CHANG, HORNG-SHING

AN ANALYTICAL COMPARISON OF CONVENTIONAL AND TAILORED TESTS

*The University of Oklahoma* PH.D. 1981

University
Microfilms
International 300 N. Zeeb Road, Ann Arbor, MI 48106

THE UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

# AN ANALYTICAL COMPARISON OF CONVENTIONAL AND TAILORED TESTS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirement for the

degree of

DOCTOR OF PHILOSOPHY

BY

HORNG-SHING CHANG

Norman, Oklahoma

1981

# AN ANALYTICAL COMPARISON OF CONVENTIONAL AND TAILORED TESTS

APPROVED BY

W. Alan Nicewander

Larry E. Toothaker

Charles Mettler

W. B. Schwarzkopf

Paula Englander-Golden

DISSERTATION COMMITTEE

# ACKNOWLEDGEMENTS

I would like to express my gratitude to Dr. Alan Nicewander and Dr. Larry Toothaker, co-chairman of my committee, for their contributions to my professional development. Their friendly guidance, encouragement, suggestions, and assistance are deeply appreciated. I am also very grateful to other members of my committee, Dr. Charles Gettys, Dr. Paula Englander-Golden, and Dr. Albert Schwarzkopf for their valuable comments and suggestions.

I wish to thank my parents, my brothers, and all my relatives for their support and confidence in me.

Lastly, I want to thank my husband, Nai-hsiang, for his valuable help and enduring patience.

# Abstract

Tailored tests were compared to 'best fixed' tests consisting of optimally chosen sets of items for all examinees. The results indicated that while the fixed tests provided very little information at the extremes of the ability distribution, the fixed tests provided as much or more information than the tailored tests in some regions of the ability distribution. The fixed tests provided average information values close to those for tailored tests. Therefore, if testing is used for selection rather than for classification, fixed tests can serve as competitors to tailored tests.

# TABLE OF CONTENTS

# AN ANALYTICAL COMPARISON OF CONVENTIONAL AND TAILORED TESTS

## CHAPTER I

### Introduction

In the past few years, tailored testing has been introduced to psychological and educational specialists. Tailored testing is an adaptive testing procedure using latent trait theory as the theoretical foundation. It is implemented by varying test items according to the characteristics of the individual being tested; thus, it represents an attempt to select test items according to the ability of the examinee.

The basic notion of adaptive testing is that an examinee is given a more difficult test item if the previous items is answered correctly; an easier item is administered following an incorrect response. This notion arose from clinical applications of individual ability tests (Space (1942); Hutt (1947); Hick (1951); Greenwood & Taylor(1965)). However, Hutt (1947) and Greenwood & Taylor (1965) found that adaptive testing did not yield higher intelligence scores in a group of well-adjusted school children, but poorly-adjusted children received reliably higher intelligence scores with the adaptive method. Their findings suggested that adaptive testing was a better procedure only for certain groups. According to Urry(1977) and Samejima (1977), tailored testing leads to dramatic improvements in ability measurement relative to standard testing methods. However, the improvement is ensured only under certain conditions (Urry,1977). These conditions will be discussed in the later sections.

In tailored testing, each item is selected for administration on the basis of the examinee's responses to previous items, therefore, examinees will be administered different tests. The purpose of this investigation is to compare tailored tests to "best fixed" tests consisting of optimally chosen sets of items for all examinees.

## (1) Latent Trait Theory

In latent trait theory, both univariate and multivariate models have been proposed. Only the univariate latent trait model will be investigated in this research. The univariate latent trait model assumes that examinees can be described by a one-dimensional ability variable, $\theta$, which is scaled to have mean zero and variance one. An item characteristic curve (ICC) is used to represent the probability of a correct answer to an item as a function of the trait , $\theta$, and the item parameters (item difficulty, discrimination and guessing parameters). Let $U_g$ represent the binary item score for the $g^{th}$ item, and denote $P_{g\theta}$ as the conditional probability of answering an item correctly.

Then according to the three-parameter latent ability model,

$$(1) \qquad P(U_g=1|\theta) = P_{g\theta} = c_g + (1-c_g)F\big(L_g(\theta)\big) .$$

where $L_g(\theta)=a_g(\theta-b_g)$; $c_g$ is the guessing parameter; $a_g$ is the item discrimination parameter, and $b_g$ is the item difficulty parameter. $F[L_g(\theta)]$ can represent either the normal CDF or the logistic CDF. For the normal ogive model,

$$F\big(L_g(\theta)\big) = \Phi\big(L_g(\theta)\big) = \int_{-\infty}^{L_g(\theta)} \phi(t)dt = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a_g(\theta-b_g)} \exp(-t^2/2)dt.$$

For the logistic model,

$$F\big(Lg(\theta)\big) = \Psi\big(Lg(\theta)\big) = \frac{1}{1+e^{-D\left(L_g(\theta)\right)}} = \frac{1}{1+e^{-D\left(a_g(\theta-b_g)\right)}} ,$$

where D=1.7. These two functions give nearly identical ICC's; in fact, the values $\Phi[L_g(\theta)]$ and $\Psi[L_g(\theta)]$ differ no more than ±0.01. Choice of the normal ogive or logistic model depends on mathematical circumstances; there is no psychological reason to prefer one to the other. In a situation where obtaining the correct answer through guessing is essentially zero, then the item characteristic function for item g is $P_{g\theta}=F[L_g(\theta)]$.

(2) Information Functions

The information function is a very important concept in latent trait theories. It is an index of the precision of measurement at all levels of the trait being measured. The item information function at ability θ is defined by Birnbaum (1968) as

$$(2) \qquad I(\theta,U_g) = \frac{(P'_{g\theta})^2}{P_{g\theta}\, Q_{g\theta}} \quad ,$$

where $Q_{g\theta}$ is $1 - P_{g\theta}$ , and $P'_{g\theta}$ represents the first derivative of $P_{g\theta}$ with respect to θ . The denominator of the item information function is the conditional variance of the item response variable , $U_g$, at a fixed value of θ. In a three-parameter normal ogive model, evaluating $P_{g\theta}$ yields the information function

$$(3) \qquad I(\theta,U_g) = \frac{\{a_g(1-c_g)\phi(L_g(\theta))\}^2}{P_{g\theta}\, Q_{g\theta}} \quad ,$$

where $\phi[L_g(\theta)]$ is the normal probability density at $L_g(\theta)=a_g(\theta-b_g)$. For a normal ogive model, where guessing is impossible, then the item information function is

$$(4) \qquad I(\theta,U_g) = \frac{\{a_g\phi(L_g(\theta))\}^2}{P_{g\theta}\, Q_{g\theta}} \quad ,$$

The test information function for a test consisting of n items is denoted by I(θ), where

$$I(\theta) = \sum_{g=1}^{n} I(\theta,U_g) \quad .$$

Birnbaum (1968) has shown that the maximum likelihood estimator $\hat{\theta}$ has, asymptotically, a normal distribution with mean zero and variance $1/I(\theta)$, as the number of items approaches infinity.

(3) Tailored Testing

Tailored testing consists of a system of choosing test items and estimating ability for an examinee until the information or number of items administered reach predetermined values. The method obviously requires a computer for implementation. Given an item bank containing items with estimated item parameters $a_g$, $b_g$, and $c_g$, tailored testing proceeds as follows:

(1) The examinee is given an item with a high value of $a_g$ and a value of $b_g$ close to zero (average ability); that is, an item is administered that is most informative for examinees of average ability.

(a) If the examinee's response to the first item is incorrect, then an easier item is given.

(b) If the examinee's response to the first item is correct, then a more difficult item is given.

(2) Repeat (1) until the examinee has at least one correct response and at least one incorrect response.

(3) Based on the examinee's r sponses, ability is estimated using the method of maximum likelihood.

(4) The estimated $\theta$ , $\hat{\theta}$, is used to compute the

information for all remaining items in the item bank at $\hat{\theta}$.

(5) The item with the highest information at $\hat{\theta}$ is selected as the next item for the examinee.

(6) After the examinee responds to each test item, repeat (3), (4), and (5), until the test information function reaches a predetermined criterion value, or a fixed number of items have been administered.

(7) The final ability estimate , $\hat{\theta}$, is the examinee's test score.

According to Urry(1977), the advantage of tailored testing over a conventional test is ensured under the following conditions: (1) the majority of the items have discrimination parameters ($a_g$) equal to or greater than 0.8; (2) the item difficulty parameters ($b_g$) are evenly and widely distributed; (3) the majority of the item guessing parameters ($c_g$) are less than 0.3, and (4) the item bank contains at least 100 items. Notice that in the tailored testing procedure, examinees are administered different tests according to their ability.

The goal of this research was to compare an optimally selected set of test items, administered to all examinees, with tailored tests. Specifically, we were interested in answering the question; Are there situations in which a selection of the "best" items from a item bank will be competitive with the tailored tests in terms of the information provided?

## CHAPTER II

## A Comparison of Fixed and Tailored Tests in an Idealized Experiment

An ideal item bank is defined as an item bank containing an infinite number of items having a single specified value of $a_g$. Any value of $b_g$ is available in an ideal item bank. In an actual testing situation, items are seldom found with $b_g$ values equal to the optimal values required for maximizing information. However, in the idealized situation, items with optimal values of $b_g$ are always available. Both tailored and fixed tests should achieve the maximum potential when constructed from ideal item banks. Comparisons made under the ideal situation should yield an estimate of the relative efficiency of these two testing procedures when they are both at their upper limit of their potential. In the present study, five values of $a_g$ (0.60, 0.80, 1.00, 1.50, 2.00) were selected to define five ideal item banks.

### (1) Perfect Tailored Tests

A perfect tailored test is defined as that collection of items that maximizes the test information function for a known ability. Given a specific ideal item bank, the value of $b_g$ yielding maximum information for a specific ability level can be found. Although the $b_g$ value that maximizes the information value at different ability levels is different, the maximum information values are the same. If the ability of the examinee were known, then the item that maximizes the information at that ability level could be chosen for the examinee. Therefore, a perfect tailored test consisting of n items drawn from an ideal item bank would contain n items of the same $a_g$ and $b_g$ value, and this test would yield the highest possible test information.

In the situation where guessing is impossible, an item with $b_g$ equal to the ability level , $\theta$, will give the highest item information at that ability level; consequently, a perfect tailored test for which guessing is impossible would be composed of n items all having $b_g=\theta$. In the present study, 20-item perfect tailored tests were constructed using the five ideal item banks.

(2) "Best Fixed" Tests

The item selection for the "best fixed" tests in this study was constructed by using average information as the criterion. In order to obtain the average information for an item, it is necessary to specify the distribution of ability. It was assumed that $\theta$ was normally distributed with mean equal to zero and variance equal to one. If the full range of ability was of interest, then items for the fixed tests were chosen to maximize

$$(5) \qquad \mu\left[I(\theta,U_g)\right] = \int_{-\infty}^{\infty} I(\theta,U_g)\phi(\theta)d\theta \qquad ,$$

the (unconditional) average information function. The integration required in (5) was done numerically using Gauss-Hermite quadrature. It was found that an item with $b_g$ =0 would yield the highest average information in the situation where guessing is impossible. Therefore, the "best fixed" test drawn from an ideal item bank would be composed of 20 items with $b_g$=0. The average information for a test is equal to the sum of the average item information functions, that is,

$$\int_{-\infty}^{\infty} \sum_{g=1}^{n} I(\theta,U_g)\phi(\theta)d\theta = \sum_{g=1}^{n} \int_{-\infty}^{\infty} I(\theta,U_g)\phi(\theta)d\theta \qquad .$$

Therefore, the average test information for a n-item test is obtained by addition of the average item information functions for the n items.

If a sub-population was of interest--for example, examinees with above average ability--then for the "best fixed" tests, we chose the items that yielded the highest conditional average information in that limited region of the ability distribution. The conditional average information for $\theta \geq 0$ is given by

$$(6)\quad \mu\left(I(\theta,U_g|\theta\geq 0)\right) = 2\int_0^{\infty} I(\theta,U_g)\phi(\theta)d\theta \quad .$$

Similarly, if only the sub-population of average and below average ability was of interest, then the items for the fixed tests that maximized

$$(7)\quad \mu\left(I(\theta,U_g|\theta\leq 0)\right) = 2\int_{-\infty}^{0} I(\theta,U_g)\phi(\theta)d\theta$$

were chosen. A modification of Gauss-Hermite graduature was used to accomplish the integration in (6) and (7).

A simple computer search procedure was used to find the $b_g$ value that maximized (6) and (7) for the five ideal item banks. For the full ability distribution, it was established that $b_g=0$ would maximize (5) for $a_g=0.6$, 0.8, 1.0, 1.5, 2.0. However, the $b_g$ values that maximized (6) and (7) were different for different $a_g$ values.

-------------------------

Insert Table 1 about here

-------------------------

In Table 1 notice that as $a_g$ increased, the smaller the optimal $b_g$ value was for the above average ability sub-population. The more discriminating the item, the easier the item must be to yield maximum conditional average information for examinees of average and above-average ability. Table 1 also indicates that the opposite was true for the sub-population of average and below average ability. For a specific $a_g$ value, the $b_g$ values that maximized the conditional average information for the upper and lower tails of the ability distribution were of the same absolute value but the signs were opposite. However, the conditional average information values were the same for both cases. Therefore, only the upper tail case was discussed in this section. In order to simplify the discussion of outcomes, a Type I fixed test is defined as that collection of items that maximizes average information for the complete ability distribution; a Type II fixed test is defined as that collection of items that maximizes the conditional average information for average and above-average ability, and a Type III fixed test is defined as that collection of items that maximizes the conditional average information for average and below-average ability.

(3) Simulated Tailored Tests

In an actual tailored testing situation, the ability of the examinee is not known and must be successively approximated throughout the testing. Define this situation as "fallible" tailored testing. In a "fallible" tailored testing situation, one starts with items that rarely yield the highest information for an individual's ability level. Therefore, one will always obtain less information than a perfect tailored test, or more items are necessary in order to yield information equal to that of a perfect tailored test.

The Monte Carlo method was used to simulate the fallible tailored test situation. The simulation was done for different ability levels (ranging from 0 to 3 by steps of 0.5). For an ideal item bank defined by a specific $a_g$ value, the first item that was given to the "examinee" was an item with $b_g=0$. The response of an "examinee" was generated by using Chen's (1971) uniform random number generator. If the number generated was less-than or equal-to $\Phi[L_g(\theta)]$, then the response of that individual was recorded as correct ($U_g=1$); otherwise, the response was recorded as incorrect ($U_g=0$). Based on the "examinee's" response to the first item, the second item was chosen. If the first response was correct, the next item was increased in difficuty by a fixed step size of k=0.693. If the response was incorrect, the next item was decreased in difficulty by the fixed step size of k=0.693. After the response to the second item was generated, if the response was different from the first, then $\theta$ was estimated through the method of maximum likelihood. If the second response was the same as the first one (either both correct or both incorrect), $\theta$ was not estimated, and the difficulty of the next item was increased or decreased by the step size k according to the examinee's previous response. For example, if the first two responses were both correct, the next item was chosen to have $b_g$=2k=1.386; if the first two responses were incorrect, the next item was chosen such that $b_g$ =-2k=-1.386. This procedure was repeated until the examinee had at least one correct and one incorrect response. After the examinee attained this state of affairs, $\theta$ was estimated by searching the log likelihood function to find the maximum likelihood estimate, $\hat{\theta}$. Once $\theta$ was estimated, the next item for the examinee was chosen with $b_g=\hat{\theta}$ (estimated $\theta$). The above procedure continued until 20 items were given to the examinee. The final estimated $\theta$ and the test information were recorded. This procedure was repeated 100 times for each specified ability level, and the average of the estimated $\theta$'s and the average test information were computed. This simulation was done for the five ideal item banks.

(4) Results:

Table 2 summarizes the conditional averages of the test information functions for fixed tests, perfect tailored tests and simulated tailored tests.

-------------------------

Insert Table 2 about here

-------------------------

Table 2 indicates that the Type I fixed tests were about 75-88 percent as efficient as the perfect tailored tests and about 82-95 percent as efficient as the simulated tailored tests for the ideal item banks defined by $a_g \leq 1.0$. For the same item banks, the Type II fixed tests were about 89-95 percent as efficient as the perfect tailored tests, and about 96-102 percent as efficient as the simulated tailored tests. Thus, on the average, fixed tests can be as efficient as tailored tests constructed from an ideal item banks defined by $a_g \leq 1.0$. For the ideal item bank defined by $a_g=2.0$, the Type I fixed test was 49 percent as efficient as the perfect tailored test and 57 percent as efficient as the simulated tailored test, while the Type II fixed test was 70 percent as efficient as the perfect tailored test and 80 percent as efficient as the simulated tailored test. However, averages can be misleading, because the fixed tests provided very little information at both extremes of the ability distribution.

The information curves for the perfect tailored tests, simulated tailored tests, and the "Type I best fixed" tests, constructed from the five ideal item banks, were plotted in Figures 1 through 5.

Insert Figure 1 to Figure 5 about here

These figures display the superiority of the perfect tailored tests across the entire ability distribution. However, the Type I fixed tests were at least 90 percent as efficient as the perfect tailored tests for abilities between -0.5 and 0.5 when the ideal item banks were defined by $a_g \leq 1.0$. For the same ideal item banks, the Type I fixed tests provided higher levels of information than simulated tailored tests for θ between -0.5 and 0.5. The region in which fixed tests had the advantage over tailored tests, however, decreased as the $a_g$ value increased. The fixed tests provided very little information for abilities greater than 2.0 or less than -2.0. As $a_g$ increased, the fixed tests provided almost no information at the extremes of the ability distribution.

Figures 6 through 10 display the information curves for the perfect tailored tests, simulated tailored tests and Type II fixed tests drawn from the five ideal item banks.

---

Insert Figure 6 to Figure 10 about here

---

The perfect tailored tests provided higher information across the entire ability distribution compared to the Type II fixed tests. However, the Type II fixed tests provided more information than simulated tailored tests between $\theta=0$ and $\theta=1.5$, and the average information for Type II fixed tests was at least 90 percent of that of the perfect tailored tests when $a_g \leq 1.0$. The region of the ability distribution in which fixed tests provided more information than simulated tailored tests decreased as $a_g$ increased. Notice that the region in which fixed tests were superior to tailored tests was broader for the Type II fixed tests than for the Type I fixed tests.

As mentioned earlier, the perfect tailored tests define the upper limit of test information for a test, because $\theta$ is known. The simulated fallible tailored tests, constructed when $\theta$ is unknown, must yield less information than perfect tailored tests. However, the simulated tailored tests provide a better picture of a real tailored testing. Table 3 lists the percentages of information loss through estimation of ability in "fallible" tailored tests.

Insert Table 3 about here

Table 3 indicates that the larger the $a_g$ value, the more test information was lost due to the misestimation of ability in the initial phase of the tailored testing. Therefore, the misestimation of ability caused more loss in test information (relative to perfect tailored tests) for more discriminating items. The information drop was greater at the extremes than in the middle of the ability distribution. However, this information-loss problem does not exist for the fixed tests, because all examinees respond to the same set of items, and ability estimation is not performed until all items have been administered.

# CHAPTER III

## A Comparison of Fixed and Tailored Tests for Certain Real Finite Item Banks

The above results were based on data generated from an idealized situation and indicated the maximum potential for both tailored and fixed tests. In this section, the comparison of testing methods was made using some actual item banks. The item banks included in this section are from Urry (1974), and Koch & Reckase (1978). Urry's item bank consisted of 200 items selected from a collection of 700 calibrated items. The selection was based on the requirements for effective tailored testing mentioned earlier. Table 4 shows the distributions of $a_g$ and $b_g$ values for both Urry's and Koch-Reckase's item banks.

-------------------------

Insert Table 4 about here

-------------------------

In Urry's item bank, the item discrimination parameter, $a_g$, ranged from 0.8 to 2.41, and the $b_g$ values ranged from -1.58 to 2.36. Urry's item bank contained more difficult items than easy items. In the Koch-Reckase item bank, the $a_g$ values ranged from 0.13 to 2.00, while the difficulty parameter , $b_g$, ranged from -3.6 to 6.0. Overall, the Koch-Reckase item bank contained more easy items than difficult items.

(1) Perfect Tailored Tests Constructed from the Urry and Koch-Reckase Item Banks

For these two actual item banks, the perfect tailored tests were constructed by choosing those items that yielded the highest information at specific ability levels. Unlike the idealized situation, in which the information functions for the perfect tailored tests were constant at all ability levels, the information functions for perfect tailored tests constructed from real item banks were different at different ability levels. This is the case because the items are finite in number, and the frequency distributions of the $a_g$ and $b_g$ values are not uniform. The item information for each item in a particular bank was computed at each of 49 values of ability, ranging from -3.0 to +3.0 in steps of 0.125. At each ability level, the items in each bank were rank-ordered according to information. At each ability level, the first $m_i$-items constituted the perfect tailored test for the ith ability, where $m_i$=Min(25, $L_i$), and $L_i$ is the number of items yielding test information of 20 or more. Therefore, perfect tailored tests constructed from the finite item banks may be of different lengths depending on the ability level. For Urry's item bank, the number of items ($m_i$) necessary for perfect tailored tests, ranged from 8 to 25, and from 11 to 25 for the Koch-Reckase item bank.

Since the perfect tailored tests were of different lengths for different ability levels, it was decided to use the average number of items in the tailored tests as the length of the "best fixed" tests. The average number of items in the tailored tests was computed using the following equation:

$$(8) \qquad \bar{m} = \sum_{i=1}^{49} m_i P(\theta_i) \quad ,$$

where $m_i$ is the number of items in the perfect tailored test for the i th ability level. $P(\theta_i)$ is defined by

$$(9) \qquad P(\theta_i) = \int_{\theta_i - 0.0625}^{\theta_i + 0.0625} \phi(t)dt \quad ,$$

where $\phi(t)$ is the normal density function, and $\theta_i$ ranged from -3.0 to +3.0 by steps of 0.125. The average value , $\bar{m}$, was rounded to the nearest integer value , M, and this value served as the number of items for the fixed test.

(2) "Best Fixed" Tests Constructed from the Urry and Koch-Reckase Item Banks

The item selection for the "best fixed" tests was based on average information values. Since the perfect tailored tests for the real item banks were constructed for each of a finite number of ability levels, a discrete approximation to the average information function was used in order to obtain the average information for all the items in these two item banks. If the target population for testing was described by the full ability distribution, a discrete approximation to the average information of any item was obtained by

$$(10) \qquad \mu\left(I(\theta,U_g)\right) = \sum_{i=1}^{49} I(\theta,U_g)P(\theta_i) \qquad ,$$

where $\theta_i$ and $P(\theta_i)$ were defined as in (9). The average information values from 200 items from Urry's item bank were computed and the M items having the highest values of $\mu[I(\theta,U_g)]$ comprised the "best Type I fixed" test. The same item selection procedure was applied to the Koch-Reckase item bank in order to choose the items for the "best" Type I fixed test. If only a portion of the ability distribution was the target population for testing, then one of the following equations was used to obtain the conditional average information values:

$$(11) \qquad \mu\left(I(\theta,U_g|\theta\geq 0)\right) = \frac{1}{\sum_{i=25}^{49} P(\theta_i)} \sum_{i=25}^{49} I(\theta_i,U_g)P(\theta_i)$$

for the upper portion of ability distribution, or

$$(12) \qquad \mu\left(I(\theta,U_g|\theta\leq 0)\right) = \frac{1}{\sum_{i=1}^{25} P(\theta_i)} \sum_{i=1}^{25} I(\theta_i,U_g)P(\theta_i)$$

for the lower portion of ability distribution. Notice that the "best fixed" test (Type I, Type II or Type III fixed tests) contained different sets of items. The perfect tailored tests, however, contained the same items regardless of the portion of the ability distribution being averaged across. Although the perfect tailored tests contained the same items for the full ability distribution or subregions of the ability distribution, the conditional average information values for these regions were different.

(3) Results:

The average test information values, which were obtained by addition of the average item information values for the fixed tests and by averaging the test information across the appropriate portion of the ability distribution for the perfect tailored tests, are summarized in Table 5.

-------------------------

Insert Table 5 about here

-------------------------

This table shows that for Urry's item bank, the Type I fixed test was about 70 percent as efficient (on the average) as the perfect tailored tests. However, the Type II and Type III fixed tests provided about 85 percent as much information as the perfect tailored tests. Urry's item bank has a slight bias in that it contained more difficult items than easy items; therefore, the average information for the

Type II fixed test is slightly higher than that for the Type I and Type III fixed tests. Also the average number of items for the Type II fixed test was smaller than that for the Type I or Type III fixed tests. Table 5 also indicates that the fixed tests constructed from the Koch-Reckase item bank were at least 98 percent as efficient, on the average, as the perfect tailored tests. In the latter item bank, the Type II and III fixed tests provided essentially the same average information as the perfect tailored tests. The Type III fixed tests provided slightly higher average information than the perfect tailored tests. This phenomenon was due to the fact that the number of items in the Type III fixed test was greater than the number of items in the perfect tailored tests for θ between -0.75 and 0.25. The Koch-Reckase item bank has a slight bias in the direction of easy items; therefore, the average test information for Type III fixed test was higher than that of the Type I or Type II fixed tests. The number of items for the Type III fixed test was smaller than that for the Type I or Type II fixed tests, because the average lengths of the perfect tailored tests for average and below-average ability were shorter than the average lengths of the perfect tailored tests for average and above-average ability and for the entire ability distribution.

The comparison of the information functions for the perfect tailored tests and the Type I fixed tests are displayed in Figure 11 and 12 for the Urry and Koch-Reckase item banks respectively.

---

Insert Figure 11 to Figure 16 about here

---

For both item banks, in a small region of the ability distribution, the Type I fixed tests provided more information than the perfect tailored tests. This outcome was due to the fact that the number of items in the fixed tests was larger than the number of items in the perfect tailored tests at these θ values.

Figures 13 and 14 show the test information for the perfect tailored tests and the Type II fixed tests constructed from the Urry and Koch-Reckase item banks respectively. The perfect tailored tests provided higher information with the exception of those θ values where the fixed tests contained more items than the perfect tailored tests. However, Figure 14 indicates that the information values for the Type II fixed tests constructed from the Koch-Reckase item bank were very close to those of the perfect tailored tests, particularly for abilities greater than zero.

Figure 15 shows that the perfect tailored tests were outperforming the Type III fixed test constructed from Urry's item bank across the enitre ability distribution. Figure 16 indicates that Type III fixed test constructed from the Koch-Reckase item bank can serve as a competitor to tailored tests.

Overall, the Type II and III fixed tests constructed from the Koch-Reckase item bank provided nearly as much information, on the average, as tailored tests. However, the tailored tests were superior to all three types of fixed tests constructed from Urry's item bank.

# CHAPTER IV

## Summary and Conclusion

Fixed tests and perfect tailored tests "constructed" from ideal and actual (finite length) item banks were compared in this study. The item selection criterion for the fixed tests was maximum unconditional average information or maximum conditional average information, depending on the region of the ability distribution that was of interest for testing.

The results generated using the real item banks were consistent with those generated from the ideal item banks. On the average, fixed tests provided average information very close to that of tailored tests; especially if only a limited region of the ability distribution was of interest for testing. However, the averages must be interpreted cautiously. The comparison of the actual information fuctions indicated that the perfect tailored tests were superior to the fixed tests across the entire ability distribution-- particularly at the two extremes of the distribution. The fixed tests provided very little information for abilities greater than two standard deviations above or below the mean.

The efficiency of the fixed tests, relative to tailored tests, depends on the characteristics of the items. With less discriminating items; that is, when the $a_g$ values were smaller than 0.8, the fixed tests were almost as efficient as the tailored tests. For more highly discriminating items, the tailored tests were superior to the fixed tests across a wide range of abilities. However, with more highly discriminating items, the information loss due to the misestimation of $\theta$ in the simulated, "fallible" tailored tests was more pronounced than with less discriminating items. The results from the ideal item banks showed that fixed tests can provide more information than the fallible tailored tests in some regions of the ability distribution. Although the fixed tests provided very little information at the extremes of the ability distribution, the fixed tests might serve as competitors to tailored tests especially if the test is used for selection rather than for classification.

# REFERENCES

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. and Novick, M. Statistical Theories of Mental Test Scores. Reading, Mass.:Addison-Wesley, 1968.

Chen, E.H. Random normal number generator for 32-bit-word computers. Journal of the American Statistical Association, 1971, 66, No. 334, 400-403.

Greenwood, D.I. & Taylor, C Adaptive testing in an older population. Journal of Psychology, 1965, 60, 193-198.

Hick, W.E. Information theory and intelligence tests. British Journal of Psychology, Statistical Section, 1951, 4, 157-164.

Hutt, M.L. A clinical study of "constructive" and "adaptive" testing with the revised Stanford-Binet. Journal of Consulting Psychology, 1947, 11, 93-103.

Koch, W.R. and Reckase, M.D. A live tailored testing study of the one and three parameter logistic models. Res. Report 78-1, 1978, Tailored Testing Research Laboratory, Dept. of Educational Psychology, University of Missouri - Columbia.

Samejima, F. Use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1, No. 2, 233-248.

Spache, G. Serial testing with the revised Stanford-Binet Scale, Form L, in the test range II - XIV. American Journal of Orthopsychiatry, 1942, 12, 81-86.

Urry, V.W. Computer-assisted testing: calibration and evaluation of the verbal ability bank. Technical Study 74-3, 1974, Personnel Research and Development Center, U.S. Civil Service Commission, Washington, D.C.

Urry, V.W. Tailored testing: a successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, No. 2, 181-195.

Table 1

Optimum Item Difficulty ($b_g$) Values that Maximized Average Information for the Upper and Lower-Portions of the Ability Distribution for the Five Ideal Item Banks

| $a_g$ value | Optimum $b_g$ for above average ability | Optimum $b_g$ for below average ability |
|---|---|---|
| 0.60 | 0.80 | -0.80 |
| 0.80 | 0.78 | -0.78 |
| 1.00 | 0.75 | -0.75 |
| 1.50 | 0.69 | -0.69 |
| 2.00 | 0.62 | -0.62 |

Table 2

Average Information for 'Best Fixed' Tests, Simulated Tailored Tests and Perfect Tailored Tests for the Five Ideal Item Banks

| | $a_g$ value | 'Best fixed' tests | Perfect tailored tests | Simulated tailored tests | Information ratio [1] | Information ratio [2] |
|---|---|---|---|---|---|---|
| Comparisons involving Type I fixed tests | 0,60 | 4.07 | 4.58 | 4.28 | 0.88 | 0.95 |
| | 0.80 | 6.70 | 8.15 | 7.52 | 0.82 | 0.89 |
| | 1.00 | 9.61 | 12.73 | 11.66 | 0.75 | 0.82 |
| | 1.50 | 17.37 | 28.65 | 25.51 | 0.61 | 0.68 |
| | 2.00 | 25.23 | 50.93 | 44.13 | 0.49 | 0.57 |
| Comparisons involving Type II fixed tests | 0.60 | 4.40 | 4.58 | 4.30 | 0.95 | 1.02 |
| | 0.80 | 7.61 | 8.15 | 7.54 | 0.92 | 1.00 |
| | 1.00 | 11.50 | 12.73 | 11.71 | 0.89 | 0.96 |
| | 1.50 | 23.51 | 28.65 | 25.62 | 0.79 | 0.88 |
| | 2.00 | 37.80 | 50.93 | 44.46 | 0.70 | 0.80 |

[1] Ratio of average information in 'best fixed' tests to average information in perfect tailored tests.

[2] Ratio of average information in 'best fixed' tests to average information in simulated tailored tests.

Table 3

The Percentage of Decrease in Information for Simulated Tailored Tests Relative to Perfect Tailored Tests in the Idealized Situation

| $a_g$ | $\theta =$ | 0.00 | 0.50 | 1.00 | 1.50 | 2.00 | 2.50 | 3.00 |
|---|---|---|---|---|---|---|---|---|
| 0.60 | | 5.17 | 5.30 | 7.11 | 9.18 | 9.76 | 11.30 | 13.37 |
| 0.80 | | 6.40 | 5.85 | 8.14 | 10.29 | 11.48 | 14.49 | 17.58 |
| 1.00 | | 7.15 | 7.44 | 8.60 | 10.38 | 13.38 | 16.83 | 20.04 |
| 1.50 | | 8.70 | 8.92 | 11.37 | 14.44 | 17.75 | 20.71 | 23.48 |
| 2.00 | | 9.46 | 12.50 | 13.57 | 16.72 | 18.71 | 24.07 | 27.70 |

Table 4

The Frequency Distributions of
$a_g$ and $b_g$ Values for Urry's and Koch-Reckase's Item Banks

| | Urry's item bank | Koch-Reckase's item bank |
|---|---|---|
| (1) $a_g$ values | | |
| 0.0 $a_g$ 0.5 | 0 | 16 |
| 0.5 $a_g$ 1.0 | 60 | 24 |
| 1.0 $a_g$ 1.5 | 105 | 14 |
| 1.5 $a_g$ 2.0 | 27 | 16 |
| 2.0 $a_g$ 2.5 | 8 | 2 |
| total number of items = | 200 | 72 |
| (2) $b_g$ values | | |
| $b_g$ -3.0 | 0 | 3 |
| -3.0 $b_g$ -2.0 | 0 | 5 |
| -2.0 $b_g$ -1.0 | 27 | 16 |
| -1.0 $b_g$ 0.0 | 60 | 31 |
| 0.0 $b_g$ 1.0 | 78 | 10 |
| 1.0 $b_g$ 2.0 | 30 | 4 |
| 2.0 $b_g$ 3.0 | 5 | 1 |
| 3.0 $b_g$ | 0 | 2 |
| total number of items = | 200 | 72 |

Table 5

The Average Information for 'Best Fixed' Tests and Perfect Tailored Tests Constructed from Urry's and Koch-Reckase's Item Banks

| | 'Best fixed' tests (No. of items) | Perfect tailored tests | Information ratio [1] |
|---|---|---|---|
| (1) Urry's item bank | | | |
| Type I fixed tests | 14.7502 (N=13) | 20.4426 | 0.72 |
| Type II fixed tests | 18.0364 (N=10) | 20.7731 | 0.87 |
| Type III fixed tests | 17.2120 (N=15) | 20.1788 | 0.85 |
| (2) Koch-Reckase's item bank | | | |
| Type I fixed tests | 17.3435 (N=19) | 17.7544 | 0.98 |
| Type II fixed tests | 16.3249 (N=21) | 16.4765 | 0.99 |
| Type III fixed tests | 19.7475 (N=15) | 19.3369 | 1.02 |

[1] Ratio of average information for 'best fixed' tests to average information for perfect tailored tests.

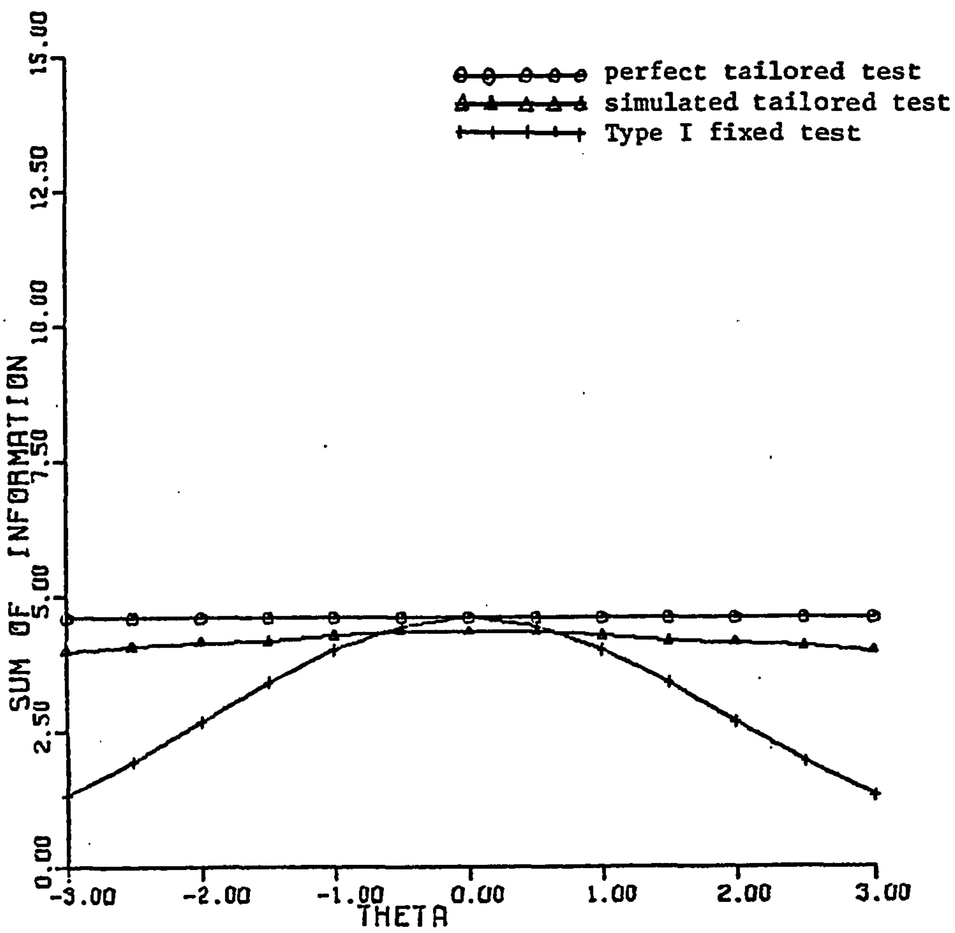

Figure 1. Information functions for perfect tailored tests, simulated fallible tailored tests, and Type I fixed tests for ideal item bank defined by $a_g$=.60 .

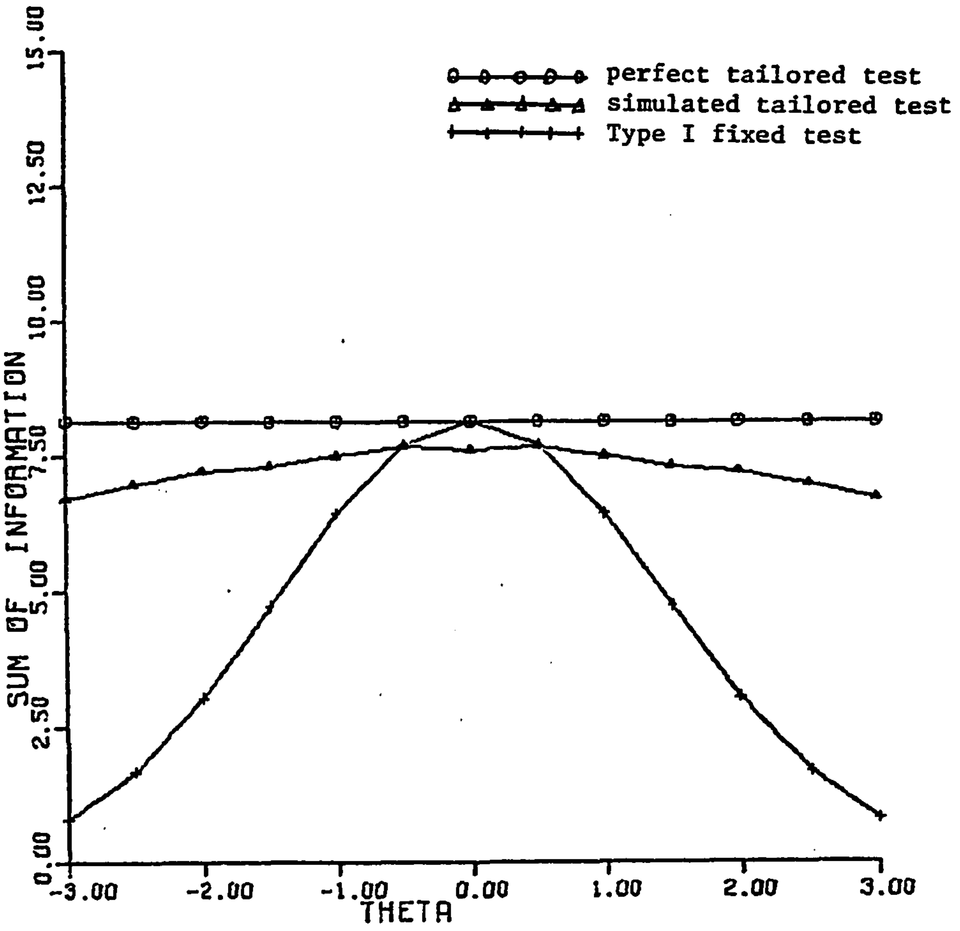

Figure 2. Information functions for perfect tailored tests, simulated fallible tailored tests, and Type I fixed tests for the ideal item bank defined by $a_g$=.80.

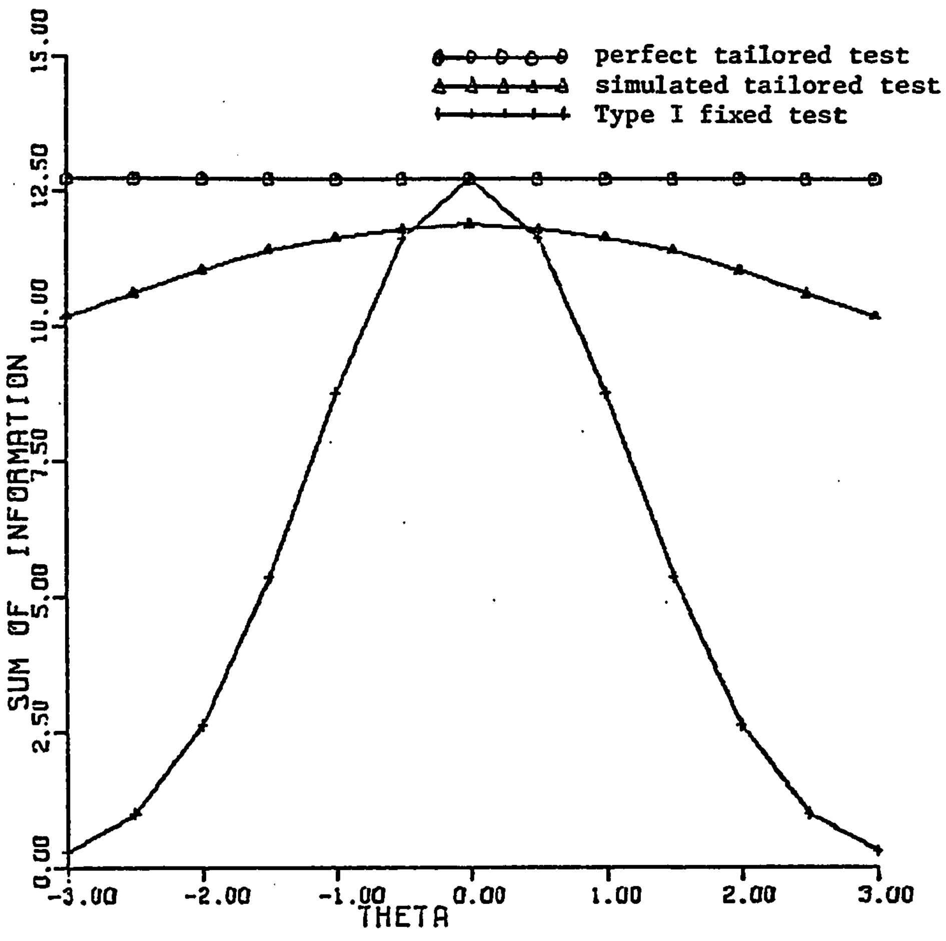

Figure 3. Information functions for perfect tailored tests, simulated fallible tailored tests, and Type I fixed tests for the ideal item bank defined by $a_g$=1.00 .

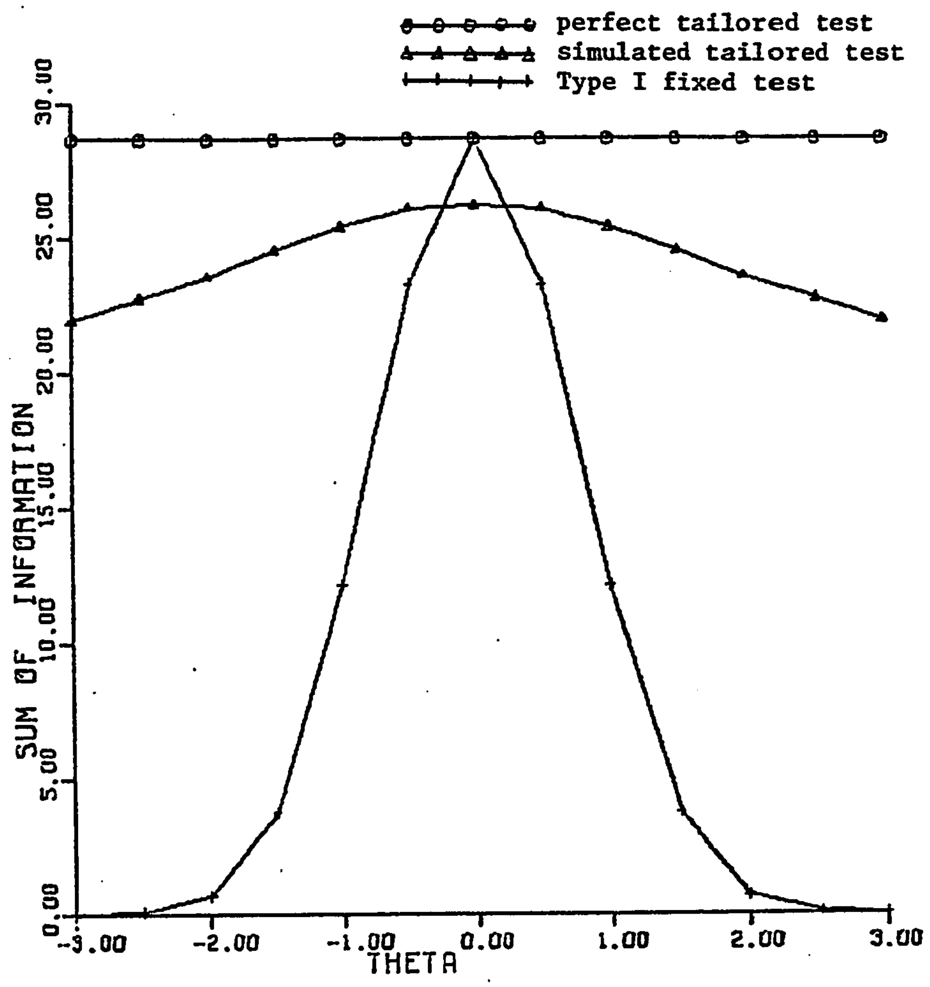

Figure 4. Information functions for perfect tailored tests, simulated fallible tailored tests, and Type I fixed tests for the ideal item bank difined by $a_g$=1.50 .

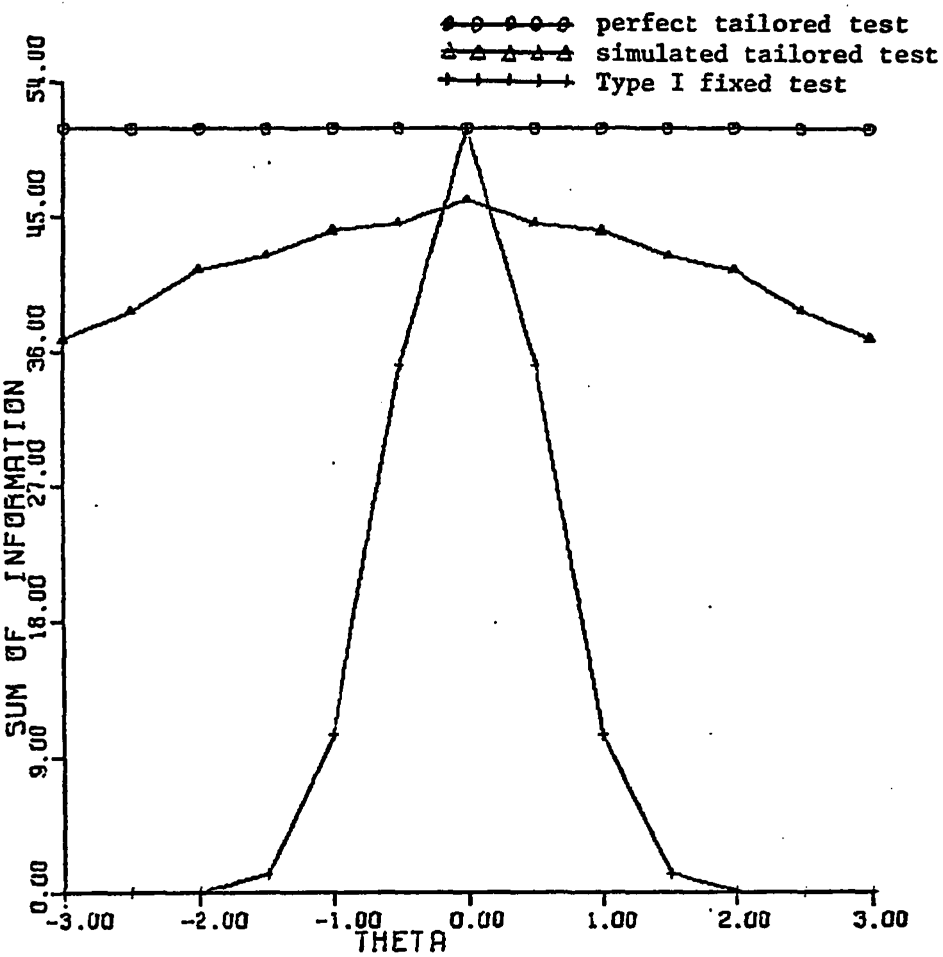

Figure 5. Information functions for perfect tailored tests, simulated fallible tailored tests, and Type I fixed tests for the ideal item bank defined by $a_g$=2.00 .

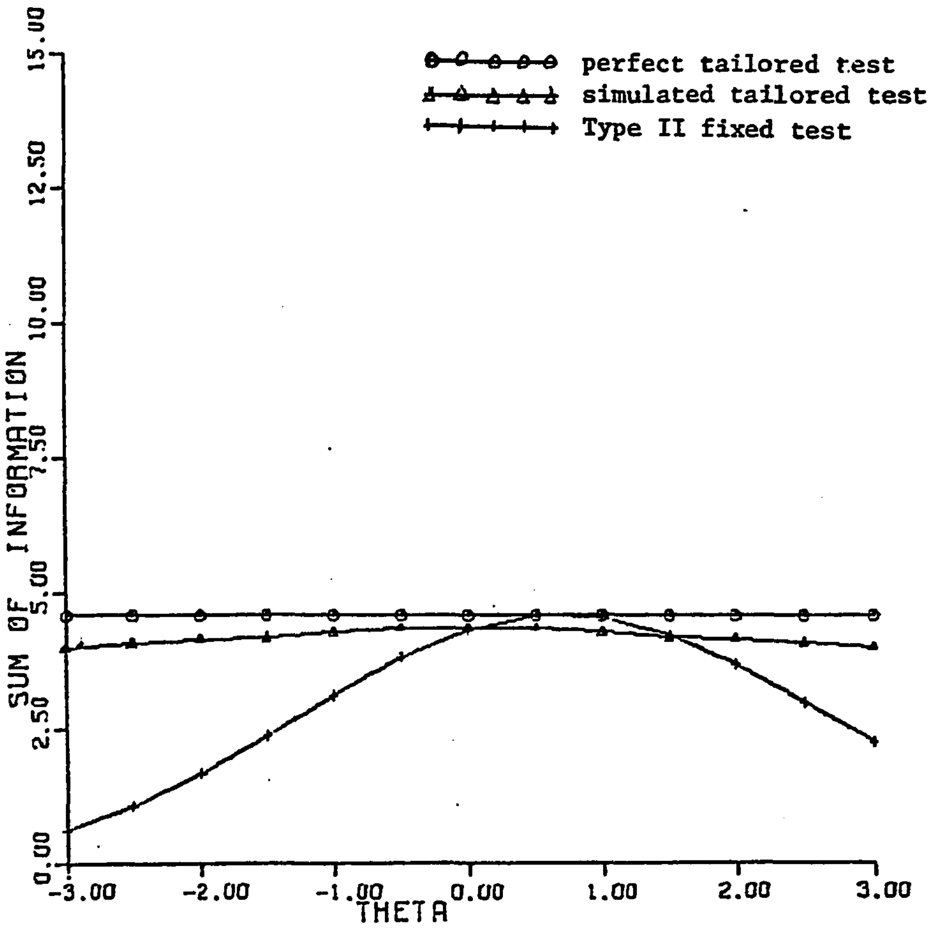

Figure 6. Information functions for perfect tailored tests, simulated fallible tailored tests, and Type II fixed tests for the ideal item bank defined by $a_g$=.60

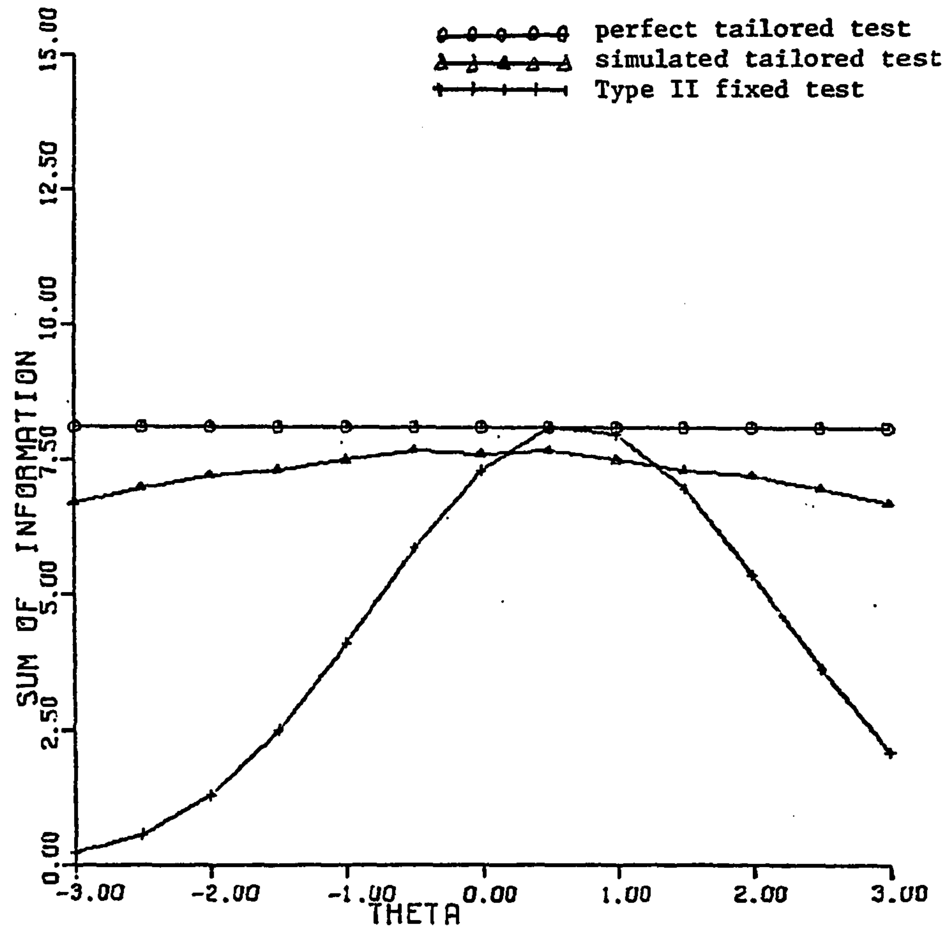

Figure 7. Information functions for perfect tailored tests, simulated fallible tailored tests, and Type II fixed tests for the ideal item bank defined by $a_g$=.80 .

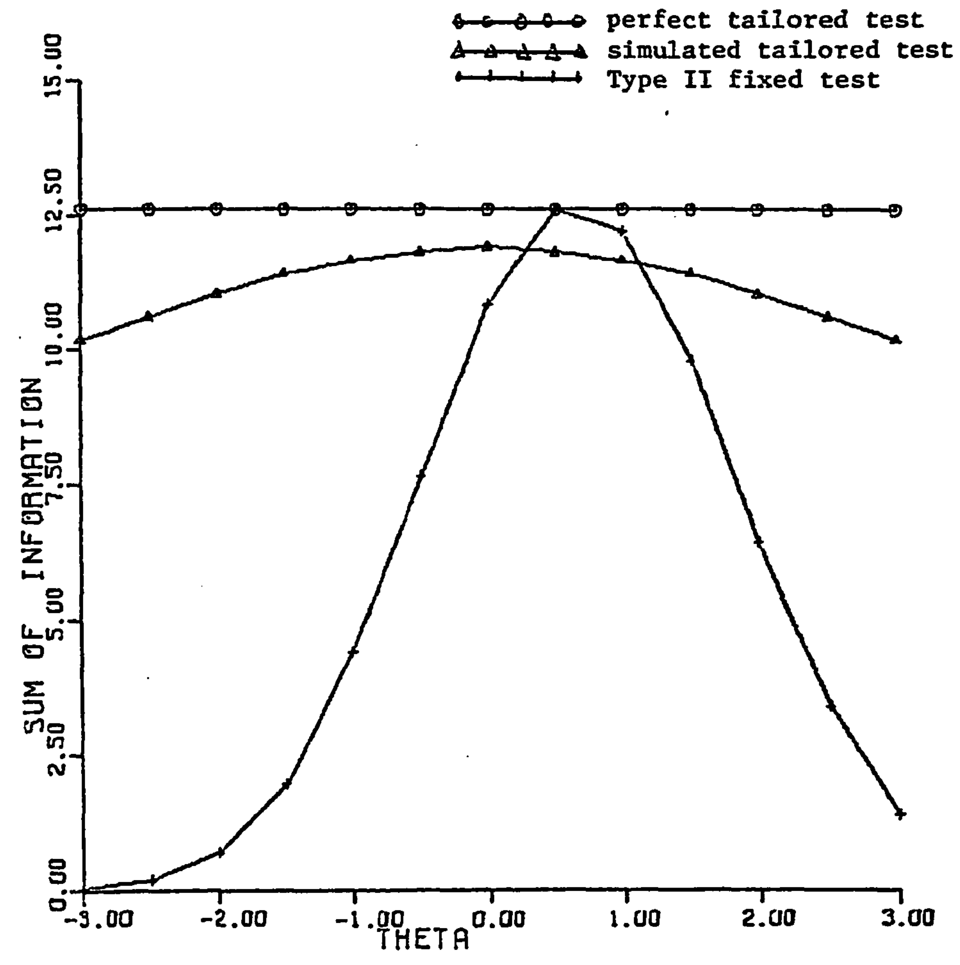

Figure 8. Information functions for perfect tailored tests, simulated fallible tailored tests, and Type II fixed tests for the ideal item bank defined by $a_g$=1.00 .

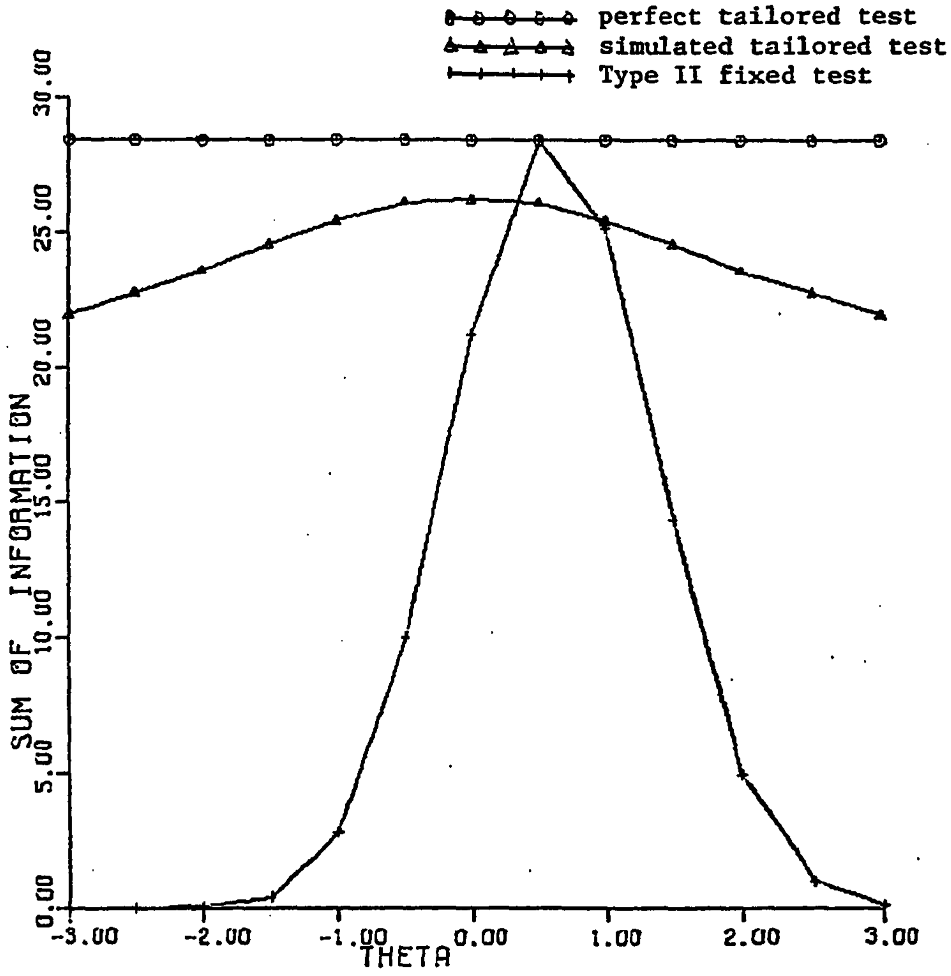

Figure 9. Information functions for perfect tailored tests, simulated fallible tailored tests, and Type II fixed tests for the ideal item bank defined by $a_g$=1.50 .

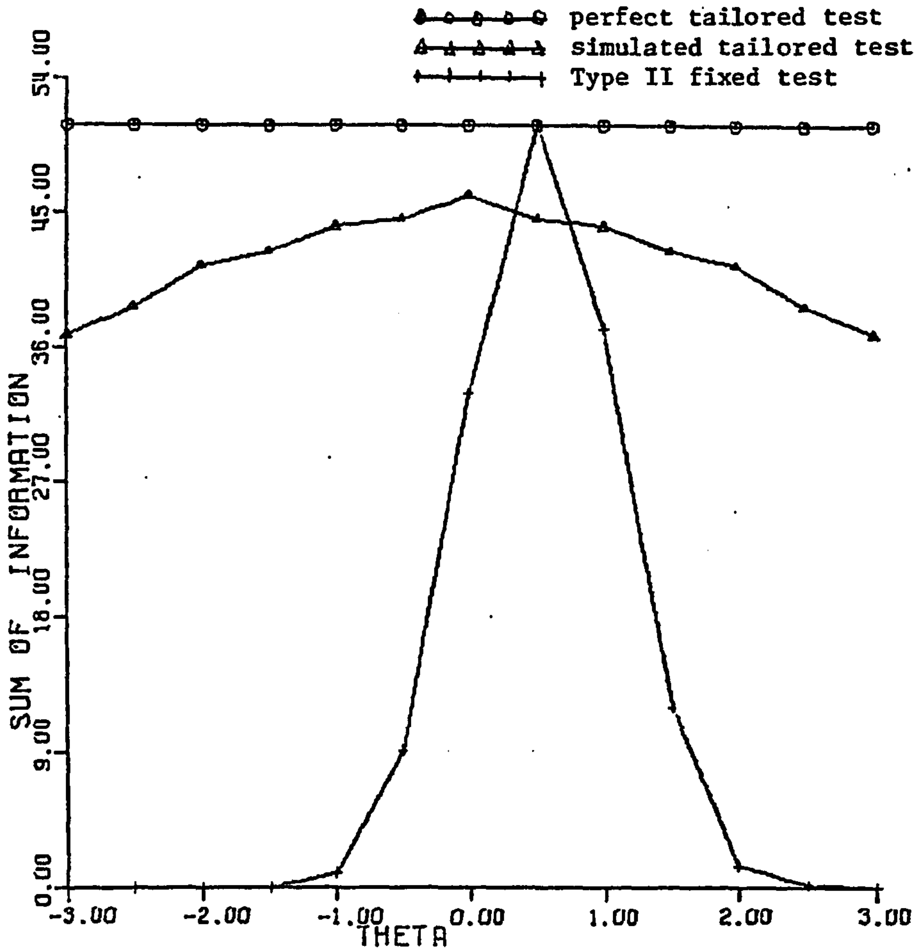

Figure 10. Information functions for perfect tailored tests, simulated fallible tailored tests, and Type II fixed tests for the ideal item bank defined by $a_g$=1.50 .

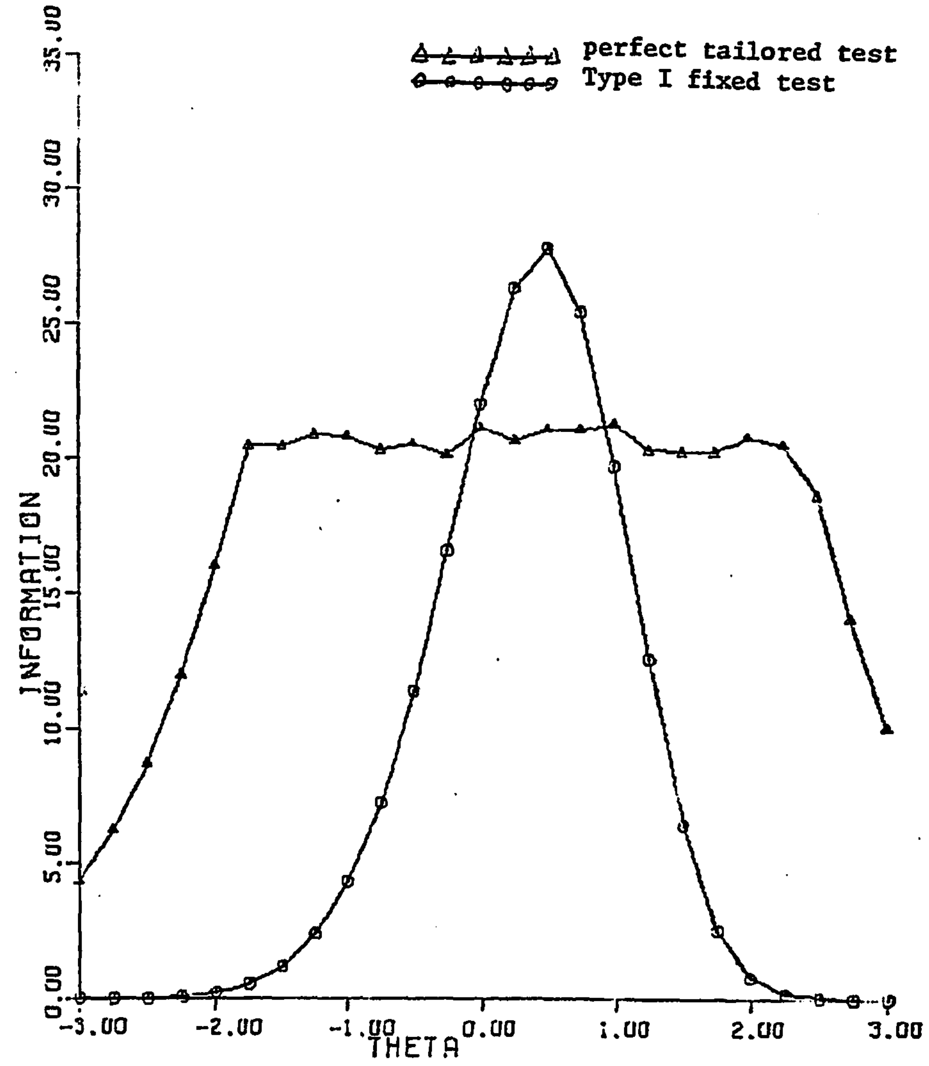

Figure 11. Information functions for perfect tailored tests, and Type I fixed tests constructed from Urry's item bank.

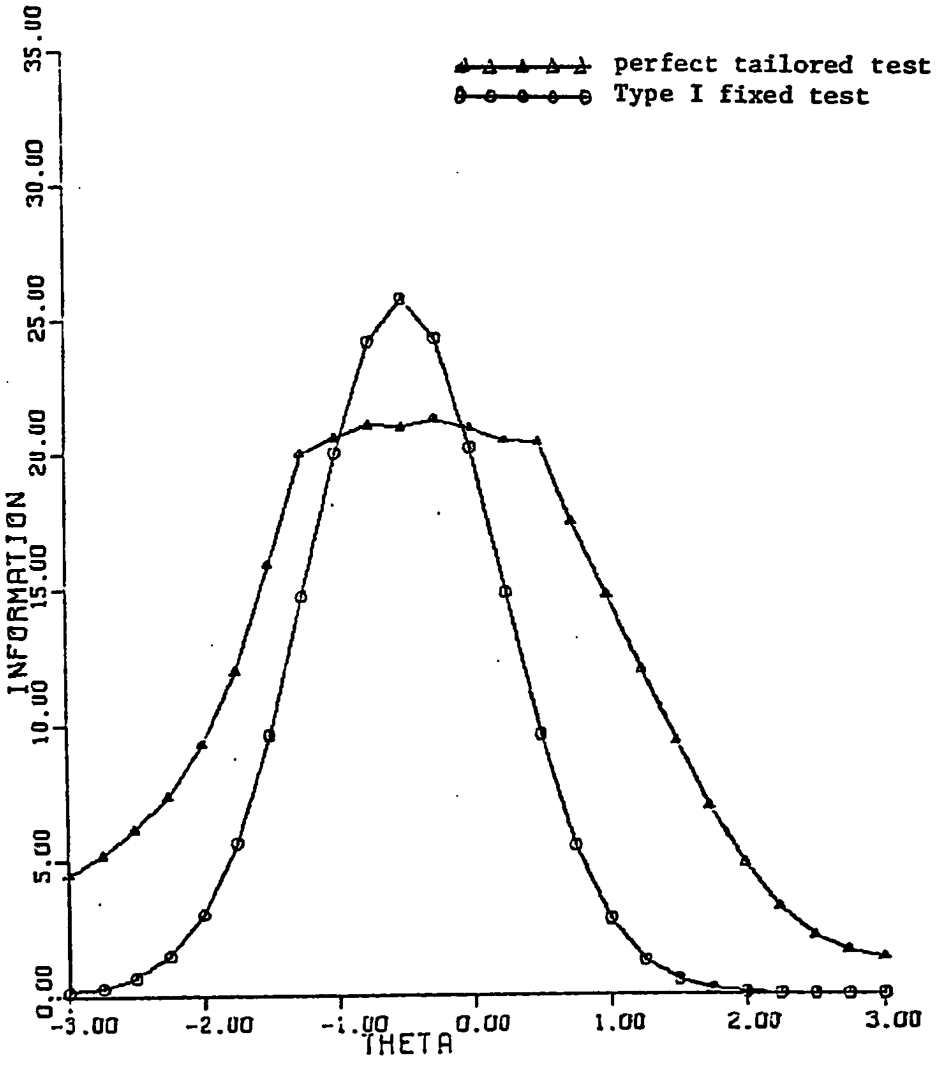

Figure 12. Information functions for perfect tailored tests, and Type I fixed tests constructed from Koch-ReKase's item bank.

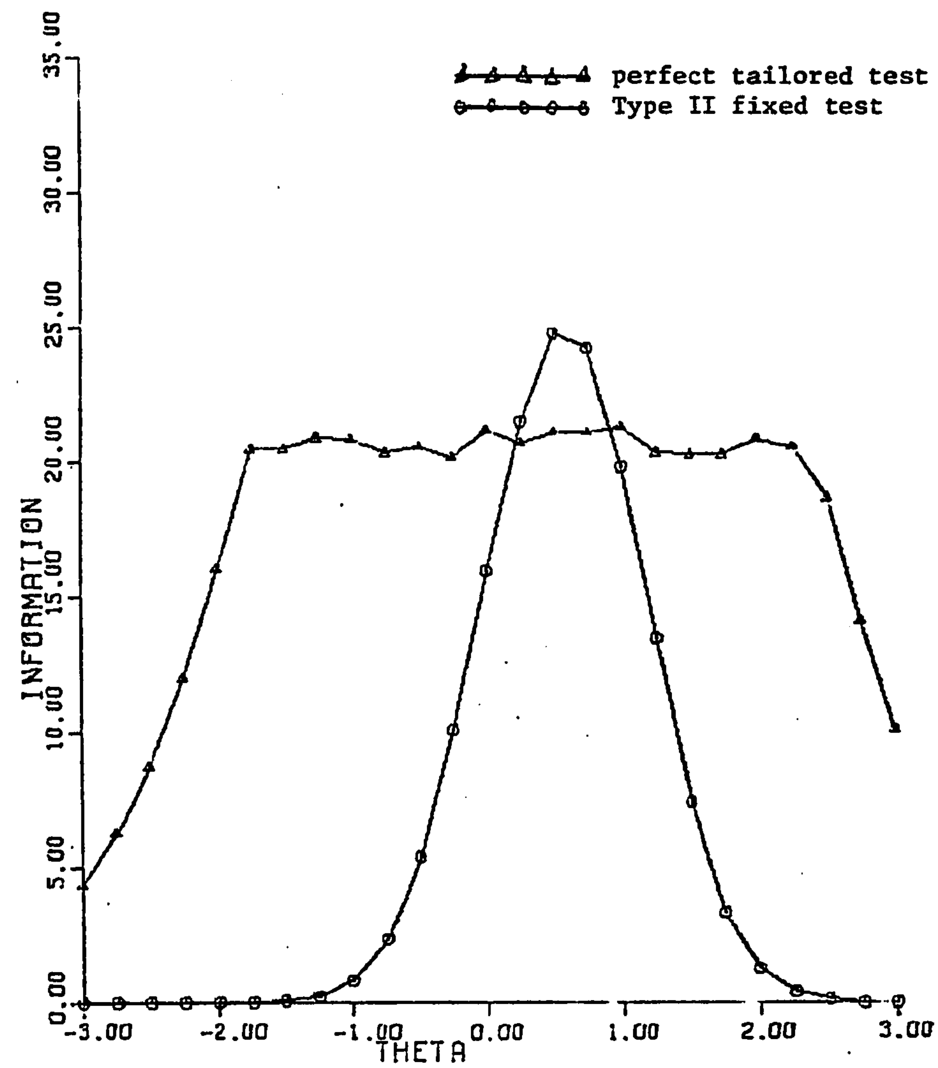

Figure 13. Information functions for perfect tailored tests, and Type II fixed tests constructed from Urry's item bank.

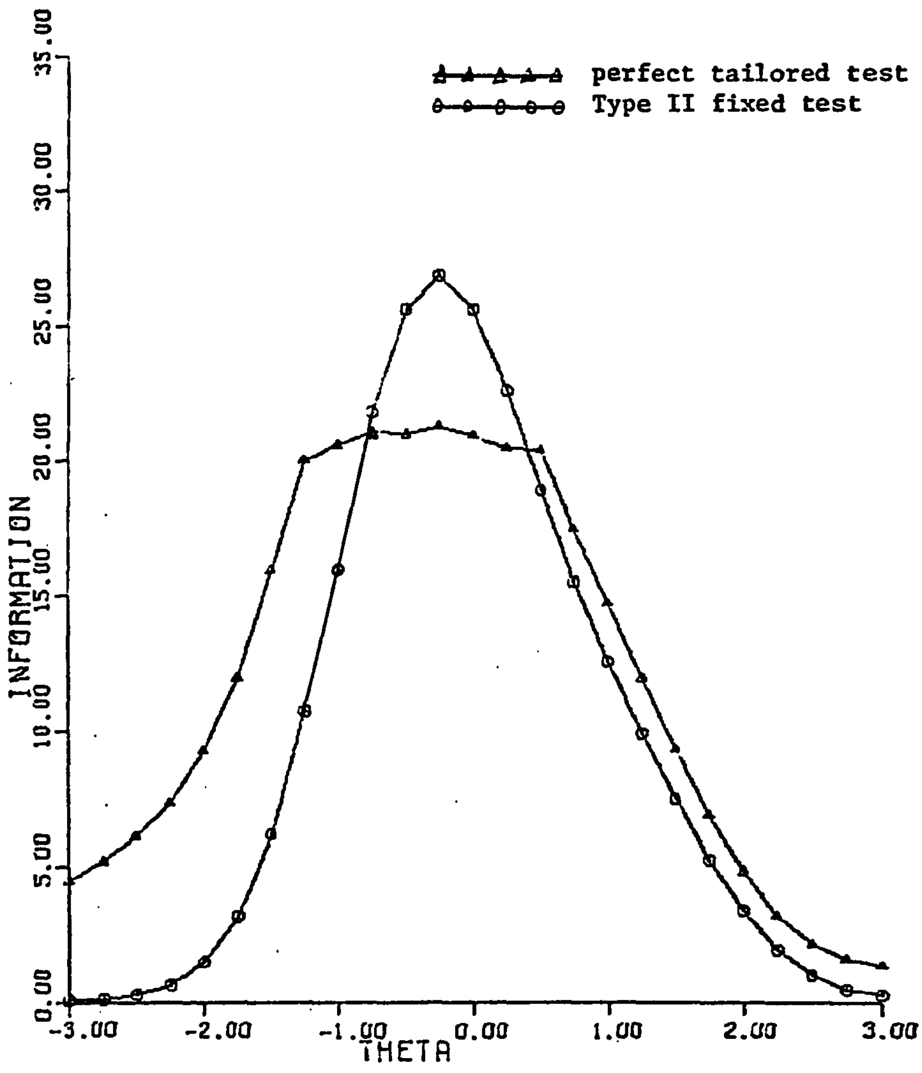

Figure 14. Information functions for perfect tailored tests, and Type II fixed tests constructed from Koch-ReKase's item bank.

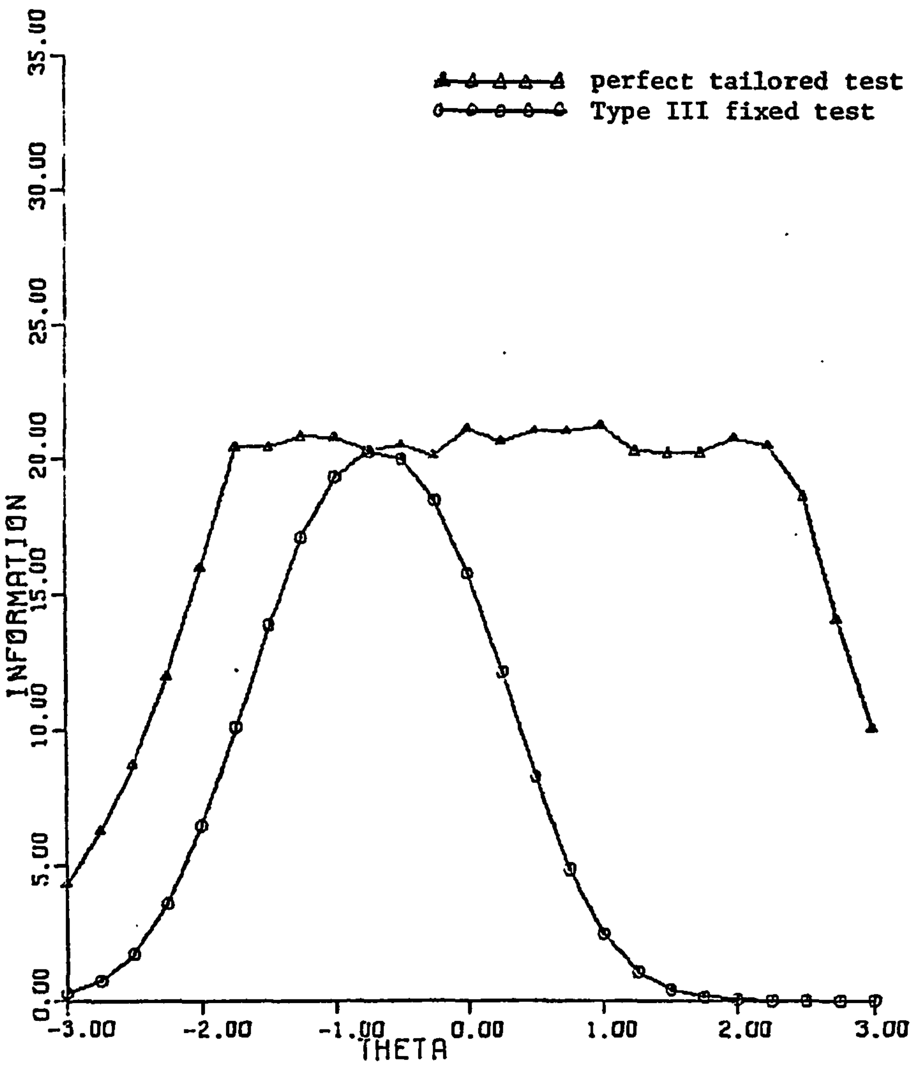

Figure 15. Information functions for perfect tailored tests, and Type III fixed tests constructed from Urry's item bank.

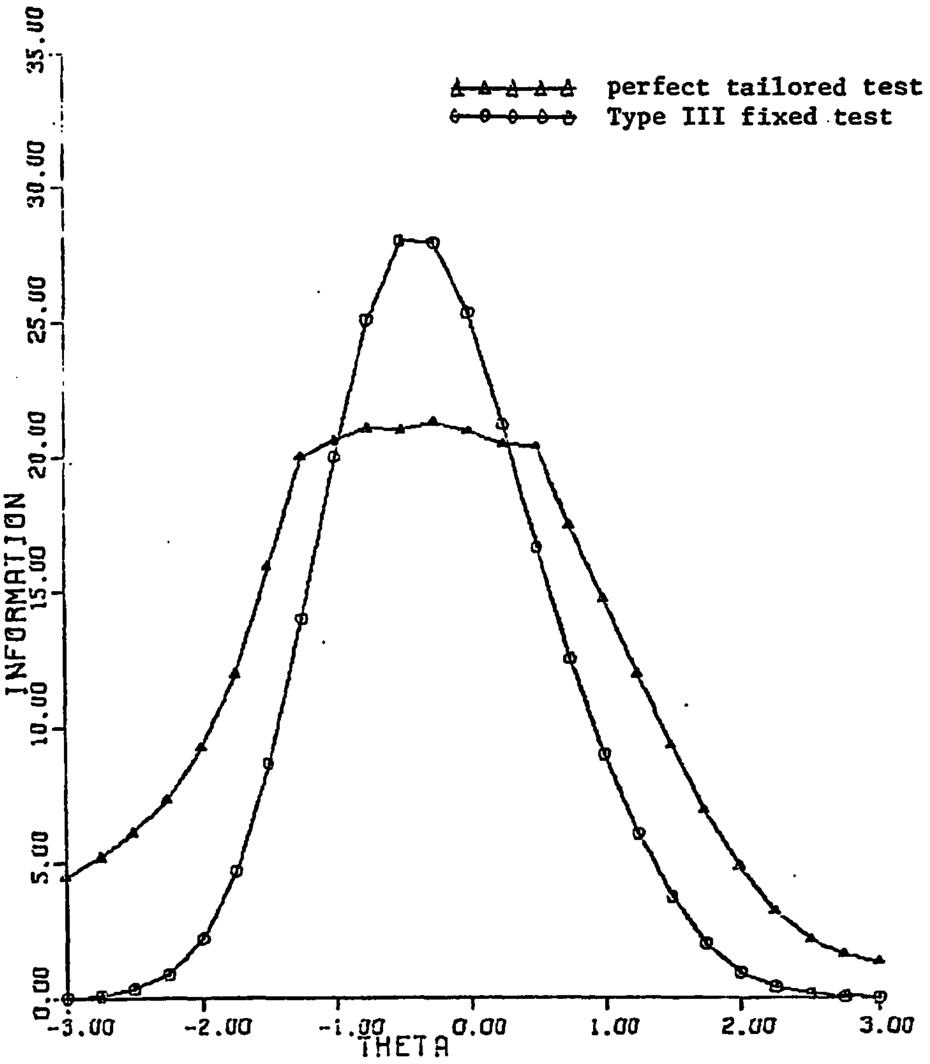

Figure 16. Information functions for perfect tailored tests, and Type III fixed tests constructed from Koch-ReKase's item bank.