BAYESIAN TAILORED TESTING AND THE INFLUENCE OF ITEM BANK CHARACTERISTICS [1]

CARL J. JENSEMA
Gallaudet College

Conventional tests are generally constructed to discriminate over a rather wide range of examinee ability. One of the consequences of this approach is that a conventional test usually contains many items which are not appropriate for a particular level of ability. Psychometricians have long been aware of this and in recent years they have increasingly turned their attention to the possibility of programming computers to design and administer tests.

Of the many computerized testing methods which have been proposed, the Bayesian process developed by Owen (1969) seems to be the most elegant and intuitively appealing method. It assumes locally independent binarily scored items and a normal ogive model (Lord and Novick, 1968, Ch. 16) in which the probability of passing a free response item g at ability level θ is expressed as

$$P_g(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a_g(\theta - b_g)} \exp\left[\frac{-t^2}{2}\right] dt \tag{1}$$

If the item is not of the free response type and c_g is the probability of guessing correctly, the probability of passing becomes

$$P_g'(\theta) = P_g(\theta) + c_g\,[1 - P_g(\theta)] \tag{2}$$

The derivation of Owen's Bayesian tailoring process has been described several times in the literature (Owen, 1969; Urry, 1971; Jensema, 1974a). We will briefly run through the fundamental formulas here for the sake of completeness.

Suppose $N(\theta_0, {\sigma_0}^2)$ expresses our knowledge of an examinee having ability θ. If we administer free response item g, which has discrimination and difficulty parameters a and b, and if the examinee responds correctly, Bayes' theorem specifies that the information available is

$$P(\theta|1) = k\,P_g(\theta)\,(\sqrt{2\pi}\,\sigma_o)^{-1} \exp\left[\frac{-(\theta - \theta_o)^2}{2{\sigma_o}^2}\right] \tag{3}$$

where $P_g(\theta)$ is defined by (1) and k is such that

$$\int_{-\infty}^{\infty} P(\theta|1)\,d\theta = 1. \tag{4}$$

The solution is

$$k^{-1} = 1/2\,(1 - \operatorname{erf} D) \tag{5}$$

where erf D is the error function

$$\operatorname{erf} D = \frac{2}{\sqrt{\pi}} \int_{0}^{D} \exp(-t^2)\,dt \tag{6}$$

and

$$D = \frac{b - \theta_o}{\sqrt{2\,(a^{-2} + {\sigma_o}^2)}}\,. \tag{7}$$

The expectation of the posterior mean is

$$E(\theta|1) = \theta_o + \frac{\sqrt{2}\,{\sigma_o}^2}{\sqrt{\pi\,(a^{-2} + {\sigma_o}^2)}} \exp(-D^2)\,(1 - \operatorname{erf} D)^{-1} \tag{8}$$

and the variance is

$$\operatorname{var}(\theta|1) = \sigma_o^2 \left[1 - \frac{\frac{2}{\sqrt{\pi}}\,2D \exp(D^2)\,(1 - \operatorname{erf} D)}{\sqrt{\pi}(1 + a^{-2}\sigma_o^{-2})\,(\exp(D^2)\,(1 - \operatorname{erf} D))^2}\right] \tag{9}$$

Similarly, if the examinee gives a wrong response to item g we have

$$P(\theta|O) = \frac{k}{k - 1}(1 - P_g(\theta))\,(\sqrt{2\pi}\sigma_o)^{-1} \exp\left[\frac{-(\theta - \theta_o)^2}{2{\sigma_o}^2}\right], \tag{10}$$

$$E(\theta|O) = \theta_o - \frac{\sqrt{2{\sigma_o}^2}}{\sqrt{\pi\,(a^{-2} + {\sigma_o}^2)}} \exp(-D^2)\,(1 + \operatorname{erf} D)^{-1}, \tag{11}$$

[1] This research was supported by the Office of Demographic Studies, Gallaudet College, Washington, D. C., 20002.

and

$$\operatorname{var}(\theta|O) = \sigma_o^2 \left[1 - \frac{\frac{2}{\sqrt{\pi}} + 2D \exp(D^2)\ (1 + \operatorname{erf} D)}{\sqrt{\pi}\,(1 + a^{-2}\sigma_o^{-2})\ (\exp(D^2)\ (1 + \operatorname{erf} D))^2} \right] \tag{12}$$

To expand this discussion a little further assume that item g is not a free response item and that it has a probability C_g of guessing correctly. If the examinee gives a correct response we have

$$P'(\theta|1) = \lambda P_g'(\theta)\ (\sqrt{2\pi}\,\sigma_o)^{-1} \exp\left[\frac{-(\theta - \theta_o)^2}{2\sigma_o^2} \right], \tag{13}$$

$$E'(\theta|1) = \theta_o + (1 - C_g)\, k^{-1}\, \lambda S, \tag{14}$$

and

$$\operatorname{var}'(\theta|1) = \sigma_o^2 - (1 - C_g)\, k^{-1}\, \lambda S^2\ \ (t - C_g\, \lambda) \tag{15}$$

where the prime is used to signify the effect of guessing, $P_g'(\theta)$ is defined by (2), and we take

$$\lambda^{-1} = C_g + (1 - C_g)\, k^{-1}, \tag{16}$$

$$S = k\,\sigma_o \exp(-D^2)\ (2\pi(1 + a^{-2}\sigma_o^{-2}))^{-1/2} \tag{17}$$

$$t = 1 - 2\sqrt{\pi}\, k^{-1}\, D \exp(D^2). \tag{18}$$

If the examinee gives a wrong response the formulas in (10), (11), and (12) hold, since our information, that the examinee does not know the correct answer, is the same as in the free response case.

Now assume we have n items and want to select the best one for administration. The expected posterior variance of θ after administration of a particular item is

$$E_u(\operatorname{var}(\theta|u)) = \theta_o^2 + \sigma_o^2 - P(O)\ [E(\theta|O)]^2 - P(1)\ [E(\theta|1)]^2$$

$$= \sigma_o^2 \left[1 - \frac{2}{\pi(1 + a^{-2}\sigma_o^{-2}) \exp(2D^2)\ (1 - (\operatorname{erf} D)^2)} \right] \tag{19}$$

when items are of the free response type and

$$E_u'(\operatorname{var}(\theta|u)) = \sigma_o^2 \left[1 - \frac{(1 - C_g)\,\lambda\,(1 + C_g\,(1 - k^{-1}))}{2\pi(1 + \sigma_o^{-2}a^{-2})\ (1 - k^{-1}) \exp(2D^2)} \right] \tag{20}$$

when the items are affected by guessing. In (19) and (20) u refers to the correctness of the examinee's response and is taken as 1 or 0. The item which leads to the smallest expected posterior variance is the most desirable one to administer. It is sufficient to select the item with the smallest value α where

$$\alpha = (a^{-2} + \sigma_o^2)\ (1 - (\operatorname{erf} D)^2) \exp(2D^2) \tag{21}$$

for free response items and

$$\alpha' = \left(\frac{1}{1 - C_g}\right)(1 + \sigma_o^{-2}a^{-2})\ (1 - k^{-1})\,\lambda^{-1} \exp(2D^2). \tag{22}$$

when guessing is present.

If we have a pool of n items and estimates of the normal ogive model parameters for each item, we may use a Bayesian sequential procedure to select items for administration to a particular examinee. Let $\hat{\theta}_{(m)}$ and $\hat{\sigma}^2_{(m)}$ be an estimate of the examinee's ability and its variance where m indicates the number of items administered. Assume the population has ability distributed as $N(0,1)$ and take $\hat{\theta}_{(o)}$ and $\hat{\sigma}^2_{(o)}$ as 0 and 1. Calculate α_i values for all (unused) items, $i = 1,2,\ldots,(n-m)$, using (22). (We will assume that the items are not free-response.) The examinee is administered the item with the smallest α_i value. If an incorrect response is given, $\hat{\theta}_{(m+1)}$ and $\hat{\sigma}^2_{(m+1)}$ are calculated from (11) and (12). If the response is correct, (14) and (15) are used. This cycle is repeated until $\hat{\sigma}_{(m)}$ is within some

pre-selected limit. The selection of a $\hat{\sigma}_{(m)}$ value for termination is, of course, arbitrary. It is usually selected to yield some expected level of validity according to

$$r_{\theta\hat{\theta}} = \sqrt{1 - \hat{\sigma}^2_{(m)}} \tag{23}$$

The characteristics of an item bank used for tailored testing are very important to the efficiency and accuracy of the process. There are four basic requirements for a good item bank. These have been mentioned in whole or part in a number of publications (i.e. Urry, 1970, 1971, 1971b, 1974; Jensema, 1972, 1974a, 1974b; etc.) and may be summarized as follows:

1) Item discrimination should be as high as possible and should not be less than .8.
2) Item guessing probabilities should be as low as possible.
3) The item bank must consist of a sufficiently large number of items.
4) Item difficulties should have a rectangular distribution.

The remainder of this paper will concentrate on demonstrating the importance of each of these four requirements.

Assume that an infinitely large item bank exists and that all items have the same discriminatory power and the same probability of guessing correctly. The assumption of an infinitely large item bank allows the selection of an item i having a difficulty level exactly equal to any given estimate of ability. When this can be done many of the formulas may be greatly simplified since we have:

$$D_i = 0 \tag{24}$$

and

$$\text{erf}\, D_i = 0. \tag{25}$$

The equations for $\hat{\sigma}^2_{(m+1)}$ for correct and incorrect responses become

$$\hat{\sigma}^2_{(m+1)} = \hat{\sigma}^2_{(m)} \left[1 - \frac{2t_i (1 - C_i)^2}{\pi (1 - C_i)^2} \right] \tag{26}$$

and

$$\hat{\sigma}^2_{(m+1)} = \hat{\sigma}^2_{(m)} \left[1 - \frac{2t_i}{\pi} \right] \tag{27}$$

where m is the number of items previously administered.

An item i's difficulty is the point at which the probability of knowing the correct answer is exactly .5. If guessing is in effect the probability of responding correctly is equal to the probability of knowing the answer plus the probability of guessing correctly. Then $\hat{\sigma}^2_{(m+1)}$ may be expected to be the sum of (26) and (27) weighted by the probabilities of a correct or incorrect response:

$$E\hat{\sigma}^2_{(m+1)} = \hat{\sigma}^2_{(m)} \left[.5 (1 + C_i) \left(1 - \frac{2t_i (1 - C_i)^2}{\pi (1 + C_i)^2} \right) + .5 (1 - C_i) \left(1 - \frac{2t_i}{\pi} \right) \right] \tag{28}$$

A little algebraic manipulation reduces this to

$$E\hat{\sigma}^2_{(m+1)} = \hat{\sigma}^2_{(m)} \left[1 - \frac{2t_i (1 - C_i)}{\pi (1 + C_i)} \right] \tag{29}$$

Inserting appropriate values for a_i and c_i in equation (29) and plotting the results against the number of items administered demonstrates the influence of item discrimination and guessing probability on the tailoring process. Figure 1 plots the expected standard error of the estimate $E\hat{\sigma}_{(m+1)}$ by the number of items administered for five levels of discrimination when guessing probability is zero and an infinite number of items are available. Notice the sharp difference in the number of items needed at different levels of discrimination. For example, if the items have discriminatory powers of 2.5 only 4 or 5 items are needed to reach a standard error of the estimate of .30 while 17 or 18 items are needed to reach this level when item discrimination is only 1.0.

Now suppose we take item discrimination to be 1.0, a rather low value which is easily obtained. Figure 2 plots the expected standard error of the estimate for various guessing values by the number of items administered. The guessing values range from .5 (i.e. true-false items) to 0.0 (i.e. free response items.) The greater the probability of guessing, the more items required to reach a specific standard error of the estimate.

To give a clear example of the combined effects of discrimination and guessing on the tailoring process, suppose we have three item banks which, for convenience, are referred to as I, II, and III. Assume Bank I items have discrimination and guessing paramenters of .5 and .33. Bank II's parameters are 1.0 and .25 while Bank III has parameter values of 2.0 and .20. These banks may be roughly classified as *unacceptable, fair,* and *excellent* for tailored testing purposes. Assuming that each bank has an infinite number of items and plotting the expected standard error of the estimate against the number of items administered, the three curves in Figure 3 are obtained.

In Figure 3, notice that Bank I would give unacceptable results. After 30 items the expected standard error of the

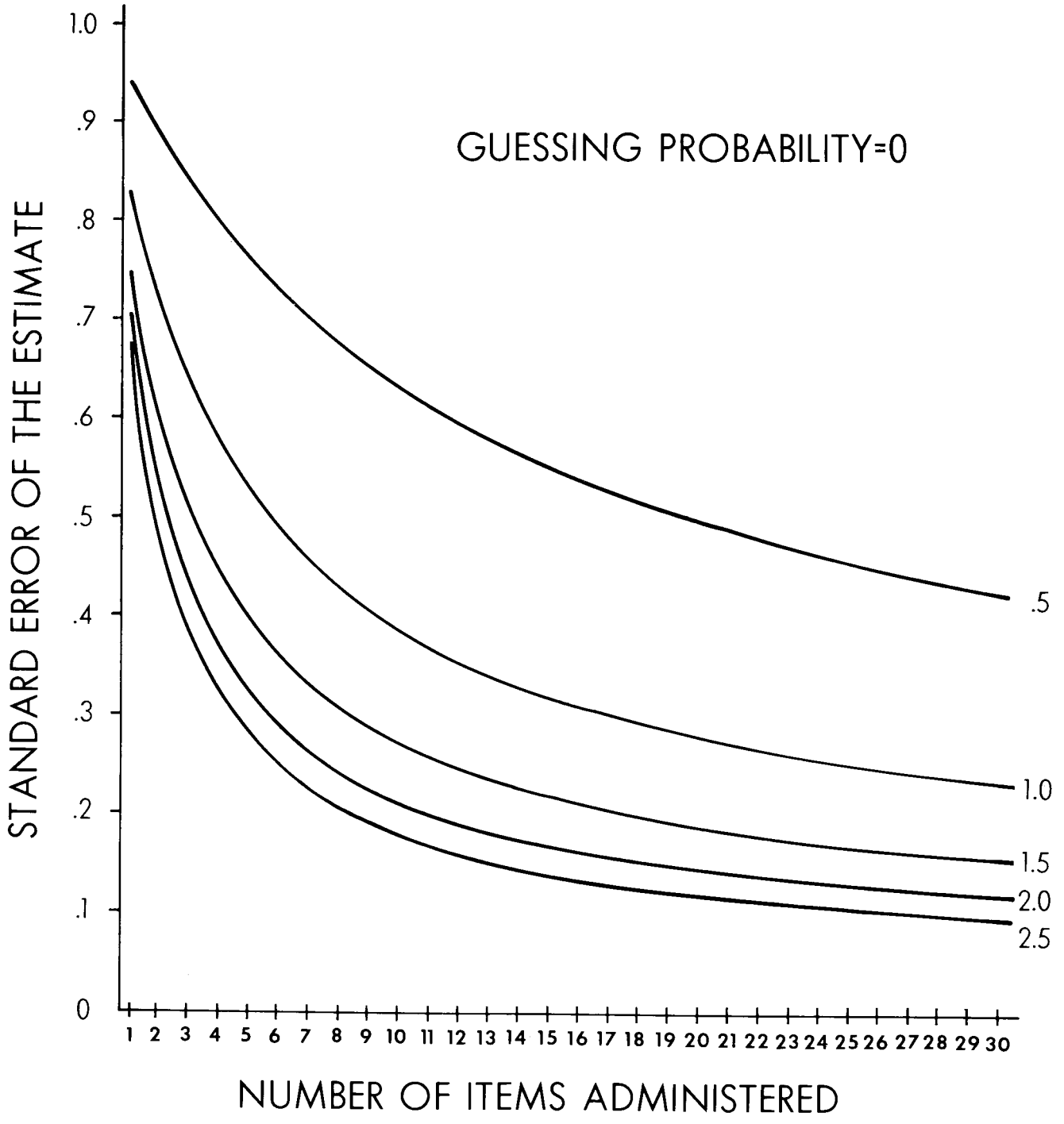

Figure 1. Expected standard error of the estimate according to number of items administered at five levels of item discrimination.

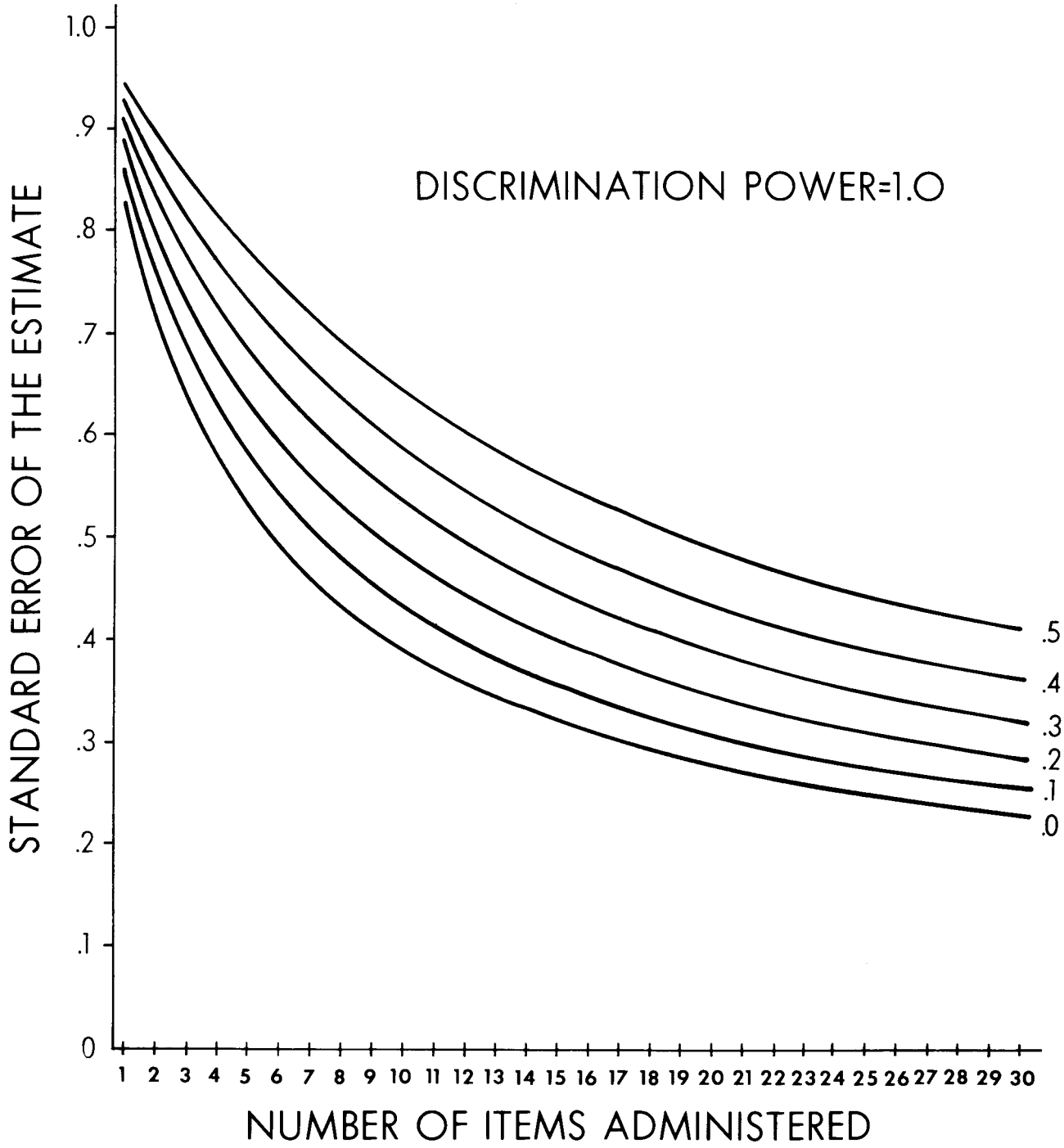

Figure 2. Expected standard error of the estimate according to number of items administered at six guessing probabilities.

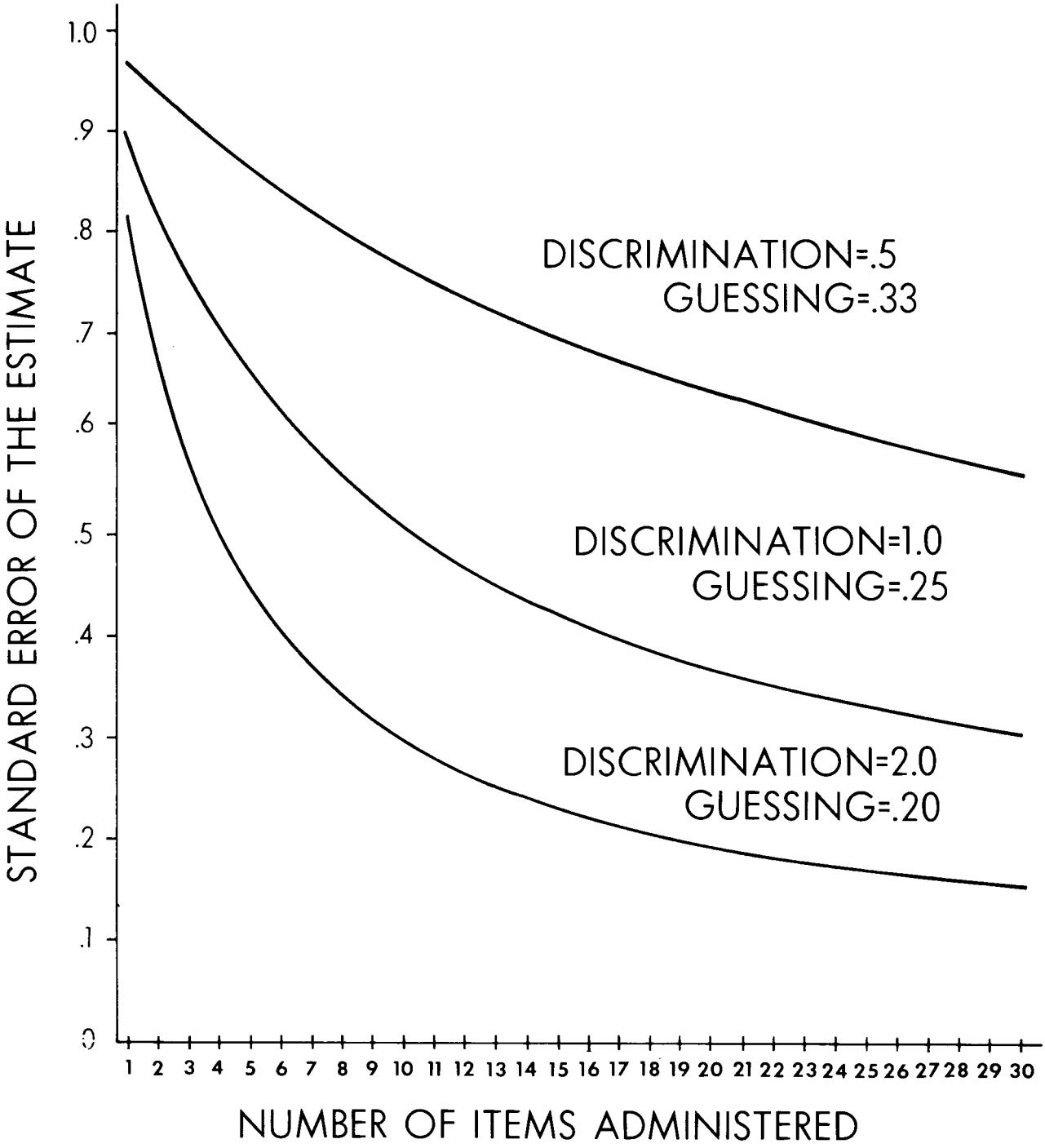

Figure 3. Expected standard error of the estimate for three item banks according to number of items administered.

estimate is only .56 (i.e. reliability = .69, validity = .83). In contrast an excellent item bank, such as Bank III, would reach this level after only 3 or 4 items. The advantage of high discrimination and low guessing probability in an item bank is obvious.

Up to this point we have discussed the behavior of Bayesian tailored testing when the item bank is assumed to be of unlimited size. The obvious question which follows is what happens when item bank sizes are within practical limits? To answer this question, Monte-Carlo data for 200 items are generated for each of 100 "examinees" using Urry's (1970) "LOGIST" program. The parameters for discrimination (1.0) and guessing (.25) were the same as for Bank II mentioned earlier. Eight sets of 25 difficulty values (-2.4, -2.2, ..., 0.0, ..., 2.2, 2.4) were employed. Bayesian tailored testing was simulated with this data using 50, 75, 100, 150, and 200 items in the bank. Since difficulty had been specified in sets of 25 values, the item banks had 2, 3, 4, 6, and 8 items at each of the 25 difficulty levels respectively.

For each of the five item banks and for each of the 100 examinees, tailoring was simulated until 30 items had been "administered". As each item was "administered" the new estimate of ability was recorded. Since the data was randomly generated, true ability (distributed as $N(0,1)$ was known and could be correlated with estimated ability. Table I gives the validity (correlation between true and estimated ability) for each item bank by the number of items "administered". The last column in Table I gives the expected validities for an item bank of infinite size as calculated from equation (32) and (23).

The Monte-Carlo data above represents items which are passable but not especially good for tailored testing. To see how item bank size would influence validity when the bank was composed of excellent items, the Monte-Carlo data tailoring simulation was repeated with higher discrimination

TABLE 1

Validity ($r_{\theta\hat{\theta}}$) Obtained With Different Size Item Banks (Monte-Carlo Data, N=100, A=1.0, C=.25)

Items Administered	ITEMS IN BANK 50	75	100	150	200	∞*
1	.53	.53	.53	.53	.53	.44
2	.59	.59	.59	.59	.59	.57
3	.65	.65	.65	.65	.65	.66
4	.72	.72	.72	.72	.72	.72
5	.78	.78	.78	.78	.78	.76
6	.81	.80	.80	.80	.80	.79
7	.83	.82	.82	.82	.82	.81
8	.84	.84	.84	.84	.84	.83
9	.85	.85	.84	.84	.84	.85
10	.86	.86	.86	.85	.85	.86
11	.86	.87	.88	.87	.87	.87
12	.87	.87	.89	.87	.87	.88
13	.89	.89	.89	.87	.88	.89
14	.90	.91	.90	.88	.88	.90
15	.91	.91	.91	.90	.90	.91
16	.91	.92	.92	.91	.91	.91
17	.92	.92	.92	.92	.91	.92
18	.92	.92	.93	.92	.92	.92
19	.92	.92	.93	.92	.92	.93
20	.93	.93	.93	.93	.93	.93
21	.93	.93	.93	.93	.93	.93
22	.93	.94	.94	.94	.93	.94
23	.93	.94	.94	.94	.94	.94
24	.93	.94	.94	.94	.94	.94
25	.93	.94	.95	.94	.94	.94
26	.94	.95	.95	.94	.94	.95
27	.94	.95	.95	.94	.95	.95
28	.94	.95	.95	.95	.95	.95
29	.94	.95	.95	.95	.95	.95
30	.94	.95	.95	.95	.95	.95

*Expected validities calculated from equations (32) and (23) for an imaginary bank having an infinite number of items.

TABLE 2

Validity ($r_{\theta\hat{\theta}}$) Obtained With Different Item Bank Sizes (Monte-Carlo Data, N=100, A=2.0, C=.2)

Items Administered	ITEMS IN BANK 50	75	100	150	200	∞*
1	.66	.66	.66	.66	.66	.58
2	.75	.75	.75	.75	.75	.74
3	.84	.84	.84	.84	.84	.82
4	.89	.89	.89	.89	.89	.86
5	.92	.92	.92	.92	.92	.90
6	.93	.93	.93	.93	.93	.91
7	.94	.94	.94	.94	.94	.93
8	.95	.95	.95	.95	.95	.94
9	.96	.95	.95	.95	.95	.95
10	.96	.96	.96	.96	.96	.96
11	.97	.96	.96	.96	.96	.96
12	.97	.96	.96	.96	.97	.96
13	.97	.97	.97	.97	.97	.97
14	.97	.97	.97	.97	.97	.97
15	.97	.97	.98	.97	.98	.97
16	.97	.98	.98	.98	.98	.98
17	.97	.98	.98	.98	.98	.98
18	.98	.98	.98	.98	.98	.98
19	.98	.98	.98	.98	.98	.98
20	.98	.98	.98	.98	.98	.98
21	.98	.98	.98	.98	.98	.98
22	.98	.98	.99	.98	.98	.98
23	.98	.98	.99	.98	.98	.98
24	.98	.98	.99	.98	.98	.98
25	.98	.98	.99	.99	.99	.98
26	.98	.98	.99	.99	.99	.99
27	.98	.98	.99	.99	.99	.99
28	.98	.98	.99	.99	.99	.99
29	.98	.98	.99	.99	.99	.99
30	.98	.98	.99	.99	.99	.99

*Expected validities calculated from equations (32) and (23) for an imaginary bank having an infinite number of items.

(2.0) and lower guessing (.20) parameter values. These configurations correspond to Bank III mentioned earlier. The results of the simulated tailoring with this new data are given in Table 2.

For practical application it is apparent that a very large number of items is not a critical item bank characteristic if the bank is good in other respects. In both Table 1 and Table 2 the Monte-Carlo data validities obtained for the five banks closely match each other and they also parallel the validities to be expected from a corresponding item bank of infinite size. However, it must be remembered that this was Monte-Carlo data and the tailoring simulation used known parameter values for discrimination, difficulty, and guessing. With real data involving imprecise parameter estimates and a possible non-uniform distribution of difficulty, it would be wise to be a bit cautious if a bank had, say, fewer than 75 items. In connection with this, there are some practical problems which arise if an item bank is too large. A large bank has more items available for administration, but the storage requirements and the increased computer processing needed for item selection also slow things down while adding to overall computer costs. (Some good cost-efficiency studies are needed on this!)

The last item bank requirement is uniform distribution of difficulty. The exact results of violating this rule are difficult to predict, since they would necessarily depend on the actual distribution of item difficulty, the discrimination and guessing parameter values, the number of items in the bank, and the criteria used to terminate the tailoring process. The essential point to remember is that the Bayesian tailoring procedure attempts to select for administration the item which will yield the most information. If, at a particular level of difficulty, there are no items available, the Bayesian process will be forced to select an item which is not appropriate and which will yield less than an optimal amount of information.

To summarize, this paper has outlined a Bayesian approach to item selection for tailored testing. Four basic requirements of a good item bank for this process have been discussed. If these requirements are met, Bayesian tailored testing will yield excellent results. The key to the process lies in careful construction of item banks. If attention is given to this, the Bayesian tailoring process gives us a fundamental tool for practical application of latent trait mental test theory.

REFERENCES

Jensema, C. J. An application of latent trait mental test theory to the Washington pre-college testing battery. Unpublished doctoral dissertation. University of Washington, 1972.

Jensema, C. J. An application of latent trait mental test theory. *Br. J. Math. Statist. Psychol.* 27, 29-48, 1974a.

Jensema, C. J. The validity of Bayesian tailored testing. *Ed. and Psychol. Meas.* 34, 757-766, 1974b.

Lord, F. M. and Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison–Wesley, 1968.

Owen, R. J. A Bayesian approach to tailored testing. *Res. Bull.* 69-92. Princeton, N. J.: Educational Testing Service, 1969.

Urry, V. W. A Monte-Carlo investigation of logistic mental test models. Unpublished doctoral dissertation. Purdue University, 1970.

Urry, V. W. Approximation methods for the item parameters of mental test models. *Res. Bull.* 0871-202. Seattle: University of Washington, Bureau of Testing, 1971a.

Urry, V. W. Individualized testing by Bayesian estimation. *Res. Bull.* 0171-177. Seattle: University of Washington, Bureau of Testing. 1971b.

Urry, V. W. Approximations of item parameters of mental test models and their uses. *Ed. and Psychol. Meas.* 34, 353-369, 1974.