A BAYESIAN RANDOM EFFECTS MODEL FOR TESTLETS

ERIC T. BRADLOW
MARKETING AND STATISTICS
THE WHARTON SCHOOL
THE UNIVERSITY OF PENNSYLVANIA

HOWARD WAINER AND XIAOHUI WANG
EDUCATIONAL TESTING SERVICE

Standard item response theory (IRT) models fit to dichotomous examination responses ignore the fact that sets of items (testlets) often come from a single common stimuli (e.g., a reading comprehension passage). In this setting, all items given to an examinee are unlikely to be conditionally independent (given examinee proficiency). Models that assume conditional independence will overestimate the precision with which examinee proficiency is measured. Overstatement of precision may lead to inaccurate inferences such as prematurely ending an examination in which the stopping rule is based on the estimated standard error of examinee proficiency (e.g., an adaptive test). To model examinations that may be a mixture of independent items and testlets, we modified one standard IRT model to include an additional random effect for items nested within the same testlet. We use a Bayesian framework to facilitate posterior inference via a Data Augmented Gibbs Sampler (DAGS; Tanner & Wong, 1987). The modified and standard IRT models are both applied to a data set from a disclosed form of the SAT. We also provide simulation results that indicates that the degree of precision bias is a function of the variability of the testlet effects, as well as the testlet design.

Key words: Gibbs sampler, Data augmentation, Testlets

1. Introduction

Most current standardized educational tests have sections comprised of coherent groups of multiple choice items based on a common stimulus. Such groupings, usually called testlets (Wainer & Kiely, 1987), are used for many reasons, principal among them are time and cost constraints. The basic idea is that the reading and understanding of the stimulus by the examinee require a considerable amount of mental processing and hence time. Therefore by pairing such a stimulus with more than a single test item the marginal time for gathering additional information is reduced. Examples of testlets can be found in tests of verbal proficiency (e.g., the stimulus might be a reading passage), mathematics (e.g., the stimulus might be a data display), and analytical reasoning (e.g., the stimulus might be a Venn diagram).

Although it is well-known that testlet items' reliance on a common stimulus, yields dependence among the item responses within an individual, this dependence is often ignored. That is, given an examinee's proficiency the responses to items are assumed to be independent. This incorrect assumption is made to facilitate the straightforward analysis of multiple choice responses. When multiple choice items are composed into testlets, the assumption of conditional independence (CI) becomes more tenuous because of subject matter expertise, misdiagramming of the stimulus, fatigue, etc. (Yen, 1993). The effect of the violation of this assumption is that items nested within the same testlet will have a dependence structure that is not exhibited across items.

The authors wish to thank Robert Mislevy, Andrew Gelman and Donald B. Rubin for their helpful suggestions and comments, Ida Lawrence and Miriam Feigenbaum for providing us with the SAT data analyzed in section 5, and to the two anonymous referees for their careful reading and thoughtful suggestions on an earlier draft. We are also grateful to the Educational Testing Service for providing the resources to do this research.

Requests for reprints should be sent to Eric T. Bradlow, Marketing Department, The Wharton School, University of Pennsylvania, 1400 Steinberg Hall-Dietrich Hall, Philadelphia PA 19104-6371.
from different stimuli. It has been repeatedly shown (Sireci, Wainer & Thissen, 1991; Wainer, 1995; Wainer & Thissen, 1996) that this dependence structure, if ignored by applying standard IRT models with a CI assumption, will result in an overstatement of precision of proficiency estimates as well as a bias in item difficulty and discrimination parameter estimates. To address this additional structure, we propose a parametric approach which involves a modification to standard IRT models that explicitly accounts for the nesting of items within the same testlets and can be applied more generally to multiple choice sections that are comprised of a mixture of independent items and testlets.

A significant amount of research from nonparametric and factor analytic perspectives (e.g., Hulin, Drasgow, & Parsons 1983; McDonald 1981, 1982; Stout et al., 1996; Stout, 1987, 1990; Zhang, 1996; and Zhang & Stout, 1999) has been written on detecting conditional independence among items, estimating the degree of that dependence, and/or assessing the number of latent factors. Our parametric approach is complimentary to this work in that our extended model allows for detection, explicit modeling, and assessing the degree of item conditional dependence (of a specific form), as does the previous work, but from a Bayesian modeling perspective.

The remainder of this paper is laid out as follows. In Section 2 we describe the general data structure, the modified IRT model to account for testlets, and the extension of the model into a larger Bayesian hierarchical framework. The Bayesian hierarchical model, although producing valid posterior inferences, leads to computational issues that are described in Section 3. Simulation results for seven test designs illustrate the effect of the proposed parameterization and are presented in Section 4. In Section 5, we apply the standard IRT and modified IRT models to a real data set obtained from the SAT. Conclusions and extensions are discussed in section 6.

2. The Data Structure and Model

This research considers a simplified data structure in which each of I examinees receives a linear test of J multiple choice items scored in a binary fashion. The I x J-dimensional response matrix Y = (yij) is assumed to be fully observed. The J items are grouped into K (1 ≤ K ≤ J) mutually exclusive and exhaustive testlets. We denote the testlet of item j by d(j), the size of each testlet by nk (1 ≤ k ≤ K), with d(1) = 1 and d(J) = K. A testlet with nk = 1 is an independent item, and similarly, if K = J, then each item is in its own testlet. Models for more complex data structures that allow for missing data, adaptive tests, and polytomous responses are areas for future consideration and are not addressed here.

We consider a probit version of the two-parameter logistic (2-PL) model (Lord & Novick, 1968) as the baseline model for this research. For computational advantage (described in section 3), we define the 2-PL model in its natural latent variable form, with latent score tij given by

\[ t_{ij} = a_j (\theta_i - b_j) + \epsilon_{ij} \]  

where

\[ y_{ij} = \begin{cases} 1 & \text{if } t_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} \]

and \( \epsilon_{ij} \) is a unit normal variate used to indicate the randomness in response \( y_{ij} \) across hypothetical replications of item \( j \) to examinee \( i \). The parameters \( \theta_i, a_j, \) and \( b_j \) are given standard interpretations as the examinee proficiency, item discrimination, and item difficulty respectively.

We complete the specification of the model given in (1) into a larger Bayesian hierarchical framework by asserting Gaussian population distributions for the ability and item parameters treating them as random effects. Assuming Gaussian distributions is done out of convention, and for computational convenience as described in Section 3. Treating the unknown parameters as random effects is consistent with current testing practice, and allows for sharing of information across persons and items in which commonalities are likely to occur. Specifically we assume