y} \beta^y (1-\beta)^{k-y}. \tag{5}$$

A Conservative Solution to the Problem of Determining the Number of Skills to Include on the Test

So far, we have merely laid the ground work for handling certain technical problems associated with so-called proficiency tests. In this section, we consider the determination of how many skills to include on the test. The analysis is made in terms of a single examinee.

For a randomly selected skill, the probability of a mastery decision is

$$\gamma = \xi_1 \zeta + \xi_2(1-\zeta). \tag{6}$$

Thus, the probability of x mastery decisions among n randomly selected skills is

$$\binom{n}{x} \gamma^x (1-\gamma)^{n-x}. \tag{7}$$

Let x_0 be the passing score for the test. In other words, the decision $\zeta \geq \zeta_0$ is made if $x \geq x_0$; if $x < x_0$, the reverse is said to be true. Here, it is assumed that x_0 is the smallest integer such that $x_0/n \geq \zeta_0$.

The goal is to find a conservative solution to the choice for n. In particular, we want to choose the smallest n so that the probability of a correct decision (CD) is reasonably close to one regardless of the actual value of ζ. To solve this problem, it is necessary for the investigator to specify an additional constant, $\delta^* > 0$. The idea is that if $\zeta \leq \zeta - \delta^*$ or if $\zeta \geq \zeta_0 + \delta^*$, we want to choose the smallest n so that

$$\Pr(CD) \geq P^*, \ 1/2 < P^* < 1. \tag{8}$$

If, however, $\zeta_0-\delta^*<\zeta<\zeta_0+\delta^*$, either decision is said to be correct. The open interval $(\zeta_0-\delta^*, \zeta_0+\delta^*)$ is called the indifference zone. The situation is similar to the one considered by Fhaner (1974) and Wilcox (1979a). Here, however, we are taking into account the errors represented by the probabilities α_i and β_i that are associated with each skill. We note that if $\delta^*=0$, it may be impossible to find an n that satisfies equation 8 for all possible values of ζ. For a more extensive discussion of the indifference zone approach to statistical problems (including the choice of δ^*), the reader is referred to Gibbons, Olkin and Sobel (1977). Further comments on the choice of δ^* are made below. In particular, it is shown that $\delta^*>0$ is a necessary, but not a sufficient condition, for solving the problem at hand.

solution to the choice of n, i.e., an n that satisfies equation 8 regardless of the value of ξ_1 or ξ_2, we need lower bounds to both ξ_1 and ξ_2. Here it is assumed that there is no data available for estimating ξ_1 and ξ_2. Thus, the investigator must specify (using nonstatistical techniques) lower bounds to ξ_1 and ξ_2 that are consistent with the types of items being used. In practice, this might be done by specifying an upper bound to α and a lower bound to β and using equations 4 and 5. This is illustrated below.

For $\zeta \leq \zeta_0 - \delta^*$ it can be seen that we require $\gamma < \zeta_0$ which implies that we must have

$$\delta^* > \zeta_0 - \frac{\zeta_0 - \xi_2}{\xi_1 - \xi_2} . \tag{12}$$

In summary, we can guarantee that the probability of a correct decision is at least P*, if equations 11 and 12 are satisfied, by choosing the smallest n so that both equations 9 and 10 are greater than, or equal to, P*. As for ξ_1 and ξ_2, this time we set $\zeta = \zeta_0 - \delta^*$ and use upper bounds to these two quantities. In contrast to the case $\zeta \geq \zeta_0 + \delta^*$, this might be accomplished by specifying a lower bound to α and an upper bound to β and again using equations 4 and 5.

An Illustration with k=1

Consider a situation in which a single item (k=1) is used to measure each skill and suppose $\zeta_0 = .8$, $\delta^* = .1$ and $P^* = .90$. For this special case, $\xi_1 = 1-\alpha$ and $\xi_2 = \beta$ (assuming, of course, $y_0 = 1$). Consider the case $\zeta \leq \xi_0 - \delta^*$. As previously explained, the Pr(CD) given by equation 9

From Wilcox (1979a) it follows that the smallest n, so that (10) has a value of at least P*=.9, is given approximately by

$$n = \lambda^2 \zeta_0(1-\zeta_0)/(\gamma_1-\zeta_0)^2 \tag{15}$$

where λ is the P* quantile of the standard normal distribution and γ_1 is the value of γ when $\zeta=\zeta_0+\delta^*$. With $\xi_1=.9$ and $\xi_2=.15$,

$$n = (1.28)^2\ (.8)\ (.2)/(.825-.8)^2 \tag{16}$$

$$= 419.$$

As for (11), the smallest n is given approximately by

$$\lambda^2 \zeta_0(1-\zeta_0)/(\zeta_0-\gamma_2)^2 \tag{17}$$

where γ_2 is the value of γ when $\zeta=\zeta_0-\delta^*$. In our illustration, we have that n=2621. Thus, n=2621 skills would be used.

It is evident that for practical purposes, n=2621 is unacceptable. Suppose, instead, we have completion items in which case guessing is virtually ruled out. For illustrative purposes, suppose $\beta = 0$, which appears to be approximately true for the test data examined by Macready and Dayton (1977), and that $0 \leq \alpha \leq .02$. In this case $\gamma_1=.882$, $\gamma_2=.7$ and n=39. If $0 \leq \alpha \leq .05$, $\gamma_1=.855$, $\gamma_2=.7$ and n=87. If we ignore errors at the item level (i.e., $\xi_1=1$ and $\xi_2=0$), the resulting value of n is approximately 29.

An Illustration with k=3

The second illustration is the same as the first, except that we assume there are k-3 items per skill. The primary purpose of this illustration is to see how much we can reduce the required number of items by

increasing k. As before, it is assumed that $.15 \leq \beta \leq .3$ and $0 \leq \alpha \leq .1$.

With $\alpha=.1$ and $\beta=.3$ and with a mastery decision for particular skill being made when the examinee gets at least 2 of the 3 items correct (i.e., $y_0=2$), expressions (3) and (4) yield $\xi_1=.972$ and $\xi_2=.216$. When $\alpha=0$ and $\beta=.15$, $\xi_1=1$ and $\xi_2=.06$. Thus, for $\zeta=\zeta_0-\delta^*$ we use $\xi_1=1$ and $\xi_2=.216$ implying that $\gamma=.7648$. Hence

$$Pr(CD) = \sum_{x=0}^{x_0} \binom{n}{x} .7648^x \binom{n}{x} .7648^x (.2352)^{n-x}. \qquad (18)$$

As for $\zeta=\zeta_0+\delta^*$, $\gamma=.88$ and

$$PR(CD) = \sum_{x=x_0}^{n} \binom{n}{x} .88^x .12^{n-x}. \qquad (19)$$

It follows that the smallest number of skills required is approximately n=212. The exact value was calculated on an IBM 360/91 computer and found to be n=219. Thus, the total number of items is decreased considerably but we would still need over 600 items on the test.

An Illustration with Tighter Bounds on α and β

To illustrate the effect of having tighter bounds on α and β, we suppose $.0 \leq \alpha \leq .02$ and $.2 \leq \beta \leq .3$ and we set $y_0=3$. Otherwise the situation is assumed to be the same as in the previous illustration. In this case, $\gamma=.848$ when $\zeta=.9$, $\xi_1=.941$ and $\xi_2=.008$. Also, $\gamma=.7027$ when $\zeta=.7$, $\xi_1=1$ and $\xi_2=.027$. It follows that the minimum n required is approximately 114. Thus, to guarantee that the probability of a correct decision is at least .9, a total of 3(114)=342 items would be used.

Retrospective Studies Using Latent Structure Models

The illustrations in the previous section demonstrate rather dramatically, that including errors at the item level might have a substantial effect on the number of items used on the test. Moreover, even with "tight" bounds on the parameters α and β, an extremely large number of items might be required. Several approaches to this problem might be used. For example, there might be a more optimal choice for k, the number of items per skill. In the case of multiple-choice items, one might consider increasing the number of distractors (cf. Lord, 1977). In this section we outline still another approach which is based on latent structure models. The approach represents a slight extension of one used by Wilcox (1979c). In contrast to the earlier sections of the paper, it is now assumed that data exists for a random sample of N examinees who have taken a test consisting of n skills with $k \geq 3$ items per skill. The reason for the restriction on k is explained below. An additional difference from previous sections is that we examine the accuracy of the test in terms of comparing ζ to ζ_0 for the typical or "average" examinee among those being tested. This alternative perspective does not affect the results previously described. If an examinee's true score is close to ζ_0, an extremely large number of items might be needed to accurately determine whether ζ is above or below ζ_0. In some situations, an investigator might also be interested in the accuracy of a test in terms of a population of examinees; for example, all the students attempting to graduate from high school. It may be that most examinees have a true score that is not close to ζ_0 or perhaps most true scores fall within

For the reasons given by Wilcox (1979b), $\hat{\mu}$ and $\hat{\sigma}^2$ may be used to estimate the mean, μ, and variance, σ^2, of the true score distribution.

Let

$$\tau_1 = \mu, \text{ if } \mu<\zeta_0-\delta^* \\ = \zeta_0-\delta^*, \text{ if } \zeta_0-\delta^*\leq\mu\leq 1. \tag{23}$$

$$m_1 = \max\left[\mu(\zeta_0-\delta^*-\mu), (\mu-\zeta_0+\delta^*)(1-\mu)\right]. \tag{24}$$

$$\phi_1 = \frac{\sigma^2}{\sigma^2+(\tau_1-\mu)^2}, \text{ if } 0<\sigma^2\leq m_1 \\ = (\mu(1-\mu)-\sigma^2)/(1-\zeta_0+\delta^*)(\zeta_0-\delta^*), \text{ otherwise.} \tag{25}$$

$$m_2 = \max\left[\mu(\zeta_0+\delta^*-\mu), (\mu-\zeta_0-\delta^*)(1-\mu)\right] \tag{26}$$

$$\tau_2 = \zeta_0+\delta^*, \text{ if } \mu<\zeta_0+\delta^* \\ = \mu, \text{ if } \zeta_0+\delta^*\leq\mu\leq 1. \tag{27}$$

$$\phi_2 = \frac{\sigma^2}{\sigma^2+(\tau_2-\mu)^2}, \text{ if } 0<\sigma^2\leq m_2 \\ = (\mu(1-\mu)-\sigma^2)/(1-\zeta_0-\delta^*)(\zeta_0+\delta^*), \text{ otherwise.} \tag{28}$$

Following Wilcox (1979c), results reported by Skibinsky (1977) can be applied to show that for $\varepsilon_1=\Pr(x\geq x_0, \zeta\leq\zeta_0)$, the probability of a false-positive decision, we have the inequality

$$\varepsilon_1 \leq \phi_1 \sum_{x=x_0}^{n} \binom{n}{x} \gamma_1^{\,x} (1-\gamma_1)^{n-x}. \tag{29}$$

where γ_1 is the value of γ when $\zeta=\zeta_0+\delta^*$. As in the previous section, it is assumed that for a specific examinee, the probability of getting x mastery decisions is given by the binomial probability function (cf. Lord and Novick, Chapter 23). As for the probability of a false-negative decision, say ε_2, it can be seen that

$$\varepsilon_2 \leq \phi_2 \sum_{x=0}^{x_0-1} \binom{n}{x} \gamma_2^x (1-\gamma_2)^{n-x} \tag{30}$$

where γ_2 is the value of γ when $\zeta=\zeta_0-\delta^*$.

To illustrate the above inequalities, we consider a situation similar to the one described in the second example of the previous section. In particular, we suppose k=3, $\zeta_0=.8$, $0\leq\alpha\leq.1$ and $.15\leq\beta\leq.3$. Further suppose that μ and σ^2 are estimated to be .75 and .10, respectively. Thus, $\tau_1=.7$, $m_1=.0125$, $\phi_1=.417$, $\tau_2=.9$, $m_2=.1125$ and $\phi_2=.4$. Hence,

$$\varepsilon_1 \leq .417 \sum_{x=x_0}^{n} \binom{n}{x} .7648^x .2352^{n-x} \tag{31}$$

and

$$\varepsilon_2 \leq .4 \sum_{x=0}^{x_0-1} \binom{n}{x} .88^x .12^{n-x}. \tag{32}$$

The smallest number of skills so that, simultaneously, $\varepsilon_1\leq.1$ and $\varepsilon_2\leq.1$, is n=59.

Concluding Remarks

This paper has examined some of the problems that occur when using the proficiency tests currently being developed by many school systems. It is evident that more investigations need to be made. As previously indicated, we need to have better methods for determining the optimal number of distractors per multiple-choice item and the optimal number of items per skill. Several other questions also occur.

It has been argued that, in terms of measuring achievement, a test should be constructed using an item sampling principle (e.g., Harris,

Pearlman and Wilcox, 1977). The authors' experience with people constructing proficiency tests is that this approach is, indeed, used in many cases. However, as pointed out by a referee, there is also the problem that, frequently, a test does not consist of a random sample of skills but rather skills are selected because they are judged to be the most important of those available. In this case, the efficacy of using the test length solution presented here might be in doubt. Alternatively, one might define "proficiency" in terms of a hypothetical domain of skills where only the most important skills are represented in the item pool. In this case, an item sampling view of the test might be acceptable and so the test length solution can be applied. We note that arbitrarily imposing the binomial error model has yielded good results using real data for certain measurement problems (e.g., Keats and Lord, 1962; Lord, 1965; Subkoviak, 1978) but that in terms of test length, the extent to which we obtain good results is not clear.

Another important point to keep in mind is that the test length solution is highly sensitive to the values of α and β. As was demonstrated, if completion items are used and $\beta=0$, a reasonably small number of items might be required even when our conservative solution to determining test length is applied. In many situations, there is the practical difficulty of physically scoring completion items and so multiple-choice items typically are used. Accordingly, it would be beneficial to have some procedure that corrects for the errors α and β in such a way that not too many multiple-choice items would be needed to ensure a reasonably high probability of making a correct decision for an examinee. For

example, we might use the usual correction-for-guessing formula score which assumes guessing is at random. In many cases, guessing is not at random but perhaps this approach will still require fewer items than would otherwise be needed. Several other possibilities are currently being investigated; the results will appear in a forthcoming paper.

References

Anderson, T. W. On estimation of parameters in latent structure analysis. Psychometrika, 1954, 19, 1-10.

Brownless, V. T. & Keats, J. A. A retest method of studying partial knowledge and other factors influencing item response. Psychometrika, 1958, 23, 67-73.

Fhanér, S. Item sampling and decision making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 27, 172-175.

Formann, A. K. A note on parameter estimation for Lazarsfeld's latent class analysis. Psychometrika, 1978, 43, 123-126.

Gibbons, J.; Olkin, I.; & Sobel, M. Selecting and ordering populations: A new statistical methodology. New York: John Wiley, 1977.

Green, B. F. A general solution for the latent class model of latent structure analysis. Psychometrika, 1951, 16, 151-166.

Goodman, L. A. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 1974, 61, 215-231.

Harper, D. Local dependence latent structure models. Psychometrika, 1972, 37, 53-59.

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin and W. James Popham (Eds.). Problems in criterion-referenced measurement. CSE Monograph No. 3, Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Harris, C. W.; Pearlman, A. P.; & Wilcox. Achievement Test Items - Methods of Study. CSE Monograph Series in Evaluation, No. 6, Los Angeles: Center for the Study of Evaluation, University of California, 1977.

Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41, 65-78.

Keats, J. A. & Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Knapp, T. R. The reliability of a dichotomous test-item: A "correlationless" approach. Journal of Educational Measurement, 1977, 14, 237-252.

Lazarsfeld, P. F.; & Henry, N. W. Latent structure analysis. New York: Houghton Mifflin, 1968.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M. Optimal number of choices per item - a comparison of four approaches. Journal of Educational Measurement, 1977, 14, 33-38.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison - Wesley, 1968.

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

Marks, E., & Noll, G. A. "Procedures and criteria for evaluation reading and listening comprehension tests." Educational and Psychological Measurement, 1967, 27, 335-348.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin and W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Skibinsky, M. The maximum probability of an interval when the mean and variance are known. Sankhya, 1977, Ser. A, 39, 144-159.

Subkoviak, M. J. Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 1978, 15, 111-116.

Wilcox, R. R. Estimating the likelihood of a false-positive or false-negative decision with a mastery test: An empirical Bayes approach. Journal of Educational Statistics, 1977, 2, 289-307.

Wilcox, R. R. Achievement tests and latent structure models. British Journal of Mathematical and Statistical Psychology, 1979, 32, 61-71(b).

Wilcox, R. R. On false-positive and false-negative decisions with a mastery test. Journal of Educational Statistics, 1979, 4, 59-73(c).

Wilcox, R. R. Applying ranking and selection techniques to determine the length of a mastery test. Educational and Psychological Measurement, 1979, 31, 13-22(a).