KM 3(1981)
pag 17-34

Estimating abilities: inference for random variables*)

Charles Lewis**)

Summary

A nonparametric definition of an individual's ability on a unidimensional scale based on binary items is proposed. Either Mokken or Rasch scales may be used. Formal Bayesian inference for abilities so defined is developed and attention is given to the problem of choosing an appropriate prior. Sensitivity of posterior inferences to a) choice of prior and b) response pattern is discussed and illustrated with artificial examples.

Introduction

The concept of a latent trait (simply referred to as an ability in the following) lies at the heart of much of psychometric theory. Consequently, the continuing arguments over the status of this concept, though largely confined to the domain of factor analysis and factor scores, are also relevant for psychometrics as a whole.

It is in the context of mental test theory that the essentially statistical nature of an ability (or true score) has been most clearly developed: from early definitions as the score on a test of infinite length, to theories about the mean observed score over

---

*) Paper presented at the Ledyard R Tucker Symposium on Psychometric Theory, University of Illinois at Urbana-Champaign, U.S.A., 24-25 April 1980. Other papers on the subject presented at the 4th International Symposium on Educational Testing, Antwerp, Belgium, 24-27 June 1980 (to appear in a book of proceedings) and at the spring meeting of the social science section of the VVS, Utrecht, 20 May, 1981

**) Vakgroep Statistiek en Meettheorie, FSW, Rijksuniversiteit Groningen.

a population of replications, occasions, or items, to modern consistency proofs for estimation in latent trait models, where both the number of individuals and the number of items are allowed to approach infinity. Although each of these developments is in harmony with the limiting relative frequency ideas on which standard statistical theory is based, their practical relevance for the common situation in which a limited group of individuals is tested once, using a limited set of items, may well be questioned.

In the following sections, an alternative statistical treatment of ability will be discussed and its implications for inference explored. The general framework adopted in this treatment is a Bayesian one. Consequently, there is no need for any reference to infinite sequences of observations or, indeed, to any observations other than those actually made, when considering an unknown quantity such as an ability. Instead, prior information regarding the value of the ability for a given individual is expressed in terms of a probability distribution and is combined via a model and Bayes' Theorem with the responses of that individual to a given set of items. The result is a modified probability distribution which describes the total information available regarding the individual's ability. It is in this sense, rather than in sampling terms, that the true value of the ability is considered to be a random variable.

Although a Bayesian approach to making inferences about abilities is quite distinct from standard treatments, it is hardly new. Developments to date, however, have typically used highly restrictive formulations of a model relating ability to responses, the author's joint work (Lewis, Wang, & Novick, 1975) being a case in point. While restrictive models may have their place at the level of test development, as a guide to the selection and modification of items, these same models become less attractive when the goal is to use an already developed test to provide information about individual abilities. In the latter case, model violations are both more difficult to detect and more likely to have serious consequences.

As a result of these concerns, a "nonparametric" latent trait model is adopted in the following sections, and used to provide Bayesian inferences about abilities. It should be noted that the model itself is also not new. It appears in the isotonic regression literature (see, for instance, Barlow, et al.,1972,and a further developed version is employed in a non-Bayesian approach to scale construction proposed by Mokken (1971). The author (Lewis, 1970) has also considered this model from a non-Bayesian point of view. A recent Bayesian analysis of a related model is given by Stewart (1979). Neither he nor the workers in the isotonic regression field give any attention to a quantity analogous to an ability, however. Thus, what is new in the following is the possibility of making Bayesian inferences about individual abilities while placing minimal restrictions on the form of the relation between ability and responses.

Model and theoretical development

Suppose there are k dichotomous items under consideration and that, for the individual of interest, the probability of success on item i is denoted by π_i. There are two basic assumptions used in the following. First, assume it is possible to order the items *a priori* according to their difficulty for the individual so that

$$1 > \pi_1 > \pi_2 > \dots > \pi_k > 0. \tag{1}$$

Second, assume that, given $\{\pi_i\}$, the individual's responses to the items are mutually independent.

Now a definition of the individual's ability on the scale defined by these items is required. If θ is used to denote this ability, let

$$\theta = \begin{cases} 0 \text{ if } \pi_1 < .5 \quad , \\ i \text{ if } \pi_i \geq .5 > \pi_{i+1}, \text{ for } i=1, \dots, k-1, \text{ and} \\ k \text{ if } \pi_k \geq .5 \quad . \end{cases} \tag{2}$$

Thus θ is the number of items for which the probability of success is at least .5. Note that this is a purely ordinal definition and does not even assume an underlying continuous scale. Different ability levels acquire meaning from the content of the items used. Also, although this model is "nonparametric", it is compatible with a "parametric" latent trait model (such as the Rasch model) which assumes parallel item characteristic curves. In such a model, the items are ordered in difficulty for all individuals once the order of the item difficulty parameters is known. Moreover, θ as defined in (2) simply locates the parametric ability in one of the k+1 intervals defined by the difficulty parameters of adjacent items.

Although primary interest is in making inferences about θ, the formal analysis must take place at the level of $\{\pi_i\}$. Let y_i equal the number of correct responses out of n_i total responses to item _i_. The most obvious case will be where all n_i equal unity, though there is no reason to exclude the possibility of additional replications using items which appear to be (at least roughly) equivalent. The case where n_i equals zero for one or more items should also be allowed. These would be items which, for one reason or another, were not answered by the individual. If these items, nonetheless, form part of the scale, there is no reason to exclude them from the analysis. These extensions might both be relevant, for example, in the case of tailored testing.

In terms of the definitions and assumptions given so far, the likelihood for $\{\pi_i\}$ satisfying (1) is given by

$$\prod_{i=1}^{k} \pi_i^{y_i} (1-\pi_i)^{n_i-y_i} \quad . \tag{3}$$

Natural conjugate prior densities for $\{\pi_i\}$ of the form

$$C_0 \prod_{i=1}^{k} \pi_i^{\alpha_i-1} (1 - \pi_i)^{\beta_i-1} \quad , \tag{4}$$

with all α_i and β_i positive to assure integrability, seem to provide a sufficiently rich choice for practical work. (Guidelines for choosing the α_i and β_i will be discussed in the following section.) The prior in (4) may be thought of as a product of beta densities for the individual π_i, truncated to conform with the order restriction (1). As a consequence of the truncation, the π_i are not mutually independent. Their dependence, however, is a direct result of prior order information and does not involve any additional structural assumption. When Bayes' Theorem is applied to (3) and (4), the result is the joint posterior density for $\{\pi_i\}$ given the y_i and n_i:

$$C_1 \prod_{i=1}^{k} \pi_i^{y_i+\alpha_i-1}(1-\pi_i)^{n_i-y_i+\beta_i-1} . \tag{5}$$

In the present application of this analysis, what is needed in place of (4) and (5) are the prior and posterior distributions of the ability θ. From the definition (2), it is clear that

$$\begin{aligned} &\text{Prob } (\theta=0) = \text{Prob } (\pi_1 < .5), \\ &\text{Prob } (\theta=i) = \text{Prob } (\pi_i \geqslant .5 > \pi_{i+1}) \quad \text{for } i=1, \ldots, k-1, \text{ and} \\ &\text{Prob } (\theta=k) = \text{Prob } (\pi_k \geqslant .5). \end{aligned} \tag{6}$$

Thus, in general, repeated integration of (4) and (5) will be necessary to achieve the desired results. The following argument shows that this is not so formidable a task as might at first be thought.

To begin with, the problem may be generalized somewhat by considering a joint density for $\{\pi_i\}$ satisfying (1) which has the form

$$C \prod_{i=1}^{k} g_i(\pi_i) , \tag{7}$$

for a sequence $\{g_i\}$ of positive integrable functions defined on the interval [0,1]. Two additional sequences of functions based on $\{g_i\}$ are also required. Let

$$f_0(u) = 1,\ 0 \leq u \leq 1.$$

For i=1, ..., k, let

$$f_i(u) = \int_u^1 f_{i-1}(t)\ g_i(t)\ dt,\ 0 \leq u \leq 1\ . \qquad (8)$$

Also, let

$$h_{k+1}(u) = 1, \quad 0 \leq u \leq 1.$$

For i=k, k-1, ..., 1, let

$$h_i(u) = \int_0^u g_i(t)\ h_{i+1}(t)\ dt, \quad 0 \leq u \leq 1. \qquad (9)$$

These sequences provide two paths for successive integration of the joint density (7). Since the total integral of the density must equal unity,

$$f_k(0) = h_1(1) = C^{-1}\ .$$

Although it is not of direct interest in the present development, it may be noted that the marginal density for any π_i is given by

$$C\ f_{i-1}(\pi_i)\ g_i(\pi_i)\ h_{i+1}(\pi_i). \qquad (10)$$

The probability distribution for θ based on (7) may be expressed as

$$\text{Prob}\ (\theta=i) = C\ f_i(.5)\ h_{i+1}(.5)\ ,\ \text{for } i=0,\ 1,\ \ldots,k\ . \qquad (11)$$

Here integration has first been carried out with respect to all π_j for $j < 1$ or $j > i+1$. This gives the bivariate density for π_i and π_{i+1}. Finally, integration is carried out with respect to π_i for all values greater than or equal to .5, and with respect to

π_{i+1} for all values less than .5 . This gives

$$\text{Prob } (\pi_i \geq .5 \text{ and } \pi_{i+1} < .5)$$

which, by (6), is the desired probability that the ability θ equals $\underline{i}$. (Obvious modifications of the above argument are necessary for $\underline{i}$ equal to zero and $\underline{k}$.)

To illustrate, consider a test with $k = 4$ items. The joint density (7) may be written as

$$p(\pi_1,\pi_2,\pi_3,\pi_4) = C\, g_1(\pi_1)g_2(\pi_2)g_3(\pi_3)g_4(\pi_4), \tag{12}$$

for

$$1 > \pi_1 > \pi_2 > \pi_3 > \pi_4 > 0 \ .$$

Obtaining, for instance, the probabiltiy that θ equals 2 may be represented as

$$\begin{aligned}\text{Prob}(\theta=2) &= C\ \overbrace{\Big(\int_{.5}^{1}\overbrace{\Big(\int_{\pi_2}^{1} g_1(\pi_1)d\pi_1\Big)}^{f_1(\pi_2)} g_2(\pi_2)d\pi_2\Big)}^{f_2(.5)} \ \cdot \\ &\quad\ \underbrace{\Big(\int_{0}^{.5} g_3(\pi_3)\underbrace{\Big(\int_{0}^{\pi_3} g_4(\pi_4)d\pi_4\Big)}_{h_4(\pi_3)} d\pi_3\Big)}_{h_3(.5)} \\ &= C\ f_2(.5)h_3(.5) \ . \end{aligned} \tag{13}$$

In general, the integrations in (8) and (9) are most conveniently carried out by numerical means, rather than analytically. It is important to note that the amount of computational effort required is a <u>linear</u> function of the number of variables ($\underline{k}$), rather than exponential, as is often the case with multiple numerical integration. This is a direct result of the product structure of the joint density (7), and implies that it is computationally feasible to analyze responses for relatively long tests. The author has written an interactive program ABILITY, which carries out the necessary computations to obtain distributions for θ. The present version operates within reasonable time and accuracy limits for tests of up to 30 items (or 30 sets of equivalent items). A non-interactive "production version" could undoubtedly double this maximum.

Although simple analytic results regarding the distributions for θ are, in general, elusive, there are exceptations. One of these occurs for the prior distribution based on (4) with all $\alpha_i = \beta_i = 1$. This gives a uniform density for $\{\pi_i\}$ over the region (1), which happens to be a member of the ordered Dirichlet family (for which standard results are available). This density may be expressed in the general form (7) with all $g_i(u) = 1$ identically. For this case,

$$f_i(u) = (1-u)^i/(i!) \text{ and}$$

$$h_{i+1}(u) = u^{k-i}/(k-i)! \text{ , for } i=0, \ldots, k.$$

Thus

$$C^{-1} = f_k(0) = h_1(1) = (k!)^{-1}$$

and, using (11),

$$\text{Prob } (\theta=i) = \frac{k!}{i!(k-i)!}(.5)^k, \tag{14}$$

the probability distribution for a binomial random variable with parameters $\underline{k}$ and .5 .

In the case where k=4, for instance, the joint prior (12) becomes

$$p(\pi_1, \pi_2, \pi_3, \pi_4) = 24$$

in the region (1). Moreover,

$$f_1(\pi_2) = \int_{\pi_2}^{1} d\pi_1 = 1 - \pi_2 \ ,$$

$$f_2(.5) = \int_{.5}^{1} (1 - \pi_2) d\pi_2 = -\tfrac{1}{2}(1 - \pi_2)^2 \Big|_{.5}^{1} = (.5)^2/2 \ ,$$

$$h_4(\pi_3) = \int_0^{\pi_3} d\pi_4 = \pi_3 \ , \text{ and}$$

$$h_3(.5) = \int_0^{.5} \pi_3 d\pi_3 = \tfrac{1}{2}\pi_3^{\,2} \Big|_0^{.5} = (.5)^2/2 \ .$$

Thus, from (13),

$$\text{Prob } (\theta=2) = \frac{24}{2 \cdot 2} (.5)^4 = .375 \ ,$$

which agrees with (14) when i=2 and k=4.

In fact, the result (14) can be generalized to the prior for θ based on (4) for any case where all $\alpha_i = \beta_i = a > 0$. To see this, define a strictly increasing transformation $\phi(\pi)$ such that

$$\phi(0) = 0 \ ,$$
$$\phi(1) = 1 \ , \text{ and}$$
$$\frac{d\phi}{d\pi} \doteq b(\pi(1-\pi))^{a-1} \text{ for some } b > 0.$$

For symmetry considerations,

$$\phi(.5) = .5 \ . \tag{15}$$

Letting

$$\phi_i = \phi(\pi_i) \ , \text{ for } i=1, \ldots, k,$$

and transforming the prior (4) to a prior for $\{\phi_i\}$, it is clear that the result will be a uniform density over the region

$$1 > \phi_1 > \ldots > \phi_k > 0 \ .$$

Moreover, because of the relation (15), the definition of θ given in (2) may be restated, replacing π_i by ϕ_i. Consequently, the conditions leading to (14) have been reproduced exactly, and θ must have as prior a binomial (k, .5) distribution whenever the prior for $\{\pi_i\}$ has the form (4) with all α_i and β_i equal and positive. The importance of this result is that the same prior for θ may arise from a whole family of priors for $\{\pi_i\}$, a fact which will be exploited in the following section.

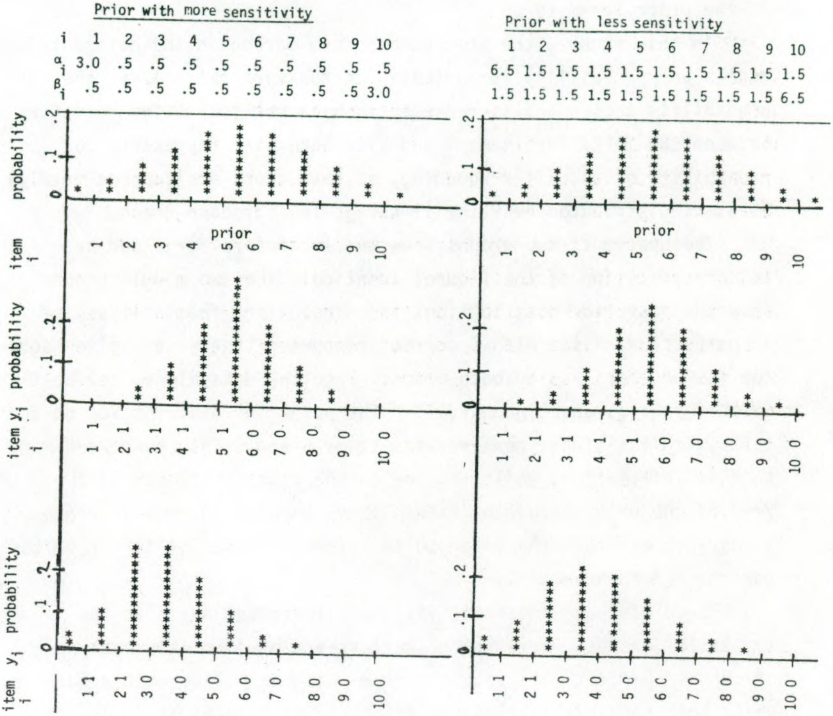
Choice of a prior distribution

With attention restricted to joint prior densities for $\{\pi_i\}$ having the form (4), the important question of selecting a prior is reduced to choosing values for the α_i and β_i (henceforth simply referred to as the prior parameters). Given the multidimensional character of (4) and the secondary interest in the π_i, it seems reasonable to concentrate on the consequences a given set of prior parameter values have for the prior distribution of the ability parameter θ. Except for the special cases mentioned in the previous section, these consequences are probably best explored with a trial-and-error procedure, using an interactive computer program such as ABILITY, mentioned earlier. Through successive adjustments of the values of the prior parameters, almost any desired shape for the prior of θ may be obtained.

As suggested by the final result of the previous section, it is often (perhaps always) possible to produce - at least roughly - a given prior distribution for θ from a whole range of priors for $\{\pi_i\}$. Far from being a problem, this introduces an important degree of flexibility in the analysis of responses to test items. Among the prior densities for $\{\pi_i\}$ yielding a given prior for θ, those with larger values of the prior parameters may be associated with greater amounts of prior information regarding the distribution of θ. The result is that the corresponding posteriors for θ will show less change from the common prior, i.e. less sensitivity to the information contained in a given set of responses.

This is illustrated in Fig. 1 for a case with 10 items and two priors

Figure 1. Two prior distributions for θ with the same form but having different sensitivities applied to two data sets.



for $\{\pi_i\}$ which produce (approximately) the same prior for θ. The form chosen is symmetric and unimodal, but has greater variance than the binomial (10, .5) which would have been obtained with equal values of the prior parameters.

In this figure, the item numbers are marked on the horizontal scales and probability is indicated on the vertical scales. The probability that $\theta = i$ is represented by a vertical column of stars between the marks for items i and i+1. Each star represents .02 probability units, after rounding, so that there are (approximately) 50 stars distributed over the 11 categories for each graph.

The uppermost two graphs show the two priors for θ and are (to the precision of the figure) identical. The two middle graphs show the posterior distributions for θ resulting from analyses of a "perfect" data set with 5 correct responses (the y_i are given above the item numbers) using both priors. From the data alone, the best guess for θ is, unambiguously, 5 . The posterior distribution on the left, with its higher peak between items 5 and 6, has responded more to this information, while the one on the right is closer to the form of the prior. From the labels above the two columns of graphs, it may be seen that the prior on the right is based on larger values for the prior parameters.

The difference in sensitivity is illustrated again in the two bottom graphs, where the posteriors are based on the analyses of another perfect data set, this time with only 2 correct responses. While both posteriors represent a compromise between prior and data, this compromise is clearly more in favor of the data in the left-hand graph. The posterior on the right retains more of both the breadth and the centering of the prior. To summarize, larger values of the prior parameters represent the fact that more is known before the test is taken. Thus, relatively less is learned from the test results and the posterior looks more like the prior.

What about the actual choice of a prior distribution for θ? A first point is that this must go hand in hand with the choice of potential items for the final test. Presumably, it would be wise -

if possible - to choose items which give a good coverage of the likely range of ability in the group of interest. This would imply a broad distribution for θ, tailing off for values close to 0 and $\underline{k}$. Of course, the chosen items should also identify ability levels of interest, since θ is only defined relative to their difficulties. Finally, there should be enough items to allow substantial revision of the prior based on test results. (If this cannot occur, why bother to give the test?)

Once items have been chosen and tested to the point where they can be ranked with confidence, and something has been learned about the abilities of potential test-takers, it may be reasonable to proceed by drawing a rough sketch of the desired prior for θ and then finding prior parameter values which produce this shape via trial and error, as mentioned earlier. A tentative choice should be tried out on a number of data sets (again using a program such as ABILITY) to observe the degree of sensitivity of the prior (as in Fig.1). If, for instance, the prior appears to be too <u>in</u>sensitive to data, one should try to produce the desired shape with smaller values of the prior parameters. This new choice should also be checked and the process continued until the results are judged satisfactory.

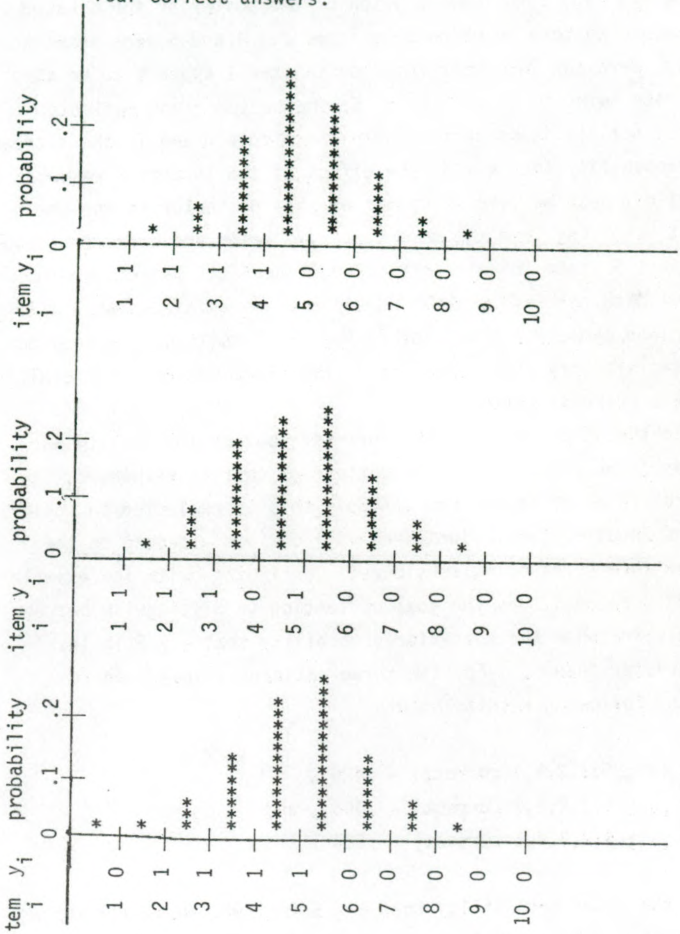
The above procedure may appear to be overly time consuming. This objection, however, cannot be maintained when the choice of prior is viewed as one step in the sequence of item construction, tryout, revision, administration, and final analysis. For any test which has been carefully prepared, it is surely worth a few hours of thought, discussion, and computer work to adequately assess the prior information available about the ability being measured. As with the rest of test development, choosing a prior for θ has a basically subjective character. The subjective choices should, however, be at least broadly defensible, with regard to both their origins and their consequences. Indeed, one of the advantages of adopting a formal Bayesian framework is that it makes subjective decisions and their implications explicit and, thus, open to examination and discussion. Although care should be exercised in the selection of a prior, the precise choice does not appear to be too critical with regard to posterior inferences about θ. For example, the corresponding prior parameters of the two priors in Fig.1 differ by factors of 2 to 3. The resulting pairs of posteriors for θ, while different in ways discussed above, still bear a strong resemblance within pairs. Based on the data set with two correct responses, for instance, an approximate 90% interval for θ ranges from 1 to 5 using the more sensitive prior and from 1 to 6 using the less sensitive one. The practical requirement that any chosen prior be substantially modifiable by the data guarantees that substantial robustness to the details of prior specification will be typical.

Posterior inferences and response pattern

One question of considerable practical interest is the sensitivity of the posterior distribution of θ to various response patterns, all having the same number of correct responses. Within the Rasch model, for instance, the number of items correct is a sufficient statistic for making inferences about an individual's ability, so the response pattern becomes ancillary information.That this is not so for the present model may be illustrated by means of an example.

Returning to the more sensitive prior illustrated in Fig. 1 for a 10 item test, three response patterns will be analyzed, each having 4 correct responses. Fig. 2 (employing a format similar to Fig. 1) shows, first, the posterior distribution of ability for an individual responding correctly to the 4 easiest items and incorrectly to the remaining 6. The posterior mode is at $\theta = 4$, reflecting the data, and the distribution is positively skewed, reflecting the prior information. The middle graph illustrates the posterior for an individual responding correctly only to the first three items and to item 5. From the data alone, values of 3 and 5 for θ seem most likely, with 4 less likely. Compared with the top graph, this

Figure 2. Posterior distribution for θ based on the more sensitive prior of Fig. 1 and three response patterns, each with four correct answers.



is reflected in the posterior: the probability that $\theta = 4$ decreases to the extent that 4 is no longer the modal value. (The fact that 5, rather than 3, assumes this role is a result of the prior information.)

Finally, Fig. 2 includes a graph of the posterior for θ based on a response pattern in which only items 2,3 4, and 5 were answered correctly. Here the incorrect response to item 1 appears to be atypical, and the data point to an ability of 5. The bottom graph reflects this. Most notable in comparison with the middle graph is the decrease in the probability that $\theta = 3$. The effect of the incorrect response to item 1 may best be seen by comparing this posterior to the one in Fig. 1 resulting from the more sensitive prior and correct responses to the first 5 items (middle left). The "imperfect" pattern produces a less definite peak and a clear negative skewness as compared with the sharp and symmetric posterior in Fig. 1. Intuitively, this seems to be a satisfactory representation of the extra uncertainty resulting from such a response pattern.

While the shape of the posterior distribution for ability has been shown to be sensitive to the pattern of correct responses, it is still of interest to see how strongly this is reflected in posterior inferences about θ. Conclusions here will obviously depend on the particular inferences being considered. Continuing with the example of Fig. 2, suppose it was the goal of testing to distinguish between individuals for whom the posterior probability that $\theta \geq 5$ is less than or greater than .5 . For the three patterns illustrated in Fig. 2, the following results hold:

$$\text{Prob } (\theta \geq 5 | 1,2,3,4 \text{ correct}) = .426 \ ,$$
$$\text{Prob } (\theta \geq 5 | 1,2,3,5 \text{ correct}) = .466 \ , \text{ and}$$
$$\text{Prob } (\theta \geq 5 | 2,3,4,5 \text{ correct}) = .524 \ .$$

Note that the prior probability that $\theta \geq 5$ is .592. Thus, all three response patterns have resulted in a reduction of this value. The first two have reduced it sufficiently to change the individual's classification, while the third maintains the prior choice.

This combination of data and inference has illustrated some sensitivity to response pattern. Now suppose the goal of testing is to obtain an interval for θ which contains the true value with a posterior probability of approximately .95. For the "perfect" response pattern, the "shortest" such interval ranges from 2 to 7 (inclusive). The variation in the probability content of this interval for the three posteriors of Fig.2 is as follows:

$$\text{Prob } (2 \leq \theta \leq 7 \mid 1,2,3,4 \text{ correct}) = .958,$$
$$\text{Prob } (2 \leq \theta \leq 7 \mid 1,2,3,5 \text{ correct}) = .951, \text{ and}$$
$$\text{Prob } (2 \leq \theta \leq 7 \mid 2,3,4,5 \text{ correct}) = .941.$$

While systematic, it is doubtful that this degree of variation could have much effect on the interpretation of the interval estimate for θ.

While only one example, the above gives some feeling for the range of results that may be expected. Sensitivity to response pattern is, in general, an increasing function of the sensitivity of the prior, the number of items, and the extremeness of the pattern. The number of correct responses remains, however, the most important single piece of information associated with a given pattern.

The results of this section have a second interpretation, namely as a study of the robustness of posterior inferences to violation of the ordering assumption (1). Thus, different patterns with the same number correct may be viewed as the same set of responses after ordering the items in different ways. Considered this way, it appears that only relatively extreme violations of (1) will have much effect on inferences (how much depending on the data and the sort of inference being made). There is, of course, also the issue of misinterpretation of θ itself if the items have been incorrectly ordered. This is a substantive matter which should be considered carefully in any testing situation where these methods might be applied. On substantive grounds, it may be safer to group together as equivalent items for which the ordering is unclear, rather than to risk working with an incorrect order.

References

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., and Brunk, H.D. Statistical inference under order restrictions. London: John Wiley & Sons Ltd., 1972.

Lewis, C. The countback method for analyzing sensitivity data. Research Bulletin 70-30. Princeton, N.J.: Educational Testing Service, 1970.

Lewis, C., Wang, M., and Novick, M.R. Marginal distributions for the estimations of proportions in m groups, Psychometrika,1975, 40, 63-75.

Mokken, R.J. A theory and procedure of scale analysis with applications in political research. The Hague: Mouton, 1971.

Stewart, L. Multiparameter univariate Bayesian analysis, Journal of the American Statistical Association, 1979, 74, 684-693.