ED 454 240 TM 032 838

AUTHOR	Schulz, E. Matthew; Wang, Lin
TITLE	The Classification Accuracy of Shortened versus Full Length Tests with Number Correct Scoring.
PUB DATE	2001-04-00
NOTE	15p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Seattle, WA, April 11-13, 2001).
PUB TYPE	Reports - Evaluative (142) -- Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Ability; *Classification; Item Response Theory; Pass Fail Grading; *Scoring; Test Format; Test Items; *Test Length
IDENTIFIERS	Accuracy; Experts; *Number Right Scoring

ABSTRACT

In this study, items were drawn from a full-length test of 30 items in order to construct shorter tests for the purpose of making accurate pass/fail classifications with regard to a specific criterion point on the latent ability metric. A three-item parameter Item Response Theory (IRT) framework was used. The criterion point on the latent ability metric corresponded to a criterion domain true score (80% correct) established by an expert panel. The shorter tests were compared to the full-length test in terms of classification accuracy. Number correct (NC) scoring was used. It was found that the classification accuracy of shorter tests met or even exceeded that of the full-length test. Results suggest that, in general, a test targeted on a specific level of ability can be about half the length of a test designed to classify examinees with regard to several (five) levels of ability, without compromising classification accuracy. For lower levels of ability, where guessing at difficult items on the test contributes more measurement error than information, tests can be shortened even more. These conclusions are limited to tests in which pass/fail decisions are based on a number correct score. (Author/SLD)

The Classification Accuracy of Shortened versus Full Length Tests With Number Correct Scoring

E. Matthew Schulz and Lin Wang

ACT, Inc.

Paper Presented at the Annual Meeting of

The National Council on Measurement in Education

April, 2001

Seattle, WA

Abstract

In this study, items are drawn from a full-length test of 30 items in order to construct shorter tests for the purpose of making accurate pass/fail classifications with regard to a specific criterion point on the latent ability metric. A three item-parameter IRT framework is used. The criterion point on the latent ability metric corresponds to a criterion domain true score (80% correct), established by an expert panel. The shorter tests are compared to the full length test in terms of classification accuracy. Number correct (NC) scoring is used. We found that the classification accuracy of shorter tests meets or even exceeds that of the full length test. In general, a test targeted on a specific level of ability can be about half the length of a test designed to classify examinees with regard to several (five) levels of ability, without compromising classification accuracy. For lower levels of ability, where guessing at difficult items on the test contributes more measurement error than information, tests can be shortened even more. These conclusions are limited to tests in which pass/fail decisions are based on a number correct score.

In this study, we were interested in constructing shorter versions of a test of Applied Mathematics with a view to maintaining or even improving the accuracy of pass/fail decisions. The test is used to assign level scores to examinees based on their number correct (NC) score. NC scores on the test range from 0 to 30. Level scores range from 0 to 5. There are three parallel forms of this test. On the particular form that we use in this study, the NC score ranges mapped to level scores 0 through 5 are, respectively, [0,11], [12,16], [17,20], [21,25], [25,28], and [29,30]. The mapping of NC scores to level scores on the other forms is very close or identical for all levels on the other forms. The lowest NC score mapped to a given level score is the "cutoff score" for that level. For example, 12 is the NC cutoff score for Level 1.

This test is often used in settings where users want to know only whether the examinee is at or above a specific level of skill. That is, the users might want to classify examinees with regard to being at-or-above Level 3, but are not interested in making further distinctions, such as whether an examinee who meets this criterion is higher than level 3.

This study addresses the very practical task of developing shorter tests, which we will call "single-level tests" (SLT), for this purpose. For security reasons, it was decided that the one SLT for each level should be constructed by drawing on the items within just one of the three available alternate forms. The SLTs would thus collectively expose items from only one test form. (There is no item overlap among the available forms.) By drawing items from just one form, the construction of a given SLT can be viewed as deleting items from that form. Since each SLT is concerned with at-or-above classifications with regard to just one level, the SLTs are comparable in their purpose to a broad array of tests such as licensure and certification tests and formative mastery tests.

Our developmental research on the SLTs is of broad interest for at least two additional reasons. First, the SLTs are similar to testlets that are used in computer-based testing. Options for computer delivery of tests includes the use of pre-constructed forms, or testlets, each of which contains relatively few items. The items on the pre-constructed form(s) might actually be a subset of the items on a given paper-and-pencil test form. Dichotomous decisions, such as which testlet to administer next, if not pass/fail decisions, might be based on the NC scores on these testlets.

Second, the criteria for making at-or-above determinations with the SLTs, involve domain scores. Pass/fail decisions on licensure and certification tests, and on many other kinds of tests as well, are typically made with regard to a criterion true score on a domain of items. The criterion score may be established by any one of several possible standard setting methods such as the modified Anghoff method. For the present set of tests, including the full-length forms as well as the SLTs, content experts decided that mastery of a level should be defined by a criterion true score of 0.8, or 80% correct, or higher on the items representing the level. For the present set of tests, each level is represented by a pool of eighteen items.

Methods

The psychometric framework for mapping NC scores on full length test forms and SLTs to level scores is based on work described in Schulz, Kolen, and Nicewander (1999). This work is based on the 3-PL IRT model, implemented by BILOG. Items from all levels and forms are calibrated to a common scale. The IRT model is used to establish a correspondence between the criterion level-domain scores (80% correct) and points on the θ scale. The criteria for mastery of levels 1 to 5 for the tests in this study correspond to θs of, respectively, -1.44, -.43, .37, 1.49, 2.41. These values define the lower boundaries of levels 1 to 5 on the θ scale. They are denoted, θ^M, M=1,…,5. Level 0 has no lower boundary.

The mean ± 1 standard deviation of the item parameter estimates for items on the form used to construct the SLTs were 1.34 ± .36 for the a parameter (slope), -0.2 ± 1.4 for the b parameter (intercept), and .176 ± .064 for the c parameter (lower asymptote or 'guessing'). Items were ordered on the test approximately by difficulty. Item p-values ranged from .216 to .965. Biserial correlations ranged from .129 to .818. Student ability in the parameter estimation model (θ) is assumed to have an approximate normal (0,1) distribution.

Construction of shortened tests: All shortened tests were assembled by drawing exclusively on the thirty items in the full length test form. For each type of classification (See Table 1), tests of length L, L=4,5,…,29, were constructed by choosing the L items that provided the most information (Lord, 1980, p 21) at the criterion theta (θ^M). The full length test corresponds to L=30. The test information function for NC scoring is (Lord, 1980, p 73):

$$I\{\theta, X\} = \frac{\left(\sum_{i=1}^{L} P_i'\right)^2}{\sum_{i=1}^{L} P_i Q_i} \qquad (1)$$

where $P_i = P_i(\theta)$ = the probability of getting item i correct conditional on θ; $Q_i = 1 - P_i$, and P_i' is the first derivative of P_i with respect to θ (Lord, 1980, p 61).

We realize that incrementally adding items ordered by the value of their NC-information function at the criterion θ to construct longer SLTs does not necessarily produce the best L-item tests for NC scoring (Hulin, Drasgow, and Parsons, 1983). The best L-item test does not necessarily contain all items from the best L-1 item test. The main points of our study, however, do not depend on how the tests were constructed. Rather, we are concerned with the classification accuracy of shorter tests, however constructed, compared with that of the full-length test for specific pass/fail classifications.

Establishing cutoff scores. Let X represent the random number correct score on a test. To find the cutoff score for assigning an examinee a level score of K (K=0,1,..., 5) on a test consisting of L items we found the minimum X that satisfied the following equation:

$$\sum_{i=1}^{L} P_i(\theta) = X, \qquad \theta \geq \theta^M, M=K. \tag{2}$$

For a given X, Equation 2 was solved for θ on the left by the iterative method of half intervals. This method provides a first-order approximation to the maximum likelihood estimate of θ from an NC score (Yen, 1984).

Estimating classification errors. Let K=0,1,...,5 and θ represent respectively the assigned level score and true θ of a given examinee. Let P^+ and P^- represent the predicted proportion of examinees whose classification is too high or too low, given their true θ. A pass/fail classification error occurs when K<M and $\theta \geq \theta^M$ (a false negative

error) and when $K \geq M$ and $\theta < \theta^M$ (a false positive error). The conditional probabilities of false positive and false negative errors are defined separately for each level, M, as (Schulz, Kolen, and Nicewander, 1999):

$$P^{+}(M,\theta) = \sum_{k=M}^{5} P[(K=k) \mid \theta], \qquad \theta < \theta_M, \quad M = 1, \ldots, 5, \tag{3}$$

and

$$P^{-}(M,\theta) = \sum_{k=0}^{M-1} P[(K=k) \mid \theta], \qquad \theta \geq \theta_M, \quad M = 1, \ldots, 5. \tag{4}$$

Marginal error rates were computed by integrating the conditional error rates over a θ-distribution. For each type of classification (See Table 1) we assumed a uniform θ distribution centered at θ^M and having a range of 3. Integrations were performed by quadriture using 31 equally-spaced points.

Results

The lower plot of Figure 1 shows that about half of the items on the full length test contained practically zero information at a θ of –1.44 (the Level 1 critical theta). The upper plot of Figure 1 shows that the test information for the number correct score, conditional on θ=-1.44, peaks at a test length of 15. Adding more items to the test after the 15th decreases information.

Figure 2 shows test information for number correct scoring as a function of θ for two tests: the 16-item test corresponding to one of the points near the peak in Figure 1, and the full-length (30-item) test. The 16-item test contains more information than the 30-item test over a considerable range of θ--from the lower boundary of the θ-distribution we assumed for computing marginal error rates (lowest asssumed θ) up to about –0.4, where the two information curves cross.

Figure 3 shows the conditional probability of being classified as "at or above Level 1" for each test (16-items and 30-items). The probability of passing should be as low as possible below the target θ and as high as possible above the target θ. On this basis, the 16-item test performs better than the 30-item test at all levels of θ, including levels above –0.4, where the 30-item test information function exceeds that of the 16-item test (Figure 2).

Figure 4 shows the theta conditional on the optimal cutscore, as a function of test length—the solution to Equation 2. At first, the cutscore increases one-for-one with increasing test length, but later the same cutscore (e.g., 11) applies to a range of test lengths. For a cutscore of 11, test lengths range from 21 to 25 items. There is an important, within-cutscore trend in Figure 4: the theta conditional on a fixed cutscore, such as 11, decreases as test length increases. The trend for a given cutscore would extent below –1.44 were it not for the rule about choosing cutscores. (This rule is represented by the "$\theta \geq \theta^{M}$" condition on Equation 2 above.

Figure 5 shows marginal classification error rates as a function of test length. Separate plots are shown for false positive, false negative, and total (sum of false positive and false negative) error rates. As expected, the 16-item test has a lower error rates of each type than the 30-item test. Also, the within-cutscore trend noted in Figure 4 above, is reflected by within-cutscore trends in false-positive and false-negative error rates. For a fixed cutoff score, the false negative rate decreases, and the false positive error rate increases with test length.

The following table summarizes the possibilities for shortening the test in any application that requires only one at-or-above classification. For each type of classification, a shortened test is identified by the number of items it contains and its marginal error rate (false positive plus false negative marginal error rates). No other tests for the same type of classification had a lower error rate or contained fewer items. It is seen that the test could be shortened by about half, on average, if one is interested only in making a pass/fail classification with regard to one level of skill.

Table 1: Classification Error Rates for Shortened vs. Full Length Test				
		Number of Items in Shortened Test	Total Error Rate	
Classification	Critical Theta		Shortened Test	Full Length Test (30 items)
≥ Level 1	-1.44	12	.095	.123
≥ Level 2	-0.43	21	.099	.117
≥ Level 3	.37	16	.102	.108
≥ Level 4	1.49	12	.087	.088
≥ Level 5	2.41	4	.142	.150

Table 1 is not meant to suggest that predicted error rates should be the only guide for constructing a test or choosing a test length. But test information may be an insufficient basis for constructing an optimal test, particularly when number-correct scoring is used. For example, compared to the 15-item test, both the 12-item test and the 16-item test had less information at the Level 1 critical theta (See Figure 1), but had lower marginal error rates (See Figure 5). The 16-item test had the same marginal error rate as the 12-item test (.095).

Educational Importance

This research shows that many tests designed to yield pass/fail results, such as licensure and certification exams, could be shortened without negatively impacting classification error rates. Under some circumstances, shortening a paper-and-pencil test could be a reasonable alternative to computerized testing. This research also has implications for the administration of fixed forms, or pre-assembled testlets, by computer, if pass/fail or stop/continue testing decisions are based on number correct scoring.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores. Reading, MA: Addison-Wesley Publishing.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). item response theory: Application to psychological measurement. Homewood, IL: Dow Jones-Irwin.

Lord, F. M. (1980) *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3. Item analysis and test scoring with binary logistic models* (2nd ed.). Mooresville, IN: Scientific Software.

Schulz, E.M., Kolen, M. & Nicewander, W.A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement. 23*, 347-362.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21*, 93-111.

Figure 1: Test and Item Information at Level 1 Critical Theta as Function of Test Length or Item Rank-by-Information (θ = -1.44)

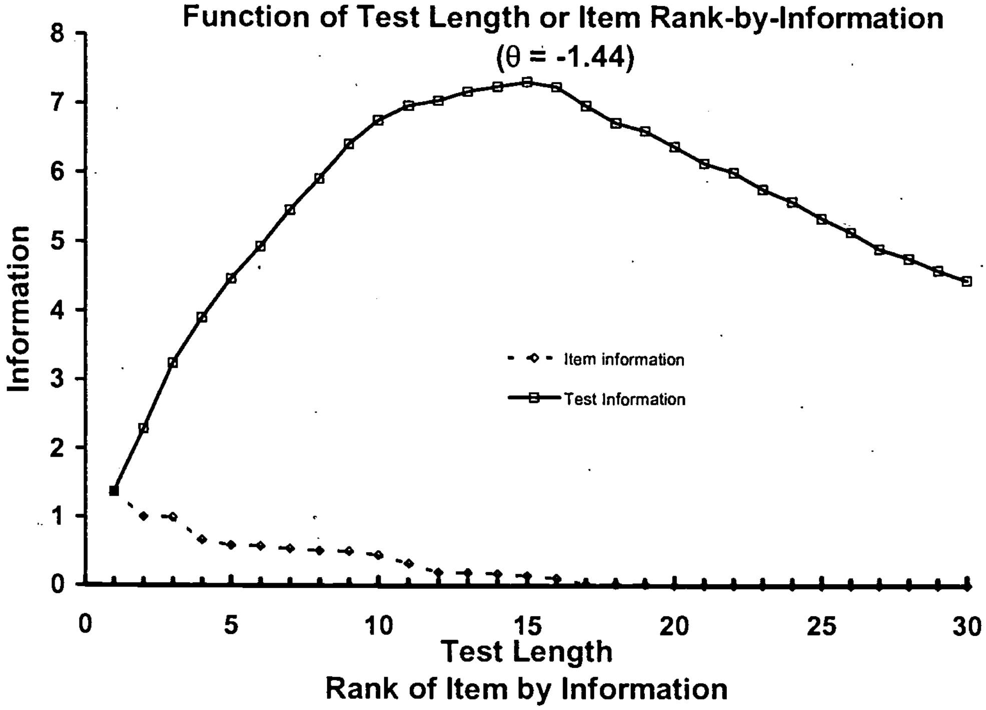
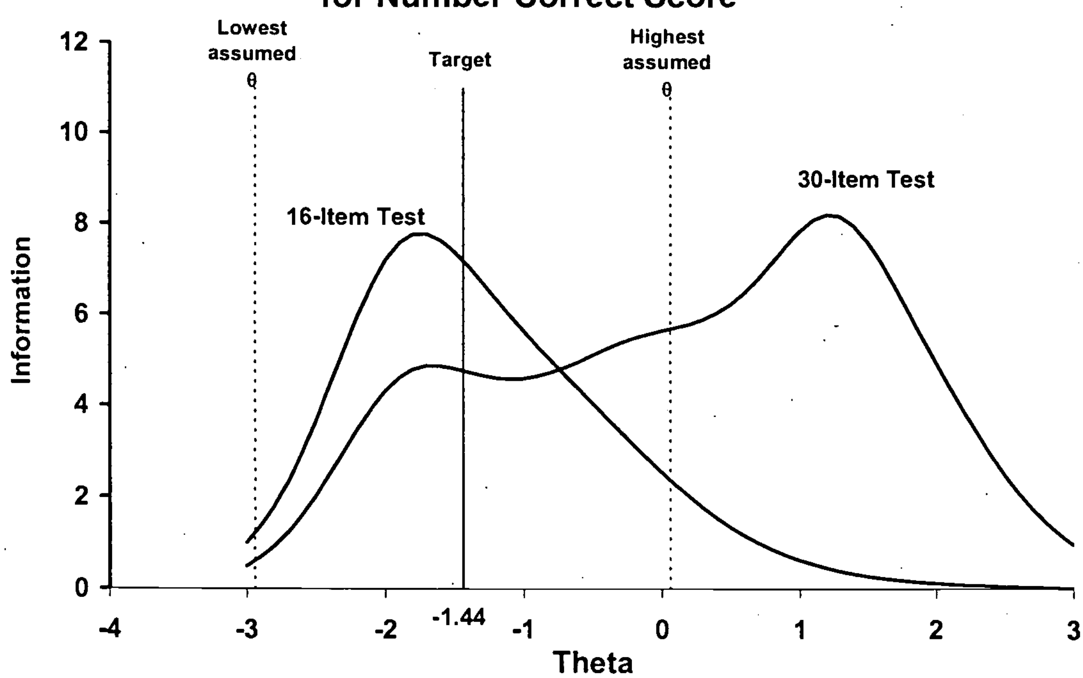


Figure 2: Test Information Functions for Number Correct Score



13

Figure 3: Probability of Passing Level 1

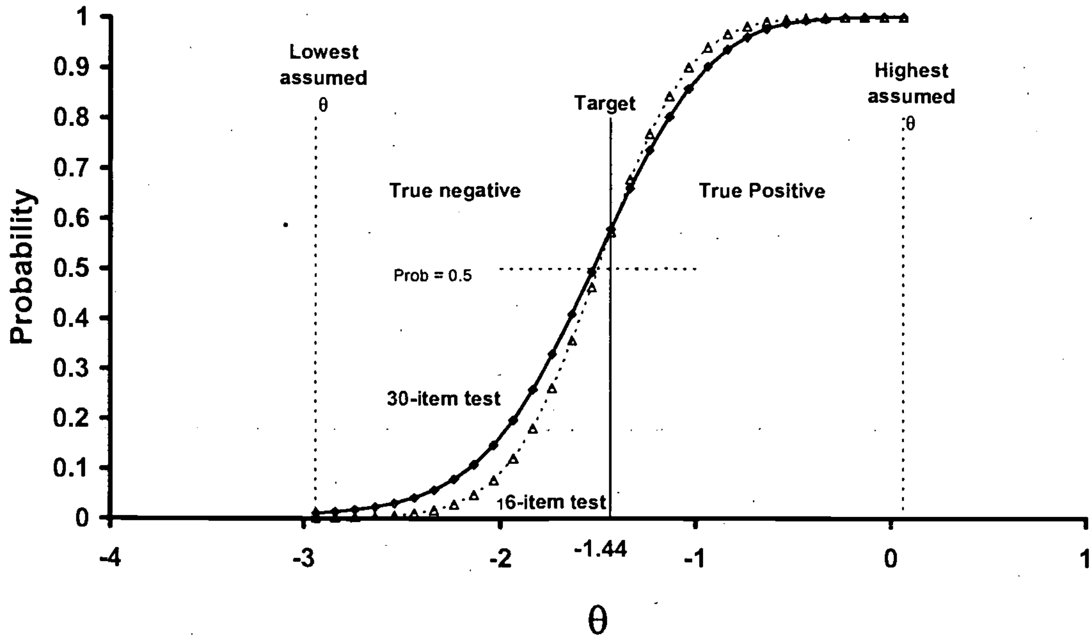


Figure 4: Theta Conditional on Cutscore

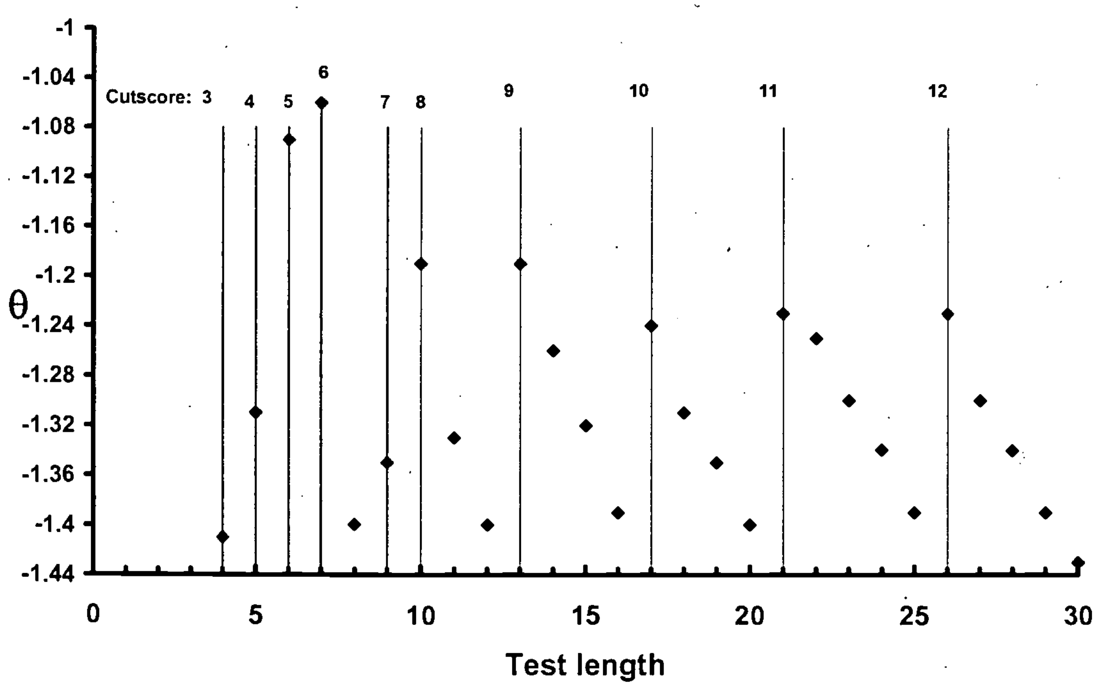
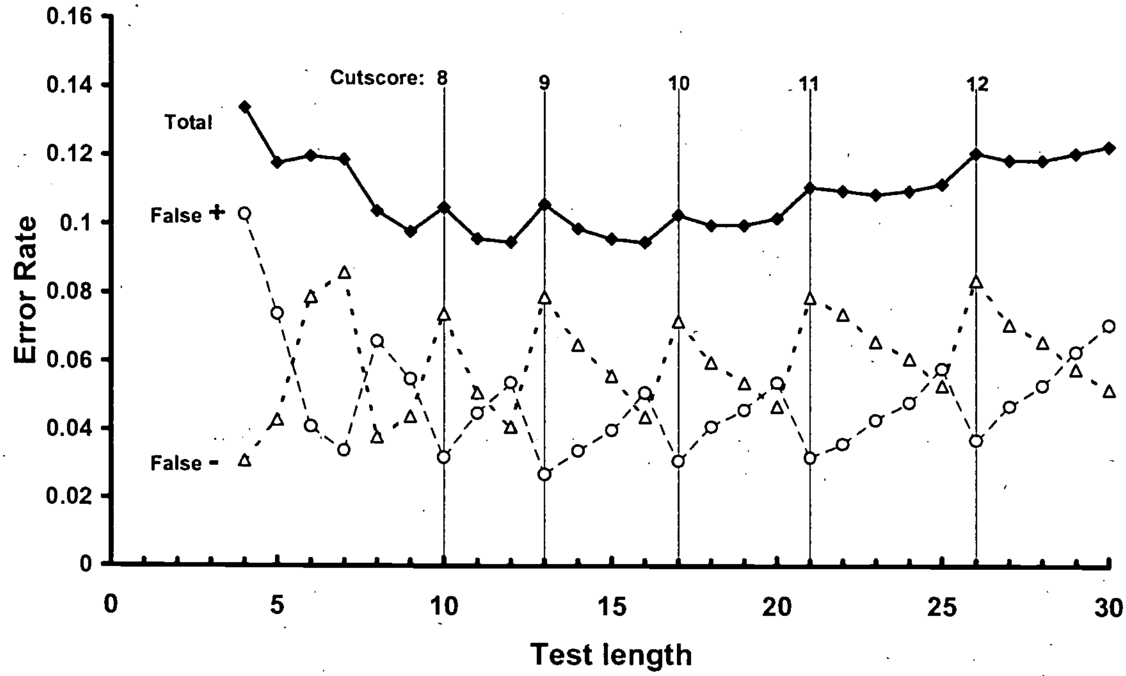


Figure 5: At-or-above Level 1 Classification Error Rates



U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

TM032838

I. DOCUMENT IDENTIFICATION:

Title: The classification accuracy of shortened versus full-length tests.	
Author(s): E. Matthew Schulz and Lin Wang	
Corporate Source: ACT, Inc.	Publication Date: April, 2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☒	Level 2A ☐	Level 2B ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: E. M. Schulz	Printed Name/Position/Title: Senior statistician	
Organization/Address: P.O. Box 168 Iowa City, IA 52243	Telephone: (319) 337-1468	FAX:
	E-Mail Address: schulz@act.org	Date: 5/3/01

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 9/97)