An Empirical Bayesian Approach to Item Banking

Wim J. van der Linden and Theo J. H. M. Eggen
University of Twente
Enschede, The Netherlands

A procedure for the sequential optimization of the calibration of an item bank is given. The procedure is based on an empirical Bayesian approach to a reformulation of the Rasch model as a model for paired comparisons between the difficulties of test items in which ties are allowed to occur. First, it is shown how a paired-comparisons design deals with the usual incompleteness of calibration data and how the item parameters can be estimated using this design. Next, the procedure for a sequential optimization of the item parameter estimators is given, both for individuals responding to pairs of items and for item and examinee groups of any size. The paper concludes with a discussion of the choice of the first priors in the procedure and the problems involved in its generalization to other item response models.

An innovative idea produced by modern test theory is the notion of a calibrated item bank. A calibrated item bank is a large collection of test items, all measuring the same trait or domain of knowledge, that is stored in a computer together with empirical estimates of its item parameters. The item parameters are defined by an appropriate response model fitting the responses of the examinees to the items in the bank. The use of a calibrated item bank has two important advantages over standardized tests. The first advantage is the introduction of flexibility in the practice of testing in education and psychology. Using a calibrated item bank, test scores obtained from any selection of items from the bank will all measure the trait on the same scale. This allows test construction solely on the basis of practical considerations. The second advantage of item banking is efficient use of response data. Any new set of data, even the responses of a single person to only a few items, can be fed back into the computer for a periodic update of the item parameter estimates.

Flexibility in item selection is most desirable in individualized instruction systems such as computer-aided instruction (CAI). In such systems, individual students' achievements are monitored by testing them regularly using short tests. The fact that the test scores are used for instructional decision-making entails the necessity of a high precision of measurement on a scale that is the same for all items.

From an optimization point of view, two different stages in item banking can be considered. The first is the *calibration* stage, in which response data are collected in order to estimate the item parameters accurately (and to assess the fit of the items to the response model). Two obvious strategies in this stage are (1) to collect the minimal number of item responses needed to guarantee a certain level of accuracy for the estimators, or (2) to maximize the accuracy for a fixed number of responses. These two strategies can be implemented by manipulating the sampling design governing the

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 4, December 1986, pp. 345–354

0146-6216/86/040345-10$1.75

collection of response data (see below). The second stage in item banking is the *measurement* stage. In this stage the items have already been calibrated and their parameters can, up to a tolerable inaccuracy, be considered as known. Now the optimization problem is to construct tests from the bank measuring the expected distribution of the persons' trait levels with prescribed precision. Procedures for this optimization problem can be derived from linear programming (Boekkooi-Timminga, in press; Theunissen, 1985; van der Linden & Boekkooi-Timminga, 1986).

The present paper deals with optimization of the item calibration stage. Typically, the design of a calibration sample has two characteristics: it is (1) structurally incomplete and (2) sequential by nature. The former property is a consequence of the fact that for item banks exceeding the size of a conventional test, it is physically impossible to administer all items to each person in the sample. Hence, when designing the calibration samples, the item bank constructor must select the item-person combinations for which responses will be collected. The latter indicates that item responses may be collected in more than one stage, notably when the first sample of persons yields some parameter estimates for which the estimated standard error of estimation is considered too large and additional sampling is required.

It is the purpose of this paper to introduce a procedure that optimizes the calibration of an item bank fully sequentially and can easily be implemented on a microcomputer. The procedure can be applied at the level of pairs of item responses; at this level, it automatically selects each next pair of items to be administered while guaranteeing any predetermined level of accuracy for the item parameter estimates with a minimum number of responses. An obvious area for this level of application is item banking in CAI systems where, for individual students connected to the system, the responses to previous items can be used to select the next items to be administered. However, the procedure can also be applied to any number of items and examinees (see below). Basically, the procedure consists of an empirical Bayesian approach to a reformulation of the Rasch model as a model for paired comparisons with ties. Because this reformulation is a unique property of the Rasch model, the procedure cannot be used without modification for the other item response models in use for item banking.

A Model for Paired Comparisons with Ties

In item response theory, the usual level of modeling is the response of a single person to a single item. For instance, in the Rasch (1960) model, the probability of a person p with ability α_p responding correctly to a dichotomous item i with difficulty δ_i is formulated as

$$P(X_{pi} = 1|\alpha_p,\delta_i) = \alpha_p(\alpha_p+\delta_i)^{-1} \quad , \tag{1}$$

where $\alpha_p \in (0,+\infty)$, $\delta_i \in (0,+\infty)$, and X_{pi} is the random variable indicating whether the response is correct ($X_{pi} = 1$) or incorrect ($X_{pi} = 0$). A different perspective, however, is to consider a person's response to a pair of items, i and j. If the values of the item parameters are unknown, the outcome of this experiment can be conceived of as the outcome of a paired-comparisons experiment in which the person "judges" the relative difficulties of the test items. Three different outcomes can be distinguished:

$\{X_{pi} > X_{pj}\}$: item i is correct and j is not;
$\{X_{pi} < X_{pj}\}$: item j is correct and i is not;
$\{X_{pi} = X_{pj}\}$: both items i and j are correct or incorrect.

The first outcome can be interpreted as a comparison showing item j likely to be more difficult than i, whereas the second outcome indicates the reverse. The last outcome represents a compound event in which the two elementary events $\{X_{pi} = 0, X_{pj} = 0\}$ and $\{X_{pi} = 1, X_{pj} = 1\}$ have been taken together because, from the point of view of a comparison between the difficulties of the items, either indicates that i and j are likely to be equally difficult. In the literature on paired comparisons, such outcomes are denoted as *ties* (Bradley, 1976; Davidson, 1970; Glenn & David, 1960; Gridgeman, 1959; Kousgaard, 1976; Rao & Kupper, 1967).

From the model in Equation 1, completed with the property of local independence, the probabilities of the three possible outcomes follow straightforwardly:

$$P(X_{pi} > X_{pj}) = \alpha_p\delta_i[(\alpha_p+\delta_i)(\alpha_p+\delta_j)]^{-1} \quad , \qquad (2)$$

$$P(X_{pi} < X_{pj}) = \alpha_p\delta_i[(\alpha_p+\delta_i)(\alpha_p+\delta_j)]^{-1} \quad , \qquad (3)$$

$$P(X_{pi} = X_{pj}) = (\alpha_p^2+\delta_i\delta_j)[(\alpha_p+\delta_i)(\alpha_p+\delta_j)]^{-1} \quad . \qquad (4)$$

Van der Linden and Eggen (1986) made a thorough study of the model specified by these equations. An important feature of the model not previously noted in the literature is that if all possible comparisons between a vector of n responses are considered, some of them are dependent. For example, if a person p has responded to three items i, j, and k and the outcome $\{X_{pi} > X_{pj}\}$ has occurred, then the outcome $\{X_{pi} < X_{pk}\}$ is impossible. In van der Linden and Eggen (1986) it is shown that, for a person with responses to n items, exactly $\min\{t, n-t\}$ of the $n(n-1)/2$ possible comparisons are independent, where $t = \sum_{i=1}^{n} X_{pi}$ is the number-correct score of the person. Algorithms to remove dependent outcomes from the data are given in the same paper. The relationship of Equations 2–4 to other models for paired comparisons with ties is examined in Eggen and van der Linden (1987). These authors indicate that one of the well-known models for ties in the literature, namely the one proposed by Davidson (1970), can be considered a simple reparameterization of the Rasch model for paired comparisons in Equations 2–4.

The outcome $\{X_{pi} = X_{pj}\}$ does not contain any information on the relative difficulties of the items. Disregarding this outcome, which is the same as conditioning the model in Equations 2–4 on the event of a non-tie $\{X_{pi} \neq X_{pj}\}$, yields the Bradley-Terry (1952) model:

$$P(X_{pi} > X_{pj} | X_{pi} \neq X_{pj}) = \theta_{ij} = \delta_j(\delta_i+\delta_j)^{-1} \quad . \qquad (5)$$

Thus, for persons responding to a series of test items, if the dependent outcomes and the ties are removed from the data, a model extensively studied in the literature is available to analyze the remaining comparisons and estimate the item parameters.

It should be noted that the parameter θ_{ij} in this model is a monotonic transformation of $\delta_i\delta_j^{-1}$ and therefore can be considered as a measure of the difficulty of item i relative to j.

An important property of Equation 5 is that, unlike the model in Equations 2–4, it does not contain the person parameter α_p. Hence, in order to estimate the parameters θ_{ij}, δ_i, or δ_j the responses from any sample of examinees may be pooled. The fact that Equation 5 can be derived from the Rasch model was already noted in Rasch (1960). Fischer (1974, 1981) has used this derivation to establish several properties of the Rasch model. Choppin (1968), in an early but not widely noted paper, pointed out the advantage of using Equation 5 instead of Equation 1 for item banking purposes (but ignored the aforementioned problem of dependency among the comparisons); the interest in the paired-comparisons design in the present paper was mainly motivated by this publication.

Parameter Estimation

From Equation 5 it follows that for a pair of items, the number of outcomes $\{X_{pi} > X_{pj}\}$ for a series of independent trials can be described as a Bernoulli experiment with success parameter θ_{ij}. Therefore, if n_{ij} $(i, j = 1, \ldots, n)$ is the number of non-ties and A_{ij} the number of times $\{X_{pi} > X_{pj}\}$ is observed, then

$$A_{ij} \sim \text{Bin}(n_{ij}, \theta_{ij}) \quad . \qquad (6)$$

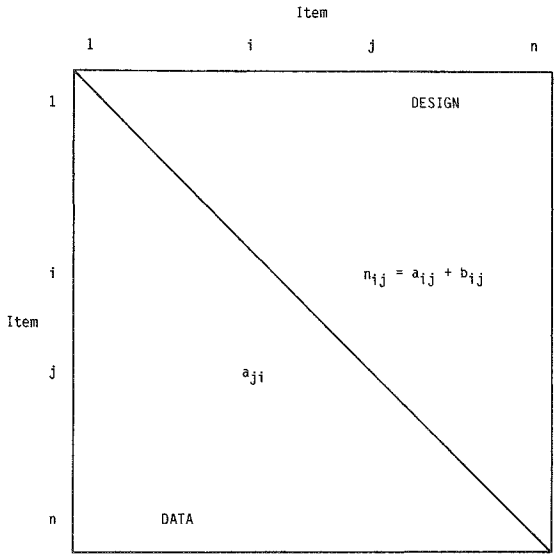
Hence, for a set I of independent comparisons between the n items, a product-binomial model with the following likelihood function holds:

$$L(\delta_1, \ldots, \delta_n; a_{ij}, n_{ij}, i < j) = \prod_I [\delta_j(\delta_i+\delta_j)^{-1}]^{a_{ij}}[\delta_i(\delta_i+\delta_j)^{-1}]^{a_{ji}} \quad . \qquad (7)$$

The following algorithm for solving the likelihood equations from Equation 7 for $(\delta_1, \ldots, \delta_n)$ has been introduced independently by Zermelo (1929) and Ford (1957) in the paired comparisons literature:

$$\delta_i^{(k+1)} = \sum_{j=1}^{n} a_{ij}\left[\sum_{j=1}^{n} n_{ij}(\delta_i^{(k)}+\delta_j^{(k)})^{-1}\right]^{-1}$$

Figure 1
A Paired-Comparisons Design for Item Banking



$$\delta_j^{(k+1)} = \delta_j^{(k)} \quad (j \neq i) \quad , \tag{8}$$

where $i = 1, \ldots, n$ cyclically.

In Equation 8 the superscript k indicates the iteration step and, as before, a_{ij} is a realization of A_{ij} and $n_{ij} = a_{ij} + a_{ji}$. The data needed to solve the likelihood equations are given schematically in Figure 1.

Existence and Uniqueness of Estimates

Zermelo (1929) and Ford (1957) gave a necessary and sufficient condition for the existence and uniqueness of a solution to the likelihood equations defined by Equation 7. The condition reads: In every possible partition of the set of items into two non-empty subsets, for some item i in the first set and some item j in the second one, the outcome $\{X_{pi} > X_{pj}\}$ has occurred for at least one value of p.

The theorem has an instructive interpretation in graph theory: If the items are mapped one-to-one on the n points of a directed graph with an arrow from i to j if and only if $\{X_{pi} > X_{pj}\}$ has occurred at least once, the condition amounts to the requirement that the resulting graph be strongly connected. The same condition has been established by Fischer (1981) for conditional maximum likelihood estimation of the item parameters directly from Equation 1. The condition is very mild and is commonly met by datasets occurring in practice.

A practical consequence of the Zermelo-Ford condition is that it is *not* necessary for the design matrix in Figure 1 to be balanced (i.e., $n_{ij} = n$ for every i,j). Therefore, the reformulation of the Rasch model given above deals with incomplete data in the original person × item matrix in an elegant fashion. When analyzing the data at the level of pairs of items, incomplete data simply return as unequal numbers of comparisons per pair, and maximum likelihood estimation is still possible as long as the Zermelo-Ford condition is met. Experiments with an unequal number of repetitions have a tradition in the paired-comparisons literature dating back to Dijkstra (1960).

An Empirical Bayesian Approach

As noted earlier, calibration designs are typified by structural incompleteness and the possibility of sequential optimization. The analysis above shows that the paired-comparisons design in Figure 1 has the simple provision of unequal numbers of repetitions n_{ij} to deal with structurally incomplete data. Since the model in Equation 5 no longer contains the ability parameter α_p, data from different examinees may be pooled together. As a consequence, data from any examinee, even when responding to only two items from the bank, can be stored in the format of Figure 1 and used in the estimation of the item parameters. The fact that this permits an efficient use of response data in item banking was already noted in Choppin (1968). From the point of view of data storage, Figure 1 also shows an efficient format. For n items, irrespective of the number of examinees, only an $n \times n$ matrix, containing the counts of the numbers of times items i and j were compared and i turned out to be easier, must be stored. Item calibration data are often collected in a sequential fashion. A paired-comparisons design would be fully appropriate for item banking purposes if sequential optimization

of the calibration procedure were possible. The following presents such a procedure based on an empirical Bayesian approach to the paired-comparisons model in Equation 5.

As A_{ij} is binomially distributed with

$$\theta_{ij} = \delta_j(\delta_i+\delta_j)^{-1} \quad (9)$$

as success parameter, the natural choice of a prior distribution for this parameter in the Bayesian framework is the beta distribution. A variable θ has a beta distribution if its density is given by

$$f(\theta) = C(a,b)\theta^{a-1}(1-\theta)^{b-1} \quad (10)$$
$$(0 < \theta < 1;\ a,b > 0) \quad ,$$

where $C(a,b) = \Gamma(a+b)[\Gamma(A)\Gamma(B)]^{-1}$ is a constant depending on the well-known gamma functions. Some properties of a beta-distributed variable useful in this context are

$$\mathrm{E}(\theta) = a(a+b)^{-1} \quad (11)$$

$$\mathrm{Mode}(\theta) = (a-1)(a+b-2)^{-1} \quad (12)$$

$$\mathrm{Var}(\theta) = ab(a+b+1)^{-1}(a+b)^{-2} \quad (13)$$

(Novick & Jackson, 1974, chap. 5). In general, the choice of the beta distribution as a prior for the binomial model is motivated by the following two properties:

1. The beta distribution is very flexible. Depending on the values for the parameters a and b, it can take many shapes in the interval [0,1]. Nevertheless, its density function contains only two parameters.
2. The beta distribution is the natural conjugate of the binomial model. This means that the family of beta distributions is closed under the application of Bayes' theorem and that the posterior distribution will always be beta again. In other words, the transition from a prior to a posterior distribution takes place only at the parametric level; if θ has $B(a,b)$ as a prior and the data consist of n observations containing k successes, then θ has $B(a+k,\ b+n-k)$ as a posterior. The fact that the parameters a and b in the prior have the same effect on the posterior as the statistics k and n, respectively, gives the information in the prior an instructive interpretation: It is equal to the information about θ in a hypothetical series of n Bernoulli experiments with k successes.

In an empirical Bayesian approach (Robbins, 1956, 1964), the Bayesian model is applied sequentially and the prior distribution for the model parameter(s) at step $k+1$ is based on the observations at steps 1, ..., k. The process is repeated until the posterior distribution meets some stopping rule (e.g., allows making a decision with a certain expected risk). In the following section an empirical Bayesian procedure optimizing the calibration of an item bank is given. For a previously determined level of uncertainty about the item parameters, the procedure automatically minimizes the number of item administrations. The procedure operates at the level of pairs of items. Roughly, it starts with a first prior for all success parameters θ_{ij} and computes the posterior each time a paired comparison is made. The posterior is used as the prior for the next observation. At each step the pair of items with the largest uncertainty about θ_{ij} is selected for administration. The process is repeated until for all pairs the posterior uncertainty is below the maximum specified in a stopping rule.

Optimal Calibration Procedure

As in the preceding sections, it is assumed that the Rasch model in Equation 1 holds for $i = 1, \ldots, n$, so that for every pair of items $i < j = 1, \ldots, n$, the comparisons not resulting in a tie can be conceived of as the outcome of a Bernoulli experiment with probability of success $\theta_{ij} = \delta_j(\delta_i+\delta_j)^{-1}$. For every θ_{ij} $(i < j = 1, \ldots, n)$, as a prior the beta distribution in Equation 10 with parameters a_{ij} and b_{ij} is assumed. The variance

$$a_{ij}b_{ij}(a_{ij}+b_{ij}+1)^{-1}(a_{ij}+b_{ij})^{-2} \quad (14)$$

in the distribution of θ_{ij} in Equation 13 is adopted as a measure for the uncertainty about the value of the parameter for the relative item difficulty θ_{ij}. Now, administering the pair (i,j) once at step k, the prior $B(a_{ij},b_{ij})$ for $\theta_{ij,k}$ will be replaced by one of the following posteriors:

$$\{X_{pi} > X_{pj}\}\colon\ \theta_{ij,k+1} \sim B(a_{ij}+1,\ b_{ij}) \quad , \quad (15)$$

$$\{X_{pi} < X_{pj}\}\colon\ \theta_{ij,k+1} \sim B(a_{ij},b_{ij}+1) \quad , \quad (16)$$

$$\{X_{pi} = X_{pj}\}\colon\ \theta_{ij,k+1} \sim B(a_{ij},b_{ij}) \quad . \quad (17)$$

Thus the outcome of a comparison between i and j is equivalent to the observation of which parameter of the prior distribution must be updated by a value of 1. The posterior uncertainty about θ_{ij} (Equation 14) is obtained by the same simple update of the parameters in Equations 15–17.

An optimal calibration procedure based on an empirical Bayesian approach is as follows:

1. Specify a first prior for the relative item difficulties θ_{ij} $(i < j = 1, \ldots, n)$ in the bank, that is, select starting values for the parameters a_{ij} and b_{ij}.
2. Administer the pair of items with maximum variance (Equation 13):

$$\max_{(i,j)} a_{ij}b_{ij}(a_{ij}+b_{ij}+1)^{-1}(a_{ij}+b_{ij})^{-2} \quad . \qquad (18)$$

   If this pair is not unique, select one of the pairs satisfying Expression 18 at random.
3. Update the prior distribution of θ_{ij} according to Equations 15–17.
4. Repeat steps 2 and 3 until the maximum variance in Expression 18 is below a prespecified level.

This procedure can easily be implemented on a computer. Only a few lines of computer programming are needed and the data to be stored have already been displayed in the matrix in Figure 1.

The parameters of the beta distribution are simply a_{ij}, the number of times i turns out to be easier than j, and $b_{ij} = n_{ij} - a_{ij}$, the remaining number of times i and j are compared in this matrix. The choice of the first priors for the model amounts to an initialization of the entries in the matrix, and the priors are simply updated by augmenting the entry a_{ij} by 1 if i was easier than j, augmenting b_{ij} by 1 if i was more difficult than j, and augmenting neither if a tie occurred. The variance used in Expression 18 can be computed directly from these entries. The determination of which pair of items must be administered at each step can be simplified by storing the data in an array ordered by the size of the variances, such that at each step the next pair to be administered is (a random selection of one of) the final element(s) in the array. The computations are not intricate because no maximum likelihood estimation of the individual item parameters δ_i is needed in each step.

The actual estimation of the item parameters is separated from the procedure optimizing its sampling design, the latter being based on the uncertainty about the relative item difficulties rather than their absolute values. However, at the same time the data needed for conditional maximum likelihood estimation are kept in an appropriate format, and each time actual estimates of the item parameters are needed the application of the Zermelo-Ford algorithm in Equation 8 to the data matrix in Figure 1 is straightforward.

As already mentioned, a natural environment to apply the above procedure at the level of pairs of items is item banking in CAI. If an item bank in such systems has to be calibrated, the sampling of item responses can be adapted sequentially to the posterior uncertainties about the item parameters. If the calibration is finished, the direction of adaptation can be reversed to tailor the selection of items to the individual students' abilities.

Preposterior Analysis

A usual principle in sequential Bayesian decision-making is preposterior analysis. For the present problem, preposterior analysis would mean that the selection of the pair of items at each step k is based not on the prior variances of θ_{ij} $(i < j = 1, \ldots, n)$, but on the expected reductions of the variances due to the outcome of the next comparison between the items. A likely criterion would be to choose the pair of items for which this reduction is maximal:

$$\max_{(i,j)} \phi(a_{ij},b_{ij}) \equiv \max_{(i,j)}\{\mathrm{E}[\mathrm{Var}(\theta_{ij,k}) - \mathrm{Var}(\theta_{ij,k+1})]\} \quad . \qquad (19)$$

It is easily shown that this expected reduction is always positive, which is in agreement with the present authors' intuition that additional observations should reduce the uncertainty about parameter values. Although preposterior analysis is in the Bayesian tradition, it offers no advantage over Expression 18 in this case, because in the beta-binomial model equal prior variances imply equal expected reductions in variance.

Suppose the prior variances of $\theta_{i_1j_1}$ and $\theta_{i_2j_2}$ are equal:

$$\mathrm{Var}(\theta_{i_1j_1}) = \mathrm{Var}(\theta_{i_2j_2}) \quad . \tag{20}$$

Then,

$$a_{i_1j_1}b_{i_1j_1}(a_{i_1j_1}+b_{i_1j_1}+1)^{-1}(a_{i_1j_1}+b_{i_1j_1})^{-2} = a_{i_2j_2}b_{i_2j_2}(a_{i_2j_2}+b_{i_2j_2}+1)^{-1}(a_{i_2j_2}+b_{i_2j_2})^{-2}. \tag{21}$$

This holds only if

$$a_{i_1j_1}b_{i_1j_1} = a_{i_2j_2}b_{i_2j_2} \tag{22}$$

and

$$a_{i_1j_1}+b_{i_1j_1} = a_{i_2j_2}+b_{i_2j_2} \quad , \tag{23}$$

that is, if

$$a_{i_1j_1} = a_{i_2j_2} \tag{24}$$

and

$$b_{i_1j_1} = b_{i_2j_2} \quad , \tag{25}$$

or if

$$a_{i_1j_1} = b_{i_2j_2} \tag{26}$$

and

$$a_{i_2j_2} = b_{i_1j_1} \quad . \tag{27}$$

Thus, Equation 20 implies Equations 24–25 or 26–27. Therefore, because

$$\phi(a,b) = \phi(b,a) \quad , \tag{28}$$

Equation 20 also implies equal reductions of the variances in Equation 19.

Because preposterior analysis in the present case does not introduce a more effective sampling of the item responses and Equation 19 is computationally more involved than Expression 18, the latter is the favored criterion.

Generalization to Tests and Groups of Any Size

Thus far, at each step in the optimization procedure, only responses of a single examinee to a pair of items have been considered. Generalizations to tests and groups of examinees of any size are, however, straightforward.

The generalization for a single examinee to a test of any length is as follows: (1) Select a pair of items, say i_0j_0, according to the above procedure; (2) suppress i_0 and j_0 in the array with prior variances and select the next pair; and (3) continue until enough items have been selected to fill the test.

A new aspect in this generalization is the independence of the outcomes of the comparisons necessary for using the likelihood equations from Equation 7. In principle, two strategies are available to satisfy the condition of independence. The first is to partition the test into pairs before the administration and to consider only the outcomes of the comparisons within these pairs. In this way independence is obtained by design. The other strategy is to use the fact that the number of independent non-ties for a single examinee and a test of n items is equal to $\min\{t,n-t\}$, and to select this number at random from all non-ties after the administration of the test under the restriction that for each examinee an item figures only once in a pair. This strategy has been followed in van der Linden and Eggen (1986) and is recommended because it guarantees the maximum amount of data from the test.

The generalization to a test of any length for a group of examinees of any size poses no further problems. The same test is administered to all examinees; the data matrix is filled less efficiently than when the test for each examinee is selected in turn while taking the responses of the previous examinees into account, but this is the tradeoff for leaving the domain of sequential optimization.

Specifying the First Priors

Four different ways of specifying the $n(n-1)/2$ first priors in the procedure are possible. The first two are empirical methods; the other two are subjective. Not all of these methods are feasible with larger item banks, however.

Empirical Information From Other Experiments

If the items were administered earlier in an experiment, the data from this experiment can be edited into the format of the matrix with paired comparisons in Figure 1; the entries of this matrix are the first values of the parameters of the priors in the procedure. Again in this approach, in order to guarantee independent data either the pairs must be chosen before inspecting the data or $\min\{t,n-t\}$ comparisons must be sampled at random.

It should be noted that this method can be used even if not all of the items have been previously administered. In that case, the priors for the other items must be specified using one of the methods below. Again, this is an example of the efficient pooling of data from different sources possible because the model in Equation 5 is free from the person parameters α_p.

Empirical Information From the Experiment Itself

An obvious variant of the above method is to begin the calibration experiment without any prior and collect some data to fill the matrix in Figure 1 with starting values. For instance, before starting the empirical Bayesian procedure all item pairs could be selected a few times in a random order to obtain empirical values for the parameters a_{ij} $(i < j = 1, \ldots, n)$. This approach might appeal to researchers who do not favor the use of subjective priors.

Noninformative Priors

The beta distribution with parameter values $a = b = 1$ is the uniform distribution on the interval [0,1]. The attractive aspect of the choice of this distribution as the first prior for all item pairs is the fact that it does not favor any of the items; each item achieves one "success" and one "failure" against every other item. Moreover, because the number of hypothetical comparisons $(n = 2)$ is small compared to the amount of data usually collected in an item calibration study, the effect of the subjectivity of this choice on the eventual estimates can be neglected. A final advantage of this method is that from the outset, the data matrix meets the Zermelo-Ford condition for the existence and uniqueness of maximum likelihood estimates; an adjacency matrix with positive (integer) values for its elements is strongly connected (Harary, Norman, & Cartwright, 1965, theorem 5.14). As the property of connectedness is maintained when data are added to the matrix, unique maximum likelihood estimates always exist with uniform priors.

Subjective Priors

The final possibility is to have human judges specify subjective priors. An obvious procedure is to use a thought experiment based on the interpretation of the parameters a_{ij} in the beta distribution as the number of successes of item i against item j in a hypothetical series of n_{ij} Bernoulli trials. In the two steps of this experiment, the judge must first estimate the percentage of times i will be easier than j, $a_{ij}n_{ij}^{-1} \times 100\%$, and then must indicate the number of empirical trials, n_{ij}, to which his/her certainty is equivalent. The subjective estimates could be smoothed by calculating the maximum likelihood estimates of δ_i $(i = 1, \ldots, n)$ using the Zermelo-Ford algorithm and replacing the original subjective values for a_{ij} $(i < j = 1, \ldots, n)$ by their estimated expected values under the model. If desired, the residuals could be used to check the judges for inconsistencies in their judgments, as is usual in paired-comparison experiments.

A problem with this method of subjective priors is that it cannot be used for larger numbers of items, because the number of comparisons would soon grow forbiddingly large. However, the method can be used in combination with one of the above methods, for example, to fill the holes in the data matrix left after data from previous experiments were used to get empirical values for the other elements.

Discussion

In the procedure of optimal item bank calibration presented in this paper, the calculations necessary to decide which items to administer next were separated from the actual maximum likelihood estimation of the item parameters. Hence, a procedure demanding only a few simple calculations was possible. Nevertheless, the data are stored in a format for which an efficient algorithm for conditional maximum likelihood estimation is available. An advantage of this procedure for conditional maximum likelihood estimation from paired comparisons data is that, as opposed to regular conditional maximum likelihood estimation, there are no numerical problems due to the presence of elementary

symmetric functions (Gustafsson, 1980). Therefore, the procedure does not have the usual limitation (to 50 or 60 items) of the standard computer programs for the Rasch model, and outperforms these considerably in required computing time (Eggen & van der Linden, 1987; van der Linden & Eggen, 1986).

A model for paired comparisons between test items without any person parameter, such as the one in Equation 5, is only possible for item response models with a sufficient statistic for the person parameter. Among the logistic models with unknown item parameters, this is a unique property of the Rasch model. Nevertheless, it is debatable whether the same sequential optimization procedure is possible where, for example, a discriminating power parameter is a feature of the model. For instance, an apparent solution could make use of double data storage—in the paired comparisons format to decide which items to administer next and in the usual (incomplete) persons × items format for item parameter estimation. However, even if maximization of the information about the difficulty parameter were possible along these lines, it is still unknown whether this would imply simultaneous maximization of the information about the discrimination parameter. The point is that with more than one item parameter in the model, the optimization problem becomes an example of decision-making with multiple objectives. A possible solution to the problem of having more than one item parameter, each imposing a different requirement on the sampling design, adopted in van der Linden (1986) is to introduce a maximin criterion in the optimization procedure. Then it is possible to calculate an optimal sampling design for simultaneous administration of the test items to the examinees. The possibility of sequential procedures along these lines has yet to be investigated.

References

Boekkooi-Timminga, E. (in press). Simultaneous test construction by zero-one programming. *Methodika*.

Bradley, R. A. (1976). Science, statistics and paired comparisons. *Biometrics, 32,* 213–232.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika, 39,* 324–345.

Choppin, B. H. L. (1968). An item bank using sample-free calibration. *Nature, 219,* 870–872.

Davidson, R. R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association, 65,* 317–328.

Dijkstra, O. (1960). Rank analysis of incomplete block designs: A method of paired comparisons employing unequal repetition on pairs. *Biometrics, 16,* 176–188.

Eggen, T. J. H. M., & van der Linden, W. J. (1987). The use of models for paired comparisons with ties. In W. E. Saris & I. N. Gallhofer (Eds.), *Sociometric research. Volume 1: Data collection and scaling*. London: Macmillan.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* [Introduction to the theory of psychological tests]. Bern: Hans Huber.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika, 46,* 59–77.

Ford, L. R., Jr. (1957). Solution of a ranking problem for binary comparisons. *American Mathematical Monthly, 64,* 28–33.

Glenn, W. A., & David, H. A. (1960). Ties in paired comparisons experiments using a modified Thurstone-Mosteller method. *Biometrics, 16,* 86–109.

Gridgeman, N. T. (1959). Pair comparison, with and without ties. *Biometrics, 15,* 382–388.

Gustafsson, J. E. (1980). A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. *Educational and Psychological Measurement, 40,* 377–385.

Harary, F., Norman, R. Z., & Cartwright, D. (1965). *Structural models: An introduction to the theory of directed graphs*. New York: Wiley.

Kousgaard, N. (1976). Models for paired comparisons with ties. *Scandinavian Journal of Statistics, 3,* 1–14.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. New York: McGraw-Hill.

Rao, P. V., & Kupper, L. L. (1967). Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association, 62,* 194–204, corrigenda *63,* 1550.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1). Berkeley CA: University of California Press.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Annals of Mathematical Statistics, 35,* 1–20.

Theunissen, T. J. J. M. (1985). Binary programming and test design. *Psychometrika, 50,* 411–420.

van der Linden, W. J. (1986, June). *Optimal sampling design for parameter estimation in item response models*. Paper presented at the annual meeting of the Psychometric Society, Toronto.

van der Linden, W. J., & Boekkooi-Timminga, E. (1986). *Algorithmic test design with practical constraints*. Enschede, The Netherlands: University of Twente, Department of Education.

van der Linden, W. J., & Eggen, T. J. H. M. (1986). *The Rasch model as a model for paired comparisons with ties*. Enschede, The Netherlands: University of Twente, Department of Education.

Zermelo, E. (1929). Die Berechnung der Turnierergebnisse als ein Maximumproblem der Warscheinlichkeitsrechnung [The computation of tournament results as a maximization problem in probability calculus]. *Mathematische Zeitschrift, 29,* 436–460.

Authors' Addresses

Send requests for reprints or further information to Wim J. van der Linden, Universiteit Twente, Faculteit der Toegepaste Onderwijskunde, Postbus 217, 7500 AE Enschede, The Netherlands. Theo J. H. M. Eggen is now at CITO, Postbus 1034, 6801 MG Arnhem, The Netherlands.