

Equivalence Between Conditional and Mixture Approaches to the Rasch Model and Matched Case-Control Studies, With Applications

Kenneth M. RICE

Analyses of data using Rasch models, including the special case of matched case-control studies, are common applications of conditional likelihood in which the usual inferential procedures are applied only after conditioning on an approximately ancillary statistic. Another common approach to the analysis of Rasch models is to integrate the nuisance parameters over a mixing distribution, using the marginal likelihood obtained as the basis for inference. We show that the full conditional likelihood can always be obtained exactly via the marginal approach, given a particular choice of mixing distribution, and derive necessary and sufficient conditions for the two approaches to agree. Previous work has shown that with sufficient flexibility in the mixing distribution, the maxima of the marginal and conditional likelihoods will be equivalent under concordance criteria. Our argument requires no such criteria, and for any dataset guarantees the equivalence of the whole of the two likelihoods, not just their maxima. This substantially enhances the previous results and provides an alternative derivation for any existing conditional analysis. We give examples of mixing distributions that guarantee the agreement of the two approaches, and explore equivalence classes of such distributions, together with some of their attractive symmetry properties. Our argument also allows for the adaption and extension of analytic techniques already widely used with Rasch data, and in particular with matched case-control studies; potential applications of these advances are illustrated with several examples. These include new numerical algorithms for evaluating the conditional likelihood without directly specifying its computationally awkward functional form, inferences about complex functions of the parameters of interest obtained using existing Markov chain Monte Carlo methods, powerful measures of goodness of fit derived from likelihood contributions that are ignored by the conditional approach, and the justifiable addition of prior knowledge to existing conditional analyses.

KEY WORDS: Conditional likelihood; Full likelihood; Matched case-control study; Mixture model; Rasch model.

1. INTRODUCTION

This article presents new results on the standard techniques used in analysis of data from Rasch models, and the special case of Rasch models typically used with data from matched case-control studies. The article comprises two distinct parts, the required theory followed by the applications. The remainder of Section 1 formally describes the Rasch model and current methodology, that is, conditional and marginal likelihood methods. Section 2 gives the main theoretical results, describing conditions for the equivalence of the two current approaches. Section 3 shows that these conditions can be met and gives some attractive examples. With this theory alone, the arguments used in three of the later examples can be justified. Section 4 extends the theoretical results to give a measure of the information lost when using the conditional approach. Section 5 exemplifies this and explores other advances. Finally, Section 6 discusses Bayesian and other alternative derivations of this work, with comparison to some related approaches. Theoretical results requiring more than a few lines of justification are given in Appendixes.

The Rasch model (Rasch 1960) is a simple logistic model for independent binary responses, typically used to analyze data from questionnaires or tests. It is assumed that each question has an “easiness” parameter, α , and that each individual has an “ability” parameter, β . These parameters are combined linearly on the logistic scale, so that the probability of a correct answer is

$$\mathbb{P}(X_{ij} = 1) = \frac{\alpha_i \beta_j}{1 + \alpha_i \beta_j}, \quad 1 \leq i \leq I, \quad 1 \leq j \leq J, \quad (1)$$

where X_{ij} is the indicator of a correct answer to question i by individual j . The responses are assumed to be independent given the ability parameters. With these assumptions alone, the number of parameters grows with the size of the dataset, and maximum likelihood parameter estimates (typically of the α_i 's) are well known to be inconsistent.

To overcome this difficulty, two analytic techniques are commonly used. The first, “classical,” method uses a conditional argument. Note that $C_j = \sum_i X_{ij}$, the total number of questions answered correctly (or total score) by individual j , is a sufficient statistic for their ability parameter β_j and is also approximately ancillary for the α_i . Conditioning on this statistic, individual j contributes

$$L_{\text{cond}} = \prod_i \alpha_i^{x_{ij}} / \sum_{\mathbf{x}' : \mathbf{1}^T \cdot \mathbf{x}' = c_j} \prod_i \alpha_i^{x'_i} \quad (2)$$

to the conditional likelihood, where summation is taken over all possible responses \mathbf{x}' that have total score identical to that of individual j . The analysis then progresses using the conditional likelihood as one would a standard likelihood. The conditional maximum likelihood estimator (CMLE) is consistent and asymptotically normal under regularity conditions (Andersen 1970). Estimates from the conditional likelihood have also been shown to be efficient, and this method of estimation is “recommended” (Molenaar 1995, p. 51). The procedure is perhaps used most commonly in the guise of conditional logistic regression for matched case-control studies, where β are the strata-specific nuisance parameters and the α parameters represent within-strata odds ratios, possibly assumed to depend on some item-specific covariables through a log-linear link function. For large values of I , the contributions L_{cond} can be difficult to compute (Mehta, Patel, and Senchaudhuri 2000).

Kenneth M. Rice is Research Associate, MRC Biostatistics Unit, Cambridge CB2 2SR, U.K. (E-mail: kenneth.rice@mrc-bsu.cam.ac.uk). The author thanks David Spiegelhalter, Pat Altham, David Clayton, Stephen Duffy, Vern Farewell, Julian Higgins, Jim Lindsey, Richard Nixon, Simon Thompson, and Jon Wakefield for their useful comments on the manuscript. Referees of this and previous versions of the article also suggested several improvements.

Another popular method of analysis uses marginal likelihood, in which the nuisance parameters β are first integrated out of the likelihood and the resultant marginal (or integrated) likelihood is maximized to obtain the maximized marginal likelihood estimator (MMLE). Implicit in this approach is a choice of mixing distribution for the β_j , say G , also known as a random-effects distribution, interpretable as a prior or hyperprior if we take the equivalent Bayesian formulation of the model. There is no standard choice for the mixing distribution, with common alternatives including some form of normal distribution with hyperparameters governing the mean and variance or a nonparametric approach in which the moments of G are allowed to vary as much as possible.

Starting with the work of the work of Tjur (1982) and Kelderman (1984), many authors have considered relations between the two methods of analysis, leading to important knowledge about the structure of item response models (Cressie and Holland 1983; Rosenbaum 1984; De Leeuw and Verhelst 1986; Follman 1988). Most recently, Lindsay, Clogg, and Grego (1991) showed that with sufficient flexibility in G , the part of the marginal likelihood depending only on C_j will have a saturated fit, which leads to the exact agreement of the MMLE and CMLE for α . Such saturation is possible if and only if the various total scores satisfy concordance conditions, but these are shown to hold asymptotically with probability 1. The flexibility property is semiparametric, but Lindsay et al. (1991) showed that it can always be attained when one takes G to be a T -class latent-structure model with at most $(I + 1)/2$ points of support. Using profile-likelihood methods on the marginal likelihood, Lindsay et al. (1991) similarly established the entire conditional likelihood as a “profile mixture likelihood” for $I = 2$ only, and again this can be done only when the data meet specified concordance conditions.

In Section 2 we derive necessary and sufficient conditions on G such that the entire marginal likelihood is exactly equivalent to the conditional likelihood for any dataset, and prove the general existence of such G for any $I \geq 2$. Like the approach of Lindsay et al. (1991), our approach is semiparametric, because the analysis requires that G is restricted to be a member of a particular class of distributions, without (necessarily) specifying it exactly. The restrictions on G here are stronger than those of Lindsay et al. (1991), but our approach leads not only to equality of MMLE and CMLE without concordance conditions, but also to equivalence of confidence intervals or other quantities typically derived from the conditional likelihood.

The formulation in Section 2 of the conditional likelihood within the marginal likelihood framework also allows it to be incorporated into Bayesian analyses by additionally placing priors on α , the parameter of interest. Although this procedure has until now not been strictly justified, it does seem intuitively sensible and as such has been implemented by some authors. For example, Diggle, Morris, and Wakefield (2000) used informative priors together with the conditional likelihood for matched case-control data; similarly, Liao (1999) implemented a Bayesian mixture model for a single odds ratio in a series of case-control studies, where the contribution of each study is its conditional likelihood. Cook and Farewell (1999) suggested combining unconditional likelihoods from large matched studies together with conditional likelihoods from smaller ones;

from a Bayesian standpoint, this can be interpreted simply as the construction of a posterior from several matched studies, where the nuisance parameters are integrated out using priors that are equivalent to the conditional or unconditional approach as appropriate.

2. MAIN THEORETICAL RESULTS

If we adopt the marginal likelihood approach and assume that all β_j are independent samples from a common mixing distribution G , then the marginal likelihood contribution that we get from the vector of data \mathbf{X}_j of individual j is

$$\mathbb{P}_G(\mathbf{X}_j = \mathbf{x}_j) = L_{\text{cond}}(\alpha, \mathbf{x}_j) \mathbb{E}_G \mathbb{P}(C_j = c_j), \quad (3)$$

(see, e.g., Lindsay et al. 1991), where $L_{\text{cond}}(\alpha, \mathbf{x}_j)$ is the conditional likelihood contribution due to individual j , and the α_i are the “easiness” parameters. We allow G to depend on α . Because G is assumed identical for all individuals, we generally drop the subscript j , writing β and C for β_j and C_j , and define as *marginal probabilities*

$$\begin{aligned} m_c &= \mathbb{E}_G \mathbb{P}(C = c) \\ &= \mathbb{E}_G \frac{\beta^c s_c}{\prod_i (1 + \beta \alpha_i)}, \quad 0 \leq c \leq I, \end{aligned} \quad (4)$$

where s_c is the c th symmetric polynomial in $\alpha_1, \dots, \alpha_I$,

$$s_c = \sum_{\mathbf{x}': \mathbf{1}^T \cdot \mathbf{x}' = c} \prod_{i=1}^I \alpha_i^{x'_i}. \quad (5)$$

The dependence of the s_c on α is assumed implicitly throughout. Also note that all distributions G that give the same functional form of \mathbf{m} lead to the same marginal likelihood, and hence such G 's form an equivalence class of distributions that are effectively indistinguishable. The formulation of (3) leads directly to the following lemma.

Lemma 1. If we assume a common mixing distribution G on the β_j , then the integrated likelihood (3) is proportional to the conditional likelihood for all datasets and all values of α , if and only if all the marginal probabilities m_c associated with G are invariant with respect to α .

Proof. If we assume the invariance property, then the result is trivial. To prove the converse, suppose that we have data \mathbf{X} such that all individuals have $C_j = c$ for some particular value of c . Then the integrated likelihood is

$$L_{\text{cond}} \times m_c^I. \quad (6)$$

To get proportionality for all α , we clearly need m_c invariant with respect to α . Applying this argument for all $0 \leq c \leq I$, the lemma is proven.

We define an *invariant distribution* as one that satisfies the conditions of Lemma 1, and a set of *invariant marginal probabilities* as a vector $\boldsymbol{\mu}$ that has elements that may be written as $\mu_c = \mathbb{E}_G \mathbb{P}(C = c)$ for some invariant distribution G . Clearly, $\boldsymbol{\mu}$ is a special case of \mathbf{m} as defined in (4), and the notation is occasionally interchangeable.

Theorem 1. For any $I \geq 2$, and any α where all α_i 's are strictly positive and finite, invariant distributions G can be found.

The proof of this theorem is given in Appendix A, with some limited discussion.

3. EXAMPLES OF INVARIANT DISTRIBUTIONS

3.1 Example: A Coin-Tossing Invariant Distribution for Pair-Matched Case-Control Studies

In the common application of the Rasch model to matched case-control studies, the “ability” parameters β correspond to effects common to each matching, where subjects may be matched on height, weight, sex, and so on. The “individual” parameters α_i are constrained to be an unknown common value ψ , or 1, according to subject i 's case or control status. The outcome of interest is usually termed “exposure,” or some other event of interest, so ψ is interpretable as the ratio of odds of exposure for the case individual(s) to the odds of exposure for the control individual(s), considered fixed across all matched sets. The simplest and most common matched case-control study has one case and one control in each matched pair. Here, restricting the α parameters to be ψ and 1 is only insisting that they be identifiable, and so the 1 : 1 matched study is in fact the general case of the Rasch model for $I = 2$.

Given the foregoing interpretation, it is natural to parameterize the problem as

$$\psi = \alpha_2, \quad p_{1j} = \frac{\beta_j}{1 + \beta_j}, \quad p_{2j} = \frac{\psi\beta_j}{1 + \psi\beta_j},$$

so that p_{1j} and p_{2j} are the probabilities that the control or case in stratum j is exposed. For given ψ , p_{1j} and p_{2j} are of course functions of each other, and hence we can define mixing distributions on just one, or use the definitions jointly.

For a 1 : 1 matched study, consider the distribution $G_{1:1,1/2}$, where

$$\begin{aligned} p_{1j} &= 1/2 && \text{with probability } 1/2, \\ p_{2j} &= 1/2 && \text{with probability } 1/2. \end{aligned} \tag{7}$$

This can be interpreted as follows. A coin is tossed to decide whether the case or control will be the “reference,” and then another coin is tossed to see whether or not the reference is exposed. The exposure probability for the nonreference individual is decided by ψ .

$G_{1:1,1/2}$ has marginal probabilities $(m_0, m_1, m_2) = (1/4, 1/2, 1/4)$, which are trivially invariant to ψ and hence, by Lemma 1, will always lead to an integrated likelihood for ψ proportional to the conditional likelihood. By defining $G_{1:1,1/2}$ as given in (7), we can clearly see that it is symmetric in p_{1j} and p_{2j} and hence that it is invariant to switching the case and control labels. This attractive symmetry is discussed further in Section 6.1.

3.2 Example: Equivalent Continuous and Discrete Invariant Distributions

The example in Section 3.1 is a discrete distribution, formally a special case of the T -point distribution with two points of support. This distribution was also used in the work of Lindsay et al. (1991); relationships between the two approaches are discussed in Section 6.2. In calculating invariant distributions for $I = 2$ and other values, the T -point distribution makes for computational simplicity. However, invariant distributions need not be discrete; an example of a continuous invariant distribution

for the 1 : 1 matched case-control study, and hence generally for $I = 2$, is given when the p_{2j} are given density function

$$f(p|\psi) = \frac{1 - p + p\psi}{1 + \psi} + \frac{\psi^2}{(1 - p + p\psi)^3(1 + \psi)}, \tag{8}$$

for $0 \leq p \leq 1$. This distribution has marginal probabilities $(m_0, m_1, m_2) = (1/3, 1/3, 1/3)$.

These marginal probabilities m are also obtained if we adopt the somewhat different mixing distribution where

$$\begin{aligned} p_{1j} &= p_{2j} = 0 && \text{with probability } 1/6, \\ p_{1j} &= 1/2 && \text{with probability } 1/3, \\ p_{2j} &= 1/2 && \text{with probability } 1/3, \\ p_{1j} &= p_{2j} = 1 && \text{with probability } 1/6. \end{aligned} \tag{9}$$

This emphasises two further points. First, even though some of the most simply constructed invariant distributions gives positive support to discrete points, including the extreme values $p = 0, 1$ ($\beta = \mathbf{0}, \infty$), not all invariant distributions have this property. Second, and more generally, the points or intervals of support of any particular invariant G are not important for inference on α . As stated in Section 2, as long as distributions for β lead to the same marginal probabilities \mathbf{m} , invariant or otherwise, then without more information we cannot discriminate between them.

4. THE SPACE OF INVARIANT MARGINAL PROBABILITIES

In the previous sections we showed that invariant distributions exist and gave examples of their use. If we simply wish to obtain the conditional likelihood for some data with given I , then any invariant distribution G will suffice. However, it is possible that this part of the model may not fit well (as seen later in the example of Sec. 5.1) in which case the observed totals scores are compared with the arbitrarily chosen invariant marginal probabilities. For full-likelihood analyses where we want to maintain the equivalence to the conditional analysis for α and retain the entire marginal likelihood (3), optimal fitting of μ to the data will require that we select G from as large a space of invariant distributions as possible.

In Section 2 we remarked that different choices of G with the same marginal probabilities \mathbf{m} form an equivalence class and are effectively indistinguishable, giving the same marginal likelihood. When using these results as part of a full-likelihood analysis as described earlier, it is therefore sufficient to consider not the whole space of invariant G , but rather the space of points μ for which we can find invariant mixing distributions.

Therefore, for a given I , let us define \mathcal{M} as the space of all invariant marginal probabilities μ for which there exist invariant mixing distributions. In a full-likelihood analysis, we maximize μ over \mathcal{M} as well as α over the positive reals. A Bayesian alternative is to place a flat prior for μ over the range of \mathcal{M} . As for general marginal probabilities \mathbf{m} , the elements of such points must be nonnegative and add to unity. By Theorem 1, we know that \mathcal{M} is nonempty and can trivially be shown to be convex.

It is also straightforward to put an “outer bound” on \mathcal{M} . By its definition, \mathcal{M} is invariant to α , for all α being considered.

Provided that the (null) value where all $\alpha_i = 1$ is one of these possible α 's, it therefore suffices to deal with that case alone. But then the total scores $C_j \sim \text{bin}(I, \beta_j/(1 + \beta_j))$, and we see immediately that μ must lie in the space of binomial mixtures,

$$\mathcal{B} = \left\{ \mu : \mu_c = \mathbb{E}_Q \left(\binom{I}{c} q^c (1 - q)^{I-c}, 0 \leq c \leq I \right), \quad (10) \right.$$

where Q can be any distribution on $[0, 1]$. Therefore, $\mathcal{M} \subseteq \mathcal{B}$.

Examples of this property can be found in the invariant distributions given earlier. In the examples of Section 3.2, for $I = 2$, the values of μ obtained can be represented as binomial mixtures through (valid) linear combinations of the points $(0, 0, 1)$, $(1, 0, 0)$, and $(1/4, 1/2, 1/4)$. Alternatively, these values of μ_c could be obtained by setting $q \sim U(0, 1)$ in (10).

The space of binomial mixtures also appears in the work of Lindsay et al. (1991), who used mixing distributions with extremely flexible moments. These authors established a necessary and sufficient concordance condition for equivalence of MMLE and CMLE—that the vector of normalized observed marginal probabilities,

$$\mathbf{M} = \{M_c : M_c^{\text{obs}} = \#\{j : C_j = c, 1 \leq j \leq J\}/J, \quad 0 \leq c \leq I\}, \quad (11)$$

must be an element of \mathcal{B} . In our approach, although the equivalence between MMLE and CMLE holds without conditions for all datasets, the restrictions that we have found on \mathcal{M} imply that a necessary condition for additionally fitting the total scores perfectly is that $\mathbf{M} \in \mathcal{B}$. In Appendixes B, C, and D we prove that it is also sufficient in some (useful) special cases, but not in general. For any particular value of \mathbf{M} , we can test whether $\mathbf{M} \in \mathcal{M}$ by evaluating the solubility of the moment problem (A.5) for general values of α , but expressing \mathcal{M} in a more concise way remains an open problem.

We have seen that use of the conditional likelihood implicitly assumes that $\mathbf{m} \in \mathcal{M} \subseteq \mathcal{B}$. A natural consequence of this is to utilize the observed \mathbf{M} to criticize the conditional “model.” The simplest and most common situation is where $\mathcal{M} = \mathcal{B}$; we know that if an observed \mathbf{M} lies outside of \mathcal{B} , this provides evidence of underdispersion; that is, the conditional independence assumption may have been violated. Thus we may conclude that the conditional approach is disregarding information about model failure relevant to the interpretation of the α parameters and in response either reject the model entirely or inflate confidence intervals for α to reflect our skepticism about the model. Conversely, $\mathbf{M} \in \mathcal{B}$ here indicates that the conditional likelihood approach may capture the totality of information available for inference on α , and so we may accept the analysis. This can be formalized if we calculate the deviance of the model in the standard way, using both parts of the likelihood in (3). An example of this deviance is given in Section 5.2, where this theory provides a potentially powerful measure of goodness of fit that can be appended to the usual conditional analysis in a natural way.

A more complex situation occurs when $\mathcal{M} \subset \mathcal{B}$. If $\mathbf{M} \notin \mathcal{B}$, then the foregoing argument holds, and we may similarly accept the conditional analysis if $\mathbf{M} \in \mathcal{M}$. However, if $\mathbf{M} \in \mathcal{B} \setminus \mathcal{M}$, then we have evidence on which to criticize the conditional analysis, but not the Rasch model assumptions. Such values of \mathbf{M} indicate that the marginal probabilities are not consistent

with the conditional analysis and contain useful information about the α parameters. If this information is in line with the conditional likelihood's inference, then we may reach the (acceptable) conclusion that the conditional approach is not completely efficient. If the two sources of information about α are in strong conflict, then this should be investigated further to ensure a reliable analysis. This idea is not formalized here, but will form the basis of future work. However, note that for a general Rasch dataset, development of this measure of goodness of fit requires that \mathcal{M} be described exactly, and so fully parameterizing the set \mathcal{M} is of more than theoretical appeal.

5. APPLICATIONS

Reinterpreting the specialized conditional likelihood for Rasch models from a mixture viewpoint is more than just an interesting fact. It makes available for Rasch analysis, and hence for analysis of matched case-control studies, any interpretive and computational tools already in use elsewhere in the full-likelihood or Bayesian analysis literature. In this section we illustrate a selection of these tools. Section 5.1 uses explicitly constructed invariant distributions, but the other applications arise directly from the full-likelihood derivation of the conditional approach. We also note that this derivation is also straightforwardly extended to allow for missing data, particularly misclassification errors in the data, and this has been explored and applied elsewhere (Rice 2003).

5.1 Markov Chain Monte Carlo Evaluation of the Conditional Likelihood

As mentioned in Section 1, evaluating the conditional likelihood can be difficult. Computational “tricks” based on Poisson regression (Tjur 1982; Agresti 1993) can be effective, and for small values of I these can be implemented in various standard statistical packages, including SPSS (TenVergert, Gillespie, and Kingma 1993), SAS (Christense and Bjorner 2003), and R/S (Lindsey 2000). For larger problems, several authors (Gail, Lubin, and Rubinstein 1981; Strawderman and Wells 1998; Mehta et al. 2000; Corcoran, Mehta, Patel, and Senchaudhuri 2001) have considered algorithms or approximations that streamline the process. An estimate of when a problem is “large” in this way has been given by Mehta and Patel (2002, p. 244). Lemma 1 arrives at the exact conditional likelihood as a marginal construction without ever specifying its full form. This directly introduces a new technique to the evaluation literature: implementing our model using standard Markov chain Monte Carlo (MCMC) techniques within a Bayesian framework, including an invariant mixing distribution as a prior on the nuisance parameters and a flat prior on α . By doing this, the posterior will be exactly proportional to the conditional likelihood for α . This result also holds for 1–1 functions of α (e.g., $\log \alpha$).

We illustrate this process with the well-known dataset of Stouffer and Toby (1951), which has been documented and analyzed by, for example, Lindsay et al. (1991) and Follman (1988). Each of 216 individuals responded to a series of four questions. We are interested in estimating the relative difficulty of the four questions, and we implement the usual conditional analysis by way of our mixture derivation.

The first requirement is an invariant distribution for $I = 4$. Appendix E gives such a distribution, with invariant marginal

probabilities $\mu_c = 1/5, c = 0, \dots, 4$. Given this distribution, it is straightforward to obtain the conditional likelihood as an integrated likelihood if a fully Bayesian approach is adopted. We use the distribution from Appendix E as a prior for the nuisance β parameters and, as in previous analyses of these data, assume that $\alpha_4 = 1$ for identifiability. We place independent, noninformative $U(-5, 5)$ priors on the other $\log \alpha$ parameters, forcing the posterior distribution obtained for $\log \alpha$ to be proportional to the conditional likelihood over the $(-5, 5)$ range and 0 elsewhere. (The log scale is used here to aid posterior normality.) In particular, we should find that the posterior mode is identical to the CMLE, and that the usual normal approximation to the conditional likelihood is closely reproduced by a normal approximation of the posterior distribution.

A total of 10,000 samples from the posterior distribution for this model was obtained using the WinBUGS software (Spiegelhalter, Thomas, and Best 2003), after burn-in of 500 samples (code available from the author on request). Table 1 summarizes the output from the MCMC sampling, together with standard estimators derived from the conditional likelihood's maximum, the observed information matrix, and profile likelihoods. We see very close concordance between the posterior mean and median estimates and the CMLEs, which of course here correspond to the posterior mode. The exact posterior credible intervals are similarly close to the asymptotic confidence intervals, as illustrated for α_1 in Figure 1. This also illustrates the approximate normality of the posterior, suggesting that the classical normal approximations are well founded.

Given an invariant prior (mixing distribution) for the value of I in question, the technique used here generalizes simply for use with any Rasch dataset. Note that the choice of an appropriate prior can be relaxed if we use an empirical Bayes argument, where for responses \mathbf{X}_j with total score C_j , we use a prior for β_j that keeps only m_{C_j} , rather than the whole vector \mathbf{m} , invariant. Although from an inferential standpoint this is "cheating" by using the data twice, the invariance required for Lemma 1 is not affected, and the required posterior is produced.

In contrast with existing asymptotic theory, here no approximations are made regarding the normality or any other property of the posterior, and thus our analysis is exact, up to the error induced by the MCMC process. Of course, efficiency, memory usage, and other technical considerations for imple-

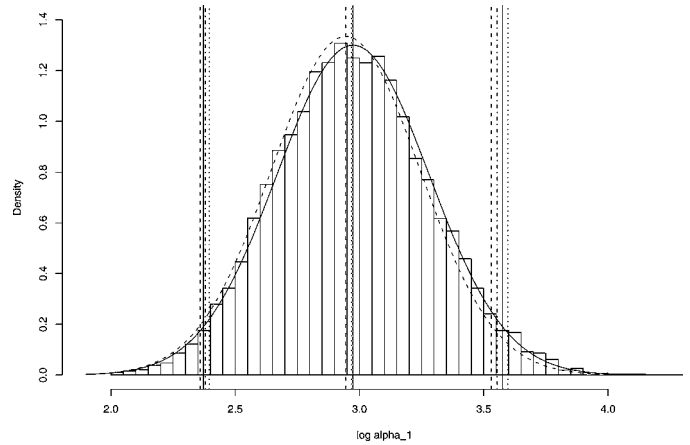


Figure 1. Histogram of MCMC Output for α_1 , From Analysis of the Stouffer Toby Data. Superimposed are point estimates, intervals, and normal densities from the MCMC data; conditional likelihood, point estimates, and intervals from the quantiles of the MCMC data; and a confidence interval derived from using profile techniques on the conditional likelihood. [— normal approx to MCMC data; --- normal approx to conditional; ···· quantiles of MCMC data; - · - · profile conditional (interval only).]

mentation require comparison with existing methods, and this work remains to be done. In addition, a general technique for finding simple invariant distributions is desirable. This is not yet available, but we do include a computationally simple algorithm for finding a discrete invariant distribution: Define the points of support as x_i with associated probabilities λ_i , where $1 \leq i \leq I + 1$ and the λ_i are nonnegative and sum to 1. Then:

1. Choose $I + 1$ points of support x_i (which may depend on α).
2. Choose a vector $\mu \in \mathcal{M}$ of marginal probabilities, invariant to α .
3. Solve the I linear equations for λ_i which ensure that $\mathbf{m} = \mu$.
4. Check that all λ_i are nonnegative, for all values of α . If they are not, then refine the choice of \mathbf{x} and repeat.

Note that appropriate μ may be chosen "automatically" by following the proof of Theorem 1, and also that, using the empirical Bayes method, in step 3 we need equate only one element of \mathbf{m} and μ . Refining this algorithm to produce automatic choice of x_i remains to be done in future work.

Table 1. Summary Statistics From the MCMC Analysis of the Stouffer Toby Data

		Parameter		
		$\log \alpha_1$	$\log \alpha_2$	$\log \alpha_3$
MCMC output	Mean	2.974	1.256	1.330
	Median	2.968	1.253	1.322
	Standard deviation	.307	.269	.268
	Mean $\pm \Phi^{-1}(.975) \times$ SD	(2.37, 3.58)	(.73, 1.78)	(.81, 1.86)
	2.5%, 97.5% quantiles	(2.40, 3.60)	(.74, 1.80)	(.82, 1.87)
Conditional approach	Climate	2.944	1.244	1.320
	Estimate(SD)	.299	.263	.263
	Mean $\pm \Phi^{-1}(.975) \times$ estimate(SD)	(2.36, 3.53)	(.73, 1.76)	(.80, 1.84)
	Profile interval	(2.38, 3.55)	(.74, 1.78)	(.82, 1.85)

NOTE: The sample's mean, variance, and associated intervals agree almost exactly with those predicted applying standard asymptotic theory to the conditional likelihood.

5.2 Deviance-Based Measures of Goodness of Fit

Several tests of goodness of fit are available for the Rasch model (for a full review, see Glas and Verhelst 1995). As discussed in Section 4, our new derivation of the conditional “model” in a full-likelihood framework allows us to use standard measures of deviance to assess goodness of fit. The test that we derive in this way for situations where $\mathcal{M} = \mathcal{B}$ checks for possible violations of the conditional independence assumptions; similar likelihood-based approaches include Kelderman’s (1984) methods of assessing potential interactions between the α parameters and Andersen’s (1973) method of testing for heterogeneity of α parameters across subgroups in the data. But these approaches do not use our restriction of \mathbf{m} to \mathcal{M} , and therefore we believe that the tests proposed here are new to the literature.

The approach is simple. We first calculate a deviance statistic in the usual way, by maximizing the full-likelihood under the restrictions that $\mathbf{m} \in \mathcal{M} = \mathcal{B}$ and, alternatively, letting \mathbf{m} take any value in the unit simplex. Twice the log-likelihood ratio can then be compared to an asymptotic distribution to give a test of goodness of fit. In situations where $\mathbf{M} \in \mathcal{M}$, the deviance will be 0, and the conditional analysis can be accepted. Otherwise, the restricted estimate of \mathbf{m} will lie on the boundary of \mathcal{M} , and so we follow the work of, for example, Self and Liang (1987) to obtain the distribution of the deviance statistic. In the case where $I = 2$, this is simply a 50 : 50 mixture of χ_0^2 and χ_1^2 . Following Lindsay et al. (1991), it may be reasonably assumed that the deviance is distributed as some form of chi-squared distribution, but the exact form for $I > 2$ may be more complicated.

Sprott (1975) gave a simple example of 1 : 1 matched case-control data where the marginal totals alone provide sufficient information on which to reject the null hypothesis that $\psi = 1$. We use a slightly more specific form of his example to illustrate our techniques. Assume a matched-pair dataset that consists of $2K$ discordant pairs and no concordant pairs, and additionally that $r_{01} = r_{10} = K$. Using the conditional likelihood alone, whatever the value of K , we always estimate ψ to be 1. The normal confidence interval for $\log \psi$ shrinks with $K^{-1/2}$, leading us to conclude that $\psi = 1$ with increasing certainty.

But using the mixture approach with an invariant mixing distribution that gives equal weight to all values of \mathbf{m} in \mathcal{M} , the marginal likelihood is

$$L_{\text{cond}}(\psi)m_0^0m_1^{2K}m_2^0, \tag{12}$$

where \mathbf{m} can take any value in \mathcal{M} . The likelihood factorizes into L_{cond} and the terms in \mathbf{m} , so the point estimation of ψ and its confidence intervals are unchanged from the conditional approach. However, as shown in Figure 2, the marginal exposure probabilities cannot be made to fit the observed marginal totals. The best fit for the marginal totals occurs when \mathbf{m} is estimated as $\{1/4, 1/2, 1/4\}$, on the boundary of \mathcal{M} , and comparing the model to one where \mathbf{m} can take any value in the unit simplex, we get a deviance of $2K \log(2)$, which grows proportionally with K . Comparing this with a 50 : 50 mixture of χ_0^2 and χ_1^2 (or, indeed, any chi-squared distribution), we see that as K increases, we obtain increasing evidence that our model fits the data poorly; the deviance will eventually exceed any threshold

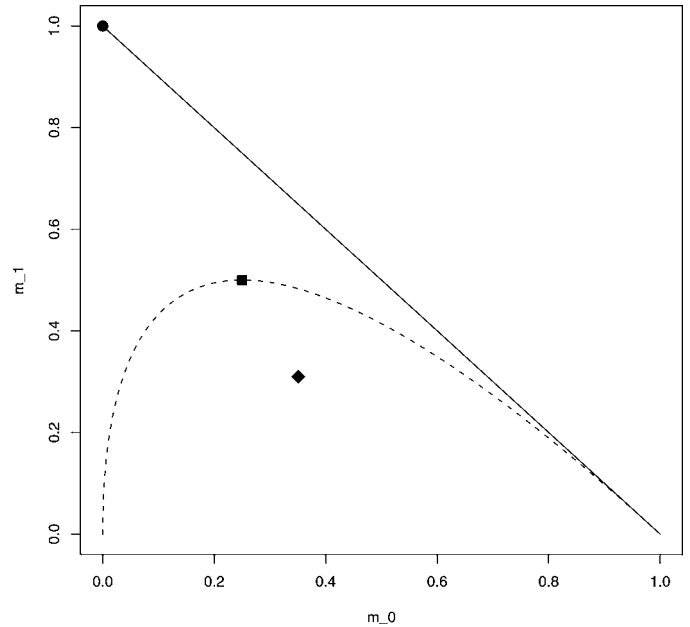


Figure 2. The Space of Marginal Probabilities for 1 : 1 Matched Studies. m_0 and m_1 are the marginal probabilities of 0 and 1 exposure per matched pair. \mathcal{M} is seen to be a subset of the unit simplex. [— upper boundary of unit simplex; --- upper boundary of space of constant m ; ● Sprott example (data); ■ Sprott example (fitted); ◆ ISIS example (data).]

of significance. For increasing values of K , despite the narrowing estimate of ψ , we thus should reject the estimate and the model with increasing certainty. Intuitively, as Sprott noted, this is because the data indicate that rows $i = 1, 2$ are not actually independent, hence violating a key assumption. This is not uncovered by the naïve conditional-based estimate, but using the equivalent marginal approach, it becomes a natural part of the analysis.

The example also motivates us to ask how often the (normalized) observed margins of the dataset \mathbf{r} lie within \mathcal{M} . Under Lebesgue measure, \mathcal{B} has been found to cover a vanishingly small proportion of the unit simplex (Wood 1992). Moreover, under sampling from a distribution with marginal totals central in \mathcal{B} , the proportion of samples of a fixed size with margins within \mathcal{B} has been found to “decrease rapidly” (Wood 1999) as the dimension I grows. It therefore appears plausible that we will often find marginal totals outside \mathcal{M} , and the deviance-based measure of fit suggested here will be useful.

5.3 Inference on Complex Functions of α Parameters

Inference on complex functions of the parameters can be difficult in the classical framework, but again a Bayesian interpretation gives us access to MCMC methods, which may make the problem tractable. We illustrate this by considering inference on the ranks of α_i , the item parameters. The rankings of the α_i are a discrete-valued, highly nonlinear, and noninvertible complex function of the vector α , but ranks, and confidence intervals around them, would be of interest in Rasch analysis of, for example, students in a competitive exam or in league tables of institutions based on performance indicators (Goldstein and Spiegelhalter 1996). Asymptotic techniques are not easily applied to obtain such confidence intervals, but the related idea

of Bayesian credible intervals has proven useful (Morris and Christiansen 1996).

We use as an example a ranking of five weather forecasters, based on the accuracy of their forecasts of maximum temperature. The data, previously considered by Brooks, Witt, and Eilts (1997), consist of the 1-day-ahead forecasts from three network TV stations, a daily newspaper, and the Oklahoma City National Weather Service, together with the observed temperature, over a period of 338 days. For the purposes of this analysis, we deem a weather forecast F_1, \dots, F_5 to be "correct" if the forecast maximum temperature is accurate to the observed value to within ± 2 degrees Fahrenheit. The full dataset is given in Table 2.

We apply the Rasch model to this data in the usual way, assigning each forecaster a parameter α_i , where we insist that $\sum_i \alpha_i = 0$ for identifiability. Assuming an invariant mixing distribution as the prior for the nuisance β parameters, we can justifiably model the data in Table 2 as a single realization from a multinomial distribution of index 338, where the cell probabilities for each response pattern are proportional to the conditional likelihood contribution from that pattern. In the WinBUGS code for this analysis, the likelihood is evaluated in just this way. We assume independent $U(-4, 4)$ priors on $\log \alpha_i$, for $1 \leq i \leq 4$.

After a burn-in of 500 iterations, we use a sample of 10,000 MCMC iterations. At each iteration, the recorded values for each of the α_i are also ranked from 1 (worst) to 5 (best), giving a

posterior distribution for the ranking of the different forecasters. The posterior distributions for the various $\log \alpha$ and their ranks are given in Figure 3. The figure shows that despite the approximate normality shared by the posteriors for the α parameters, their ranks are nonnormal, highly skewed in some cases, with great variation in general between all of the rank posteriors.

Figure 4 plots the posterior means together with the credible intervals bounded by the 2.5% and 97.5% quantiles of the posterior distributions for each α_i . This figure clearly illustrates the coarsening of the inference that we must accept if we restrict the analysis to the rank ordering alone; for example, forecasters F_1 and F_3 have identical rank-credible intervals, although the intervals around α_1 and α_3 are obviously different. These findings follow the general pattern of other studies of ranking (Goldstein and Spiegelhalter 1996; Marshall and Spiegelhalter 1998), showing that the ranks alone, although of interest, may be a rather crude and unstable measure of relative performance.

Using the ranks does clarify some points of inference, however. We see that the credible interval for the rank of forecaster F_2 does not extend down to the third position, which is not automatically clear if we examine just the credible intervals for the α parameters. Similarly, the interval around α_5 suggests that forecaster F_5 is performing relatively well, but the parameter uncertainty in α_5 and the other parameters combines to give a rather large interval for that forecaster's rank. At the 95% level, all ranks are credible for forecaster F_5 except the absolute lowest. Using this method therefore gives us a better idea of how much of the forecaster's ranking is due to chance and how much is due to his or her relative merit. Also note that, although not shown here, a measure of uncertainty of the overall ranking of $\{F_1, \dots, F_5\}$ could be calculated from the same MCMC chain by tabulating the number of posterior samples with ranking F_1, F_2, F_3, F_4, F_5 ; F_2, F_1, F_3, F_4, F_5 ; and so on.

5.4 Prior Information Added to a Case-Control Study

Our model allows a Bayesian interpretation of conditional likelihood, and thus the use of informative prior distributions with standard analyses of matched-case control studies. We illustrate how this may be useful in some recent work—the matched case-control section of the (ISIS) study of the gastric infection *Helicobacter pylori* as a risk factor for heart disease, as reported by Danesh et al. (1999). We are interested in $\theta = \log \alpha$, the log-odds ratio of *pylori* exposure between cases who suffered heart attacks and controls. Danesh et al. also reported a summary of five similar previous studies with which their ISIS analysis is "compatible," although they did not find a statistically significant odds ratio estimate. Summaries of the previous studies and the ISIS data are plotted together, but not formally combined. Here we use the data from the previous studies to construct prior distributions for θ , and find that the collected studies suggest a more significant result than the ISIS data alone. The ISIS data and the point estimates and standard errors from the previous studies are given in Table 3. (Note that in the analysis by Danesh et al., $\hat{\theta}$ was adjusted to allow for other recorded risk factors, to very slight effect. For adjustments, potential biases, inclusion criteria and other details, see that article and the references therein.)

We summarize the prior evidence using two sets of assumptions, corresponding to "exchangeable" and "historical bias"

Table 2. Raw Data Extracted From the Brooks et al. Weather Forecasting Dataset, Recording Correct Maximum Temperature Forecast From 5 Forecasters Over 338 Days

Total score	Count	Response pattern					Count
		F1	F2	F3	F4	F5	
0	98	0	0	0	0	0	98
1	49	1	0	0	0	0	7
		0	1	0	0	0	16
		0	0	1	0	0	11
		0	0	0	1	0	10
		0	0	0	0	1	5
2	39	1	1	0	0	0	3
		1	0	1	0	0	2
		0	1	1	0	0	7
		1	0	0	1	0	4
		0	1	0	1	0	8
		0	0	1	1	0	2
		1	0	0	0	1	2
		0	1	0	0	1	2
		0	0	1	0	1	3
		0	0	0	1	1	6
3	42	1	1	1	0	0	4
		1	1	0	1	0	4
		1	0	1	1	0	1
		0	1	1	1	0	2
		1	1	0	0	1	9
		1	0	1	0	1	5
		0	1	1	0	1	8
		1	0	0	1	1	2
4	58	0	1	0	1	1	3
		0	0	1	1	1	4
		1	1	1	1	0	6
		1	1	1	0	1	23
		1	1	0	1	1	10
5	52	1	0	1	1	1	9
		0	1	1	1	1	10
		1	1	1	1	1	52

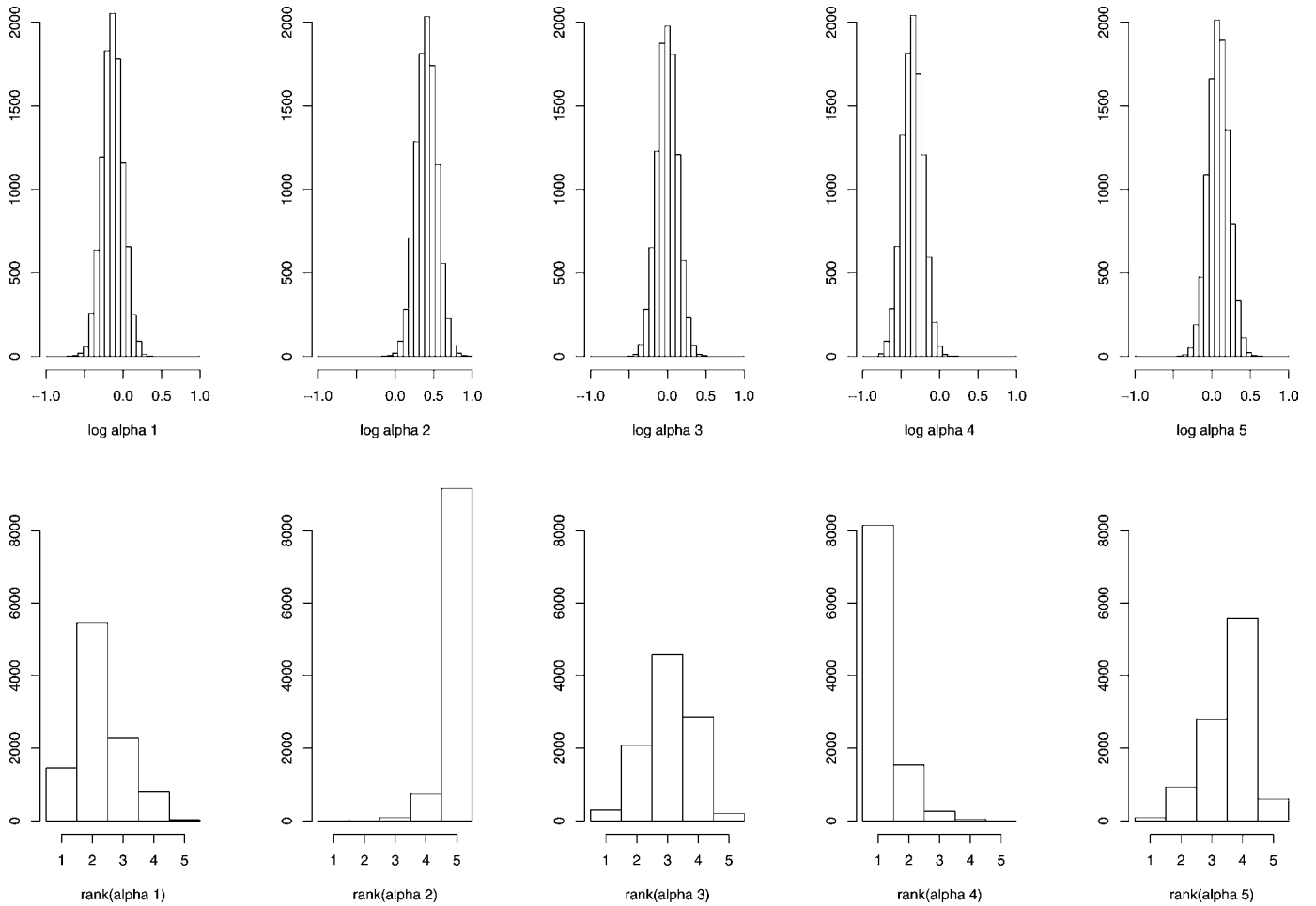


Figure 3. Histograms of MCMC Output From Weather Forecaster Data, With Posterior Distributions of Raw α Parameters and Their Rankings.

models (Speigelhalter, Abrams, and Myles 2003). In each set, the previous studies' data are assumed to approximate a normal likelihood on the log scale. Denoting the historical data by y_1, \dots, y_5 , this can be written as

$$y_h \sim N(\theta_h, \sigma_h^2), \quad 1 \leq h \leq 5.$$

First, assuming that the θ_h and θ are exchangeable with a normal distribution, so that

$$\theta_h, \theta \sim N(\mu, \tau^2),$$

the appropriate prior for θ is the predictive distribution of θ in

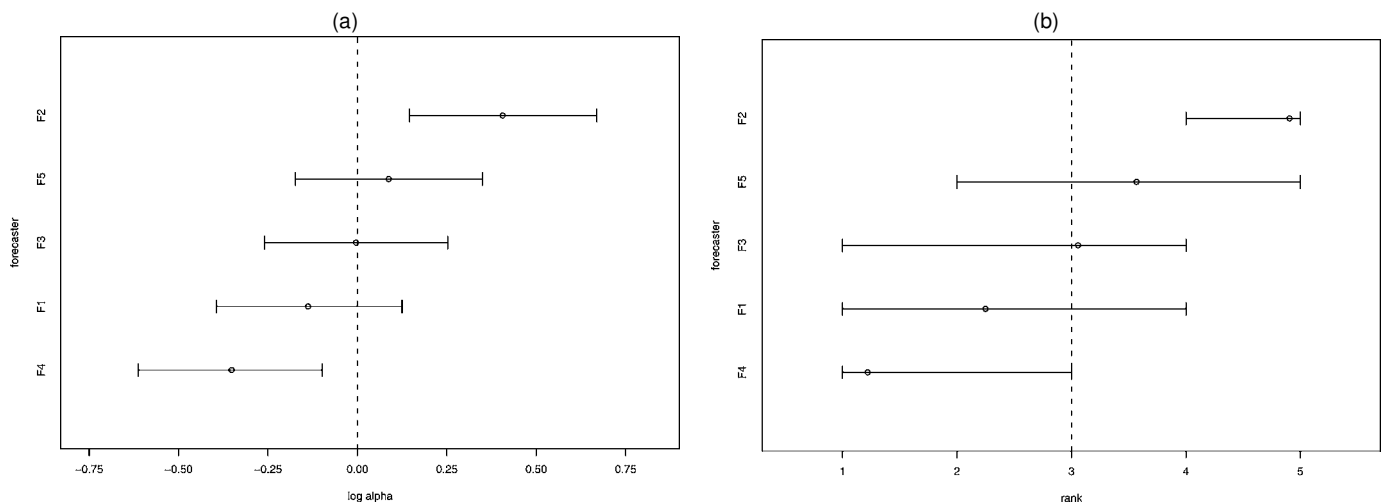


Figure 4. Boxplots of Point Estimates and Credibility Intervals for Comparing Weather Forecasters. Shown are the 2.5% quantile, mean, and 97.5% quantile from the posterior distribution for (a) each α parameter and (b) the ranked α values.

Table 3. Matched-Pair Data From the ISIS Trial

	Case exposed	
	-	+
Control exposed	179	91
	67	173

a new study, given by

$$\theta \sim N\left(\frac{\sum_h y_h w_h}{\sum_h w_h}, \frac{1}{\sum_h w_h} + \tau^2\right), \tag{13}$$

where $w_h = 1/(\sigma_h^2 + \tau^2)$. We estimate τ^2 using a profile likelihood argument (Hardy and Thompson 1996) on the joint likelihood for μ and τ , and use this value in (13) to give what we term the “exchangeable” prior. Second, assume that each θ_h is biased by an amount δ_h , so that $\theta_h = \theta + \delta_h$ and $\delta_h \sim N(0, \sigma_\delta^2)$. This leads to prior

$$\theta \sim N\left(\frac{\sum_h y_h w_h}{\sum_h w_h}, \frac{1}{\sum_h w_h}\right), \tag{14}$$

where $w_h = 1/(\sigma_h^2 + \sigma_\delta^2)$. Using the past data, σ_δ is estimated identically to τ earlier, and, using this estimate in (14), we obtain the “bias prior.”

Medians for the two prior distributions, together with their .5% and 99.5% quantiles, are plotted in Figure 5, which also contains the prior information used. As might be expected given its construction, the exchangeable prior is much more diffuse than the bias prior, although they have the same median.

The posteriors given by combining the ISIS likelihood with the two priors are also shown in Figure 5, and they are seen to be closely concordant, suggesting reasonable robustness to the priors used. Figure 2 shows that the marginal totals can be fitted well by using invariant distributions, suggesting that there is no need to adjust for poor fit as described in Section 5.2.

The upper extent of the 99% credible intervals formally confirms the conclusions of Danesh et al. (1999) that any association of *H. pylori* infection with heart attack incidence is “moderate.” Using just the likelihood from the ISIS study and a flat $U(-\infty, \infty)$ prior for θ , we find that 2.75% of the posterior lies below 0, the null value. Using the exchangeable prior for θ , this shrinks to 1.38%, and using the less diffuse bias prior yields .12%. In a formal way, the use of the combined studies’ information has therefore allowed us to reach a more precise conclusion; both posteriors suggest stronger evidence for an association than the ISIS study alone, which is just nonsignificant at the usual 5% level.

Table 4. Summary of Results From Previous Studies for Use in Prior Elicitation

Study (first author, journal, year)	Estimate θ_j	Standard error σ_h
Balaban, <i>Gastroenterology</i> , 1996	.76	.47
Rathbone, <i>Heart</i> , 1996	.05	.19
Aceti, <i>British Medical Journal</i> , 1996	1.72	.55
Morgando, <i>Lancet</i> , 1995	1.51	.46
Ponzetto, <i>British Medical Journal</i> , 1996	1.71	.62

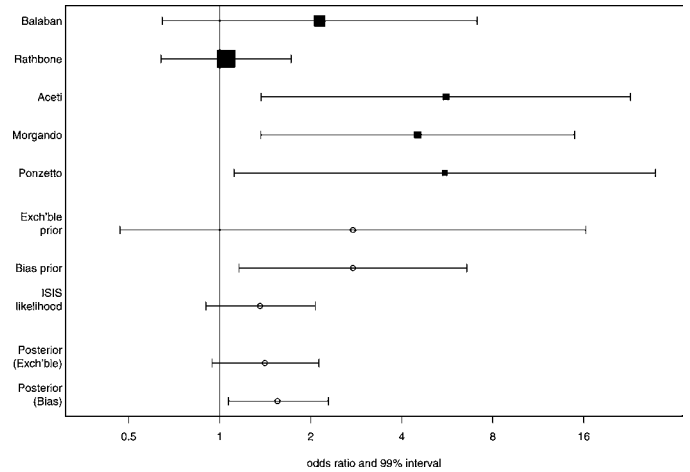


Figure 5. Summary Information for ISIS Example. This comprises five previous estimates of the odds ratio of helicobacter pylori infection among heart attack cases, two prior distributions derived from these previous studies, the likelihood from the matched case-control part of ISIS, and resultant posterior likelihoods. Box sizes in the summary of previous studies are proportional to number of cases; lines represent 99% confidence or credible intervals as appropriate.

6. DISCUSSION

6.1 Parameterisation and Symmetry Properties

As noted in Section 1, the present work follows a series of articles relating marginal and conditional methods in which it usually was assumed that the mixing distribution on β is free of α . Our new results are possible only because we have relaxed this assumption; even though it may be written implicitly, as in (7), our invariant mixing distributions have a nontrivial dependence on α .

From a Bayesian perspective, such an assumption means that if one gave the same set of subjects a new test with different relative difficulties, then the assumed ability distribution would change. Some analysts may find this property undesirable. However, the same property holds for apparently innocuous priors that might be used in Bayesian analysis of Rasch models. Consider a $U(0, 1)$ prior for p_{1j} in Section 3. If we retain the same set of absolute prior beliefs, but reparameterize with p_{2j} as the nuisance parameter, then the prior’s density now depends on α . Therefore, the freedom from α is seen to be dependent on the parameterization and is not inherent in the prior beliefs themselves. In Rasch models, we are trying to compare individuals relatively, without a natural absolute “baseline” level, and so this parameterization is a particular problem. We therefore suggest that for Rasch models, mixing distributions with some explicit α dependence at least be considered. From a much more pragmatic standpoint, the general acceptance of the conditional likelihood as a good basis for inference also suggests that such considerations are reasonable.

Accepting the notion of α dependence in the mixing distribution for β , the use of invariant distributions means that we need choose only marginal probabilities μ and not any particular nuisance parameter, which is attractive in light of the foregoing argument. Our invariant distributions also give some indication of what type of dependence may be analytically sensible.

Consider the elements of \mathbf{m} , given by

$$m_c = \mathbb{E}_G \frac{\beta^c s_c}{\sum_{c'} \beta^{c'} s_{c'}}, \quad 0 \leq c \leq I. \quad (15)$$

As we have seen, any invariant distribution for β keeps all m_c constant with respect to α . Therefore, if we wish to find a distribution G with fixed marginal probabilities $\mathbf{m} = \boldsymbol{\mu}$, then we need to satisfy a series of I equations, all of which depend on the α_i only through the various $s_{c'}$. If the general form of “candidate” distributions considered has exactly I free parameters, then an invariant solution G will therefore depend on α_i only through the values of $s_{c'}$, and hence will be entirely unchanged by relabeling the α_i . This symmetry property seems natural and desirable for G . Intuitively, we do not wish our beliefs about the ability of individuals to be affected by relabeling the questions that they are asked, and, therefore, interpreting G as a Bayesian prior, we can still reasonably claim that invariant G ’s are non-informative about the α parameters.

The relabeling property again directly challenges the non-informativeness of many seemingly straightforward Bayesian priors. For example, if we again use a “noninvariant” $U(0, 1)$ prior for p_{1j} in Section 3, this leads to marginal probabilities m_0, m_1, m_2 and that are functions of the odds ratio ψ . These are different from the marginal probabilities that would be obtained if we placed the $U(0, 1)$ prior on p_{2j} , which could be equally well chosen as the nuisance parameter to be eliminated. In fact, \mathbf{m} is reversed if we choose this alternative prior, and so we should expect the subsequent analyses to give different results for any dataset that does not have exactly equal numbers of concordantly exposed and unexposed pairs.

Any invariant G found in the way mentioned earlier, by fixing I free parameters in some candidate distribution, can of course be thought of as a special case of some yet more general candidate distribution, with more free parameters. As long as the marginal probabilities m are unchanged, these extra parameters could be set arbitrarily, potentially to terms that are not symmetric in the α_i . For example, on the H “scale” used in the proof of Theorem 1, the moments above $\mathbb{E}X^I$ are not fixed, apart from the restriction that they are indeed moments of a positive distribution, and so one could choose them to depend arbitrarily on α . It is therefore possible to construct mixing distributions that are invariant in the sense of Lemma 1 but do not have the relabeling property. However, because the moments above $\mathbb{E}X^I$ do not feature in the marginal likelihood terms m , such asymmetry is irrelevant for practical purposes. The marginal likelihood (and subsequent analysis) will be identical to that obtained using a distribution G that remains entirely unchanged by relabeling the α_i ’s.

The relabeling property, although attractive, is additionally not sufficient to ensure that a distribution is invariant in the sense of Lemma 1. This is illustrated by the work of Severini (1999) for the special case of $n : n$ matched case-control studies, which we view as a Rasch model with an $I = 2n$ items, comprising n “cases” with $\alpha_i = \psi$, and n “controls” with $\alpha_i = 1$. Then, similarly to the example of Section 3, we can define the “ability” parameters β in terms of the case and control “probabilities of success,”

$$p_{\text{case}} = \frac{\psi\beta}{1 + \psi\beta}, \quad p_{\text{control}} = \frac{\beta}{1 + \beta}. \quad (16)$$

Under a rather different parameterization, Severini (1999) showed that the mixing distribution that assumes that

$$\begin{aligned} p_{\text{case}} &\sim U(0, 1) && \text{with probability } 1/2, \\ p_{\text{control}} &\sim U(0, 1) && \text{with probability } 1/2 \end{aligned} \quad (17)$$

leads to a marginal log-likelihood that is $O(n^{-3/2})$ equivalent to the conditional log-likelihood. Severini’s distribution is clearly unchanged by relabeling cases and controls but is not invariant in the sense of Lemma 1. Although not explored here, the obvious similarity between the mixing distributions given by (17) and (7) suggests that even though exact equivalence may not be achieved, good asymptotic approximations to the conditional likelihood can be obtained if we insist on the relabeling property.

6.2 Relationship With Other Work

The reader will note numerous citations here to the work of Lindsay et al. (1991), to which this analysis is strongly related. The approach of Lindsay et al. (1991) is to semiparametrically assume that the mixing distribution for β is free of α and fits the marginal probabilities m exactly to the marginal totals \mathbf{M} . This can be done only when the normalized \mathbf{M} lies in \mathcal{B} . Our approach here is instead to first semiparametrically assume only that the marginal probabilities \mathbf{m} are invariant to α , and also to let the invariant value of \mathbf{m} take any value in the space \mathcal{M} . For some applications, this second criterion will not affect the analysis and can be omitted. Thus when $\mathcal{M} = \mathcal{B}$, using both of our criteria, we will fit \mathbf{M} exactly, and although our mixing distribution depends on α , we have satisfied the main criterion of Lindsay et al. (1991). Therefore, the two approaches are related but are not equivalent.

Another similarity comes when specific choices of G are considered. Lindsay et al. (1991) showed that the T -point distribution for β , with $(I + 1)/2$ unspecified points of support, fulfills their semiparametric requirements. In particular, the T -point distributions considered by Lindsay et al. (1991) do not depend on α . In this article we have shown that specific instances of the T -point distribution that do have some dependence on α can fulfill one or both of our criteria. Therefore, some invariant distributions will also be similar in form to those considered by Lindsay et al. (1991), but again the two are not equivalent.

A final relationship is with the work of Lauritzen (1988), where the conditional distribution (2), together with some degenerate measures, are shown to make up the extremal family for the Rasch model. The extremal family consists of measures that give degenerate distributions to the total scores, ensuring that the total scores can take only the value observed. The proportionality that we seek in Lemma 1 is a weaker condition than this, insisting only that the distribution of total scores be free of α . Although in this article we do not further explore the connections between these two results, we do note that because invariant distributions G are not analytically affected by different values of α , they may be interpreted as “extreme points” of the space of possible mixing distributions. In the examples of Sections 3 and 5.1, this interpretation can still be retained when the set of possible mixing distributions is restricted to those with the relabeling property described in Section 6.1.

6.3 Conclusions

Derivation of the the full conditional likelihood for the Rasch model via the alternative mixture approach is, we believe, both theoretically interesting and practically relevant, because the results justify extending the “tried and tested” conditional system of analysis with a range of novel and practical innovations. These innovations are relatively simple to implement once the mixture framework is in place. We have mentioned several opportunities for further research, notably the general description of \mathcal{M} given in Section 4 and the algorithm for finding invariant distributions discussed in Section 5.1. However, the results presented here also suggest the possibility of alternative derivations for other nonstandard likelihoods, perhaps the most commonly used being Cox’s partial likelihood for survival analysis (Cox 1975). The proportionality argument used in Lemma 1 appears to be applicable in almost any situation where one wants to extract some pseudolikelihood for the parameters of interest from a full-likelihood description of the data.

APPENDIX A: PROOF OF THEOREM 1

To show that a distribution G is invariant, we need to show that it satisfies

$$m_c = \mathbb{E}_G \frac{\beta^c s_c}{\prod_i (1 + \beta \alpha_i)} = \mu_c, \quad 0 \leq c \leq I, \quad (\text{A.1})$$

where the μ_c are nonnegative constants that we can choose, although for nontrivial proportionality in Lemma 1 we require that μ_c be strictly positive. Summing both sides of (A.1) over all values of c gives 1 throughout for any choice of G , and so at least one of these equations is redundant. We therefore divide through by m_0 and consider only the ratios m_c/m_0 ; hence we are trying to solve

$$\mathbb{E}_G \frac{\beta^c s_c}{\prod_i (1 + \beta \alpha_i)} \bigg/ \mathbb{E}_G \frac{1}{\prod_i (1 + \beta \alpha_i)} = \frac{\mu_c}{\mu_0}. \quad (\text{A.2})$$

But writing the expectations as integrals, this becomes

$$\int \frac{\beta^c s_c}{\prod_i (1 + \beta \alpha_i)} g(\beta|\alpha) d\beta \bigg/ \int \frac{1}{\prod_i (1 + \beta \alpha_i)} g(\beta|\alpha) d\beta = \frac{\mu_c}{\mu_0}, \quad (\text{A.3})$$

and writing

$$h(\beta|\alpha) = \frac{g(\beta|\alpha)}{\prod_i (1 + \beta \alpha_i)} \bigg/ \int \frac{g(\beta|\alpha)}{\prod_i (1 + \beta \alpha_i)} d\beta, \quad (\text{A.4})$$

we see that we are seeking solutions H on $\beta \geq 0$ so that

$$\mathbb{E}_H \beta^c = \frac{\mu_c}{\mu_0 s_c}, \quad 0 \leq c \leq I, \quad \forall \alpha. \quad (\text{A.5})$$

Expressing invariant distributions on this “scale” has two clear benefits. First, if an invariant H can be found with moments $\mu_c/(\mu_0 s_c)$, then it can be trivially transformed to give an invariant distribution with moments $t^c \mu_c/(\mu_0 s_c)$, for any $t \geq 0$. Second, we have shown that invariant distributions can be found as solutions of the truncated version of the classical moment problem, due to Stieltjes (1894/1895), for which a substantial literature exists. We use a recent result on the moment problem to prove the general existence of invariant distributions.

By the work of Craven and Csordas (1998), a sufficient condition for the existence of such a distribution with moments given by (A.5) is that

$$\frac{\mu_{c-1} \mu_{c+1}}{s_{c-1} s_{c+1}} \geq C_0 \frac{\mu_c^2}{s_c^2}, \quad 1 \leq c \leq I - 1, \quad (\text{A.6})$$

where C_0 is the only real root of $x^3 - 5x^2 + 4x - 1$, approximately 4.08. Now by, for example, Marcus and Minc (1964), we know that

$$s_c^2 \binom{I}{c-1} \binom{I}{c+1} \geq s_{c-1} s_{c+1} \binom{I}{c}^2, \quad (\text{A.7})$$

with equality if and only if all of the α_i ’s are equal, and so it is sufficient to show that we can choose μ_i such that

$$\frac{\mu_{c-1} \mu_{c+1}}{\mu_c^2} \geq C_0 \binom{I}{c-1} \binom{I}{c+1} \bigg/ \binom{I}{c}^2. \quad (\text{A.8})$$

Rewriting this using

$$\mu_1 = \nu_1, \quad \mu_i = \nu_i \mu_{i-1}, \quad \mu_i = \prod_1^i \nu_i, \quad (\text{A.9})$$

we get

$$\frac{\nu_c}{\nu_{c-1}} \geq C_0 \binom{I}{c-1} \binom{I}{c+1} \bigg/ \binom{I}{c}^2, \quad (\text{A.10})$$

and hence any sequence of ν_i with $\nu_i/\nu_{i-1} \geq C_0$ will do. Thus we have proved the existence of invariant mixing distributions for any I .

The condition on the ν_i can be thought of as “increasing sufficiently quickly.” If we pick $\nu_i/\nu_{i-1} = C_0$ throughout, then we get $\mu_c = \mu_0 C_0^{c(c+1)/2}$, where μ_0 becomes a normalizing constant. As noted previously, if we have a set of moments to which a distribution can be fitted, then we can multiply them by t^c to trivially find another, so picking $t = 1/\sqrt{C_0}$ here yields $\mu_c \propto C_0^{c^2/2} \approx 2c^2$.

APPENDIX B: PROOF THAT $\mathcal{B} = \mathcal{M}$ FOR $I = 2, 3$

To prove equivalence of \mathcal{B} and \mathcal{M} here and in Appendix C, we first return to (10) and note that for any I , if there exists an invariant distribution that leads to μ such that

$$\mu_c = \binom{I}{c} q^c (1-q)^{I-c}, \quad 0 \leq c \leq I, \quad (\text{B.1})$$

then putting “hyperprior” Q on q leads to the construction of \mathcal{B} as seen in (10). Therefore, to prove the equivalence of \mathcal{B} and \mathcal{M} , it will be sufficient to describe or prove the existence of a mixing distribution that leads to margins on the boundary of \mathcal{B} , as given in (B.1).

As noted in Section 3, for $I = 2$, we can interpret Rasch data as a 1 : 1 matched case-control study. Using the definitions of Section 3.1, we find that the mixing distribution where

$$\begin{aligned} p_{1j} &= q && \text{with probability } \frac{1-q+q\psi}{1+\psi}, \\ p_{2j} &= q && \text{with probability } \frac{q+(1-q)\psi}{1+\psi} \end{aligned} \quad (\text{B.2})$$

leads to margins $((1-q)^2, 2q(1-q), q^2)$, which are invariant and meet the bound of (B.1).

For the case $I = 3$, we transform the problem as in Appendix A. An invariant distribution for β that meets the bound of (B.1) can be found by solving the moment problem

$$\mathbb{E}_H \beta^c = \left(\frac{q}{1-q} \right)^c \binom{3}{c} \bigg/ s_c, \quad 0 \leq c \leq 3, \quad (\text{B.3})$$

for general values of q and α . By Appendix A, it is sufficient to prove this for $q = 1/2$, where

$$\mathbb{E}_H \beta^c = \binom{3}{c} \bigg/ s_c, \quad 0 \leq c \leq 3. \quad (\text{B.4})$$

To solve this, we use the Hankel matrix method given by, for example, Lindsay et al. (1991). The existence of a boundary meeting G for $I = 3$ can be ascertained by checking the positivity of two matrices,

$$\mathbf{R}_2 = \begin{pmatrix} 1/s_0 & 3/s_1 \\ 3/s_1 & 3/s_2 \end{pmatrix} \tag{B.5}$$

and

$$\mathbf{R}_3 = \begin{pmatrix} 3/s_1 & 3/s_2 \\ 3/s_2 & 1/s_3 \end{pmatrix}. \tag{B.6}$$

Now some algebra shows that

$$\begin{aligned} |\mathbf{R}_2| &= \frac{3}{4}((\alpha_1 - \alpha_2)^2 + (\alpha_1 - \alpha_3)^2 + (\alpha_2 - \alpha_3)^2)s_1^{-2}s_2^{-1}, \\ |\mathbf{R}_3| &= \frac{3}{4}((\alpha_1 - \alpha_2)^2\alpha_3^2 \\ &\quad + (\alpha_1 - \alpha_3)^2\alpha_2^2 + (\alpha_2 - \alpha_3)^2\alpha_1^2)s_1^{-1}s_2^{-2}s_3^{-1}, \end{aligned} \tag{B.7}$$

both of which are strictly positive unless all α_i 's are equal. In the case of strict inequality, this is sufficient to show that a boundary meeting G exists. If all α_i 's are equal, then we know that one exists trivially, and hence the equivalence is proven.

APPENDIX C: PROOF THAT $\mathcal{B} = \mathcal{M}$ FOR $1 : n$ CASE-CONTROL STUDIES

The $1 : n$ matched case-control study is a common form of study design and can be considered a very simple form of the Rasch model, where $I = n + 1$, $\alpha_1 = \psi$, and all other $\alpha_i = 1$. Hence we get

$$\begin{aligned} s_c &= \psi \binom{I-1}{c-1} + \binom{I}{c} - \binom{I-1}{c-1} \\ &= \psi \binom{I-1}{c-1} + \binom{I-1}{c}, \end{aligned} \tag{C.1}$$

and for a distribution satisfying (B.1), we must show that there exists an invariant G such that

$$\begin{aligned} \mu_c^I &= \binom{I}{c} / \left(\psi \binom{I-1}{c-1} + \binom{I-1}{c} \right) \\ &= I / (\psi c + I - c) \\ &= \frac{1}{(\psi - 1)c / I + 1}, \quad 0 \leq c \leq I. \end{aligned} \tag{C.2}$$

A proof of existence of a solution can be constructed using the Hankel matrix technique of Appendix B. The Hankel matrices take Cauchy form, and their determinants are easily calculated. Alternatively, the result may be proved directly by considering the distribution with density

$$f(x) = (\tau + 1)x^\tau, \quad 0 \leq x \leq 1, \quad \tau > -1. \tag{C.3}$$

This distribution has moments

$$\mathbb{E}(X^c) = \frac{1}{c/(\tau + 1) + 1}, \tag{C.4}$$

and so putting

$$\tau = \frac{I}{\psi - 1} - 1 \tag{C.5}$$

for $\psi > 1$, we are done (note that τ thus obeys $\tau > -1$). For $\psi < 1$, we use

$$f(x) = -(\tau + 1)x^\tau, \quad 1 \leq x \leq \infty, \quad \tau < -1 \tag{C.6}$$

with

$$\tau = -1 - \frac{I}{1 - \psi} \tag{C.7}$$

to get the same result.

APPENDIX D: PROOF THAT $\mathcal{B} \neq \mathcal{M}$ FOR $2 : 2$ CASE-CONTROL STUDIES

To prove that the equivalence between \mathcal{M} and \mathcal{B} does not hold in general, consider the (simple) Rasch model associated with a $2 : 2$ matched case-control study. Here $\alpha_1 = \alpha_2 = \psi$ and $\alpha_3 = \alpha_4 = 1$. The Hankel matrix with elements $\{1/s_0, 4/s_1, 6/s_2, 4/s_3, 1/s_4\}$ always has a negative determinant (unless $\psi = 1$), and hence we cannot find a distribution to satisfy (B.1).

This counterexample also illustrates the dependence of \mathcal{M} on the space of α being considered. Under the ‘‘strong’’ restrictions on α used in the context of $1 : 3$ matched case-control studies in Appendix C, we obtain $\mathcal{B} = \mathcal{M}$, which we know by Section 4 to be the largest space consistent with $\alpha_i = 1, i = 1, \dots, I$. The restrictions on α for a $2 : 2$ matched study or, more generally, for the usual Rasch model with $I = 4$, are thus interpretable as ‘‘weaker,’’ leading to a space of invariant marginal probabilities that is a proper subset of \mathcal{B} .

APPENDIX E: INVARIANT DISTRIBUTION FOR $I = 4$

This Appendix give details of the invariant distribution used in Section 5.1, with the (arbitrary) marginal probabilities $\mu_c = 1/5, c = 0, \dots, 4$. On the transformed H scale, this corresponds to moments $1/s_c$, where s_c is the c th symmetric polynomial in $\alpha_1, \dots, \alpha_4$. We solve this moment problem by using the simplest distribution with sufficient flexibility, namely a discrete distribution with three points of support, one of which is fixed at 0. From (A.5), we obtain a series of five equations, which we solve to find the two nonzero points of support and the three respective weights.

The points of support are

$$\begin{aligned} x_0 &= 0, \\ x_1 &= (s_2s_3(s_1s_4 - s_2s_3) + (s_2s_3(s_2^3s_3^3 + 2s_1s_3^2(2s_1s_3 - 3s_2^2)s_4) \\ &\quad + s_1s_2(4s_2^2 - 3s_1s_3)s_4^2))^{1/2} / (2s_3(s_1s_3 - s_2^2)s_4), \\ x_2 &= (s_2s_3(s_1s_4 - s_2s_3) - (s_2s_3(s_2^3s_3^3 + 2s_1s_3^2(2s_1s_3 - 3s_2^2)s_4) \\ &\quad + s_1s_2(4s_2^2 - 3s_1s_3)s_4^2))^{1/2} / (2s_3(s_1s_3 - s_2^2)s_4), \end{aligned} \tag{E.1}$$

and on the H scale, the weights assigned to each are

$$\begin{aligned} \lambda_0 &= 1 - \lambda_1 - \lambda_2, \\ \lambda_1 &= \frac{s_3 - s_4x_2}{s_3s_4x_1^4 - s_3s_4x_1^3x_2}, \\ \lambda_2 &= \frac{s_3 - s_4x_1}{s_3s_4x_2^4 - s_3s_4x_2^3x_1}. \end{aligned} \tag{E.2}$$

This distribution has moments $\mathbb{E}_H X^c = 1/s_c$ for $c = 0, \dots, 4$, and so (transformed to the G scale) gives $\mu_c = 1/5$.

Note that in the data of Stouffer and Toby (1951), the observed totals scores are not distributed evenly over their five possible values, as would be expected from this distribution. However, this lack of fit does not affect conditional inference on the α parameters, a phenomenon explored more fully in Section 4.

[Received February 2003. Revised December 2003.]

REFERENCES

Agresti, A. (1993), ‘‘Computing Conditional Maximum Likelihood Estimates for Generalized Rasch Models Using Simple Loglinear Models With Diagonal Parameters,’’ *Scandinavian Journal of Statistics*, 86, 96–107.
 Andersen, E. (1970), ‘‘Asymptotic Properties of Conditional Maximum-Likelihood Estimators,’’ *Journal of the Royal Statistical Society, Ser. B*, 32, 283–301.
 ——— (1973), ‘‘A Goodness of Fit Test for the Rasch Model,’’ *Psychometrika*, 38, 123–140.

- Brooks, H., Witt, A., and Eilts, M. (1997), "Verification of Public Weather Forecasts Available via the Media," *Bulletin of the American Meteorological Society*, 77, 2167–2177, dataset available at <http://www.nssl.noaa.gov/brooks/feda/datasets/snelli.html>.
- Christense, K., and Bjorner, J. (2003), "SAS Macros for Rasch-Based Latent Variable Modelling," Technical Report 03/13, University of Copenhagen, Dept. of Biostatistics.
- Cook, R., and Farewell, V. (1999), "The Utility of Mixed-Form Likelihoods," *Biometrics*, 55, 284–288.
- Corcoran, C., Mehta, C., Patel, N., and Senchaudhuri, P. (2001), "Computational Tools for Exact Conditional Logistic Regression," *Statistics in Medicine*, 20, 2723–2739.
- Cox, D. R. (1975), "Partial likelihood," *Biometrika*, 62, 269–276.
- Craven, T., and Csordas, G. (1998), "A Sufficient Condition for Strict Total Positivity of a Matrix," *Linear and Multilinear Algebra*, 45, 19–34.
- Cressie, N., and Holland, P. W. (1983), "Characterizing the Manifest Probabilities of Latent Trait Models," *Psychometrika*, 48, 129–141.
- Danesh, J., Youngman, L., Clark, S., Parish, S., Peto, R., and Collins, R. (1999), "Helicobacter pylori Infection and Early Onset Myocardial Infarction: Case-Control and Sibling Pairs Study," *British Medical Journal*, 319, 1157–1162.
- De Leeuw, J., and Verhelst, N. (1986), "Maximum-Likelihood Estimation in Generalized Rasch Models," *Journal of Educational Statistics*, 11, 183–196.
- Diggle, P. J., Morris, S. E., and Wakefield, J. C. (2000), "Point-Source Modelling Using Matched Case-Control Data," *Biostatistics*, 1, 89–105.
- Follman, D. (1988), "Consistent Estimation in the Rasch Model Based on Non-parametric Margins," *Psychometrika*, 53, 553–562.
- Gail, M., Lubin, J., and Rubinstein, L. (1981), "Likelihood Calculations for Matched Case-Control Studies and Survival Studies With Tied Death Times," *Biometrika*, 68, 703–707.
- Glas, C., and Verhelst, N. (1995), "Testing the Rasch Model," in *Rasch Models: Foundations, Recent Developments and Applications*, eds. G. Fischer and I. Molenaar, New York: Springer-Verlag.
- Goldstein, H., and Spiegelhalter, D. (1996), "League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance" (with discussion), *Journal of the Royal Statistical Society, Ser. A*, 159, 385–442.
- Hardy, R., and Thompson, S. (1996), "A Likelihood Approach to Meta-Analysis With Random Effects," *Statistics in Medicine*, 15, 619–629.
- Kelderman, H. (1984), "Loglinear Rasch Model Tests," *Psychometrika*, 49, 223–245.
- Lauritzen, S. L. (1988), *Extremal Families and Systems of Sufficient Statistics*, New York: Springer-Verlag.
- Liao, J. (1999), "A Hierarchical Bayesian Model for Combining Multiple 2×2 Tables Using Conditional Likelihood," *Biometrics*, 55, 268–272.
- Lindsay, B., Clogg, C., and Grego, J. (1991), "Semiparametric Estimation in the Rasch Model and Related Exponential Response Models, Including a Simple Latent Class Model for Item Analysis," *Journal of the American Statistical Association*, 86, 96–107.
- Lindsey, J. (2000), "Directly Modelling Matched Case-Control Data," *Statistics in Medicine*, 19, 35–44.
- Marcus, M., and Minc, H. (1964), *A Survey of Matrix Theory and Matrix Inequalities*, Boston: Allyn and Bacon.
- Marshall, E., and Spiegelhalter, D. (1998), "Reliability of League Tables of in vitro Fertilisation Clinics: Retrospective Analysis of Live Birth Rates," *British Medical Journal*, 316, 1701–1705.
- Mehta, C., and Patel, N. (2002), *LogXact5 for Windows. Software for Exact Logistic Regression*, Cambridge, MA: Cytel Software Corp.
- Mehta, C., Patel, N., and Senchaudhuri, P. (2000), "Efficient Monte Carlo Methods for Conditional Logistic Regression," *Journal of the American Statistical Association*, 95, 99–108.
- Molenaar, I. (1995), "Estimation of Item Parameters," in *Rasch Models: Foundations, Recent Developments and Applications*, eds. G. Fischer and I. Molenaar, New York: Springer-Verlag.
- Morris, C., and Christiansen, C. (1996), "Hierarchical Models for Ranking and for Identifying Extremes, With Applications," in *Bayesian Statistics*, Vol. 5, eds. J. Bernardo, J. Berger, A. Dawid, and A. Smith, Oxford, U.K.: Oxford University Press, pp. 277–296.
- Rasch, G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen: Danmarks Paedagogiske Institut, expanded edition (1980), Chicago: University of Chicago Press.
- Rice, K. (2003), "Full-Likelihood Techniques for Misclassification of Exposure in Matched Case Control Studies," *Statistics in Medicine*, 22, 3177–3194.
- Rosenbaum, P. R. (1984), "Testing the Conditional-Independence and Monotonicity Assumptions of Item Response Theory," *Psychometrika*, 49, 425–435.
- Self, G., and Liang, K.-Y. (1987), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," *Journal of the American Statistical Association*, 82, 605–610.
- Severini, T. (1999), "On the Relationship Between Bayesian and Non-Bayesian Elimination of Nuisance Parameters," *Statistica Sinica*, 9, 713–724.
- Spiegelhalter, D., Abrams, K., and Myles, J. (2003), *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Chichester, U.K.: Wiley.
- Spiegelhalter, D., Thomas, A., and Best, N. (2003), *WinBUGS Version 1.4 User Manual*, MRC Biostatistics Unit.
- Sprott, D. (1975), "Marginal and Conditional Sufficiency," *Biometrika*, 62, 599–606.
- Stieltjes, T. (1894/1895), "Recherches sur les fractions continues," *Annales de la Faculté de Sciences de Toulouse*, 8, J1–122; 9, A1–47 (in French).
- Stouffer, S., and Toby, J. (1951), "Role Conflict and Personality," *American Journal of Sociology*, 56, 395–406.
- Strawderman, R., and Wells, M. (1998), "Approximately Exact Inference for the Common Odds Ratio in Several 2×2 Tables," *Journal of the American Statistical Association*, 93, 1294–1307.
- TenVergert, E., Gillespie, M., and Kingma, J. (1993), "Testing the Assumptions and Interpreting the Results of the Rasch Model Using Log-Linear Procedures in SPSS," *Behaviour Research Methods Instruments & Computers*, 25, 250–259.
- Tjur, T. (1982), "A Connection Between Rasch's Item Analysis Model and a Multiplicative Poisson Model," *Scandinavian Journal of Statistics*, 9, 23–30.
- Wood, G. (1992), "Binomial Mixtures and Finite Exchangeability," *The Annals of Probability*, 20, 1167–1173.
- (1999), "Binomial Mixtures: Geometric Estimation of the Mixing Distribution," *The Annals of Statistics*, 27, 1706–1721.