Bayesian Treatment of Nonstandard Problems in Test Analysis

Rajitha M. Silva, Yuping Guan and Tim B. Swartz [*]

Abstract

This paper extends the methods of Silva, Guan and Swartz (2017) in an attempt to handle nonstandard problems in test analysis. The approach is based on a Bayesian framework where test characteristics are treated as random parameters for which posterior probability assessments are available. The generality of the approach permits straightforward analyses of problems that may be difficult using standard classical test theory and standard item response theory. We first illustrate the methods on aviation test scores where the test outcomes are not dichotomous (i.e. correct and incorrect responses). Instead, the approach is modified to handle questions with answers on a five-point ordinal scale. The second problem addresses the complication of the assessment of instructors in addition to the assessment of test questions and students.

Keywords: Empirical Bayes, Markov chain Monte Carlo, JAGS programming language.

[*] Rajitha Silva is Senior Lecturer, Department of Statistics, University of Sri Jayewardenepura, Sri Lanka (rsilva@sjp.ac.lk), Yuping Guan is President of Astrom Aviation Big Data Inc., 2270 - 21331 Gordon Way, Richmond BC, Canada V6W1J9 (ypguan@AstromAviation.com), and Tim Swartz is Professor, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6 (tim@stat.sfu.ca). Swartz has been partially supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors thank the Editor and an anonymous reviewer for their contributions to an earlier version of the manuscript.

1 INTRODUCTION

The analysis of tests and questionnaires has an extensive literature which addresses problems across a spectrum of disciplines including educational testing, customer questionnaires, opinion polls and social science surveys.

Historically, the two main approaches in test analysis have been *classical test theory* (CTT) and *item response theory* (IRT). Both approaches are used in practice and numerous comparative studies have been undertaken (e.g. Hambleton and Jones 1993, Fan 1998, Guler, Uyanik and Teker 2014, Kohli, Koran and Henn 2015, Raykov and Marcoulides 2016). As research developments have progressed, the distinction between classical test theory and item response theory has narrowed. However, in a very brief and perhaps oversimplified comparison of the two approaches, CTT is the original testing framework which investigates the results of test questions on a specific sample of respondents and has few (if any) modeling assumptions. One of the appealing aspects of CTT is that the corresponding statistics are relatively simple and guidelines have been introduced for the assessment of these statistics. In the IRT framework, statistical models form the backbone of the approach. The statistical models involve parameters that distinguish particular aspects of both test questions and respondents. IRT relies on statistical theory and is less accessible to some practioners. IRT has grown in many directions where various models have been proposed. Notably, Bayesian implementations of IRT now exist (Fox 2010, Levy and Mislevy 2016), and these require another level of statistical sophistication on the part of the practitioner.

Silva, Guan and Swartz (2017) attempted to incorporate some of the best features of both CTT and IRT in their approach to test analysis. Specifically, they attempted to retain some of the simplicity of CTT which is appealing to practitioners. Also, like IRT, they developed an approach which has inferential capability. In this paper, we demonstrate how the approach of Silva, Guan and Swartz (2017) can be easily extended to address nonstandard problems in test analysis. The use of the JAGS programming language (Plummer 2015) facilitates extensions to these more complex testing scenarios.

In Section 2, we review the approach of Silva, Guan and Swartz (2017) in the context of dichotomous responses arising from test questions. In particular, we point out some advantages of the testing framework. In Section 3, we generate a dataset intended to mimic data from the aviation industry. In this case, the responses are no longer dichotomous but lie on a five-point ordinal scale. We illustrate an analysis of this dataset based on small modifications to the basic model. In Section 4, a dataset is considered which also contains the identification of the instructor.

Here, a nonstandard inferential question involves the impact of the quality of the instructor on the test results. Again, we illustrate an analysis of this dataset based on small modifications to the basic model. Finally, a short discussion and concluding remarks are provided in Section 5.

2 THE BASIC BAYESIAN MODEL

We consider test data consisting of an $n \times k$ matrix $X = (x_{ij})$ where the n rows correspond to respondents and the k columns refer to test questions. The data are dichotomous (binary) where $x_{ij} = 1(0)$ specifies that the ith respondent provided a correct (incorrect) answer to the jth question. Therefore, the setup is applicable in various scenarios including true/false questions and multiple choice questions.

Our approach is based on a simple Bernoulli model where $x_{ij} \sim \text{Bernoulli}(\theta_{ij})$. The model stipulates that the probablity of a correct answer by the ith respondent to the jth question is $\text{Prob}(x_{ij} = 1) = \theta_{ij}$ which leads to the joint probability mass function

$$f(x \mid \theta) = \prod_{i=1}^{n} \prod_{j=1}^{k} \theta_{ij}^{x_{ij}} \; . \tag{1}$$

An immediate reaction to (1) may be that the model is problematic since there are as many parameters nk as there are data values. However, in a Bayesian approach, prior information concerning the parameters is often available and parameters may "borrow" from one another such that the effective parametrization is reduced. For example, two parameters may have a common prior distribution; learning about one of the parameters is effectively learning about both parameters.

What makes our Bayesian approach different from CTT is that the focus changes from calculations performed on the data x_{ij} to calculations performed on the parameters θ_{ij}. We suggest that the θ_{ij} are intrinsically of more interest than the observed x_{ij}. For example, there is typically more interest in the quantity $\theta_{i\cdot} = \sum_{j=1}^{k} \theta_{ij}$ which is an ability measure of the ith subject rather than the subject's one-off test score $x_{i\cdot} = \sum_{j=1}^{k} x_{ij}$. Similarly, $\bar{\theta}_{\cdot j} = \sum_{i=1}^{n} \theta_{ij}/n$ describes the difficulty of test question j where the formulation acknowledges the randomness of test questions. One may think of test question j as arising from a population of questions. The idea of focusing on population quantities (i.e. parameters) rather than statistics (i.e. data) has been previously explored; see for example Swartz (2011) in the context of clustering.

And herein lies a second possible reaction - the θ_{ij}'s are unknown. How can one learn about the parameter matrix $\theta = (\theta_{ij})$, especially in cases where the posterior distribution is complex and high-dimensional? Under a simulation-based Bayesian approach, θ's are generated from the posterior distribution, from which posterior quantities of interest can be calculated. With samples generated from the posterior, variability of the posterior quantities can also be determined. For example, we have mentioned that the quantity $\theta_{i\cdot} = \sum_{j=1}^{k} \theta_{ij}$ is an ability measure of the ith student and its posterior mean may be a quantity of interest. If we denote $\theta_{ij}^{(l)}$ as the lth realization of θ_{ij} generated from the posterior distribution, $l = 1, \ldots, M$, then the posterior mean $\mathrm{E}(\theta_{i\cdot} \mid x)$ can be estimated by $\hat{\theta}_{i\cdot} = \sum_{l=1}^{M} \sum_{j=1}^{k} \theta_{ij}^{(l)}/M$. We may also obtain the corresponding standard error of $\hat{\theta}_{i\cdot}$ by $SD = [\ \sum_{l=1}^{M} (\sum_{j=1}^{k} \theta_{ij}^{(l)})^2/M - \hat{\theta}_{i\cdot}^2\ ]^{1/2}$.

The only additional ingredient that is required for the Bayesian implementation is the specification of a prior distribution on the parameters. A prior density $\pi(\theta)$ is combined with the Bernoulli sampling model to yield the posterior density

$$\begin{aligned} \pi(\theta \mid x) &\propto f(x \mid \theta)\ \pi(\theta) \\ &\propto \prod_{i=1}^{n} \prod_{j=1}^{k} \theta_{ij}^{x_{ij}}\ \pi(\theta)\ . \end{aligned} \tag{2}$$

Although prior distributions may be subjective and take into account expert knowledge, we utilize empirical Bayes priors in the applications presented in Section 3 and Section 4.

One of the added advantages of a Bayesian approach is the elegance and ease with which missing data can be handled. For example, there are exams where test questions are randomly generated from a databank for each student or subsets of students. In these situations, individual students answer only some of the questions. In this sense, there is missing data. We therefore distinguish between the observed data x_{obs} and the missing data x_{mis}. Letting f denote generic functions, the relevant posterior distribution in this case is given by the posterior density

$$\begin{aligned} \pi(\theta, x_{\mathrm{mis}} \mid x_{\mathrm{obs}}) &\propto f(\theta, x_{\mathrm{mis}}, x_{\mathrm{obs}}) \\ &= f(x_{\mathrm{mis}}, x_{\mathrm{obs}} \mid \theta)\ \pi(\theta)\ . \end{aligned} \tag{3}$$

The key observation from (3) is that $f(x_{\mathrm{obs}}, x_{\mathrm{mis}} \mid \theta)\ \pi(\theta)$ is the unnormalized posterior density that one would obtain if x_{mis} were actually observed. Therefore, one simulates as before except that x_{mis} takes the role of a random parameter rather than a fixed data value.

Above, we have alluded to simulation-based Bayesian software. For this purpose, we use the JAGS programming language which is relatively simple and avoids the need of special purpose

Markov chain Monte Carlo code. JAGS is open source software (www.mcmc-jags.sourceforge.net) which is similar to WinBUGS. To handle missing data in JAGS, we need only code the unobserved data values with the NA symbol. Details on WinBUGS and an introduction to the Bayesian approach are given by Lunn et al. (2013).

3 EXTENSION 1: ORDINAL OUTCOMES

We now consider a generated dataset intended to mimic data arising from the aviation industry. The benefit of a generated dataset is that we know the "truth" (i.e. the underlying parameters) and can assess whether the proposed methods provide accurate estimates. Here, $n = 80$ pilots are evaluated on $k = 20$ tasks from a flight simulator. In this case, the data consists of an $n \times k$ matrix $X = (x_{ij})$ where the outcomes are no longer dichotomous (0/1) as in Section 2. Rather, the outcomes are measured on a five-point scale where $1 \equiv$ very poor, $2 \equiv$ not meeting expectations, $3 \equiv$ normal, $4 \equiv$ exceeding expectations and $5 \equiv$ outstanding.

In this setting, the first challenge is to assign a sampling distribution to the data. We choose

$$x_{ij} \sim 1 + \text{Binomial}(4, \theta_{ij}) \ . \tag{4}$$

Technically, the specification in (4) is not ideal as the Binomial distribution is appropriate for *ratio* data and our data are *ordinal*. However, the Binomial distribution is appealing since it is based on a single unknown parameter θ_{ij} which describes the difficulty of the jth question for the ith pilot. Furthemore, the distribution in (4) has the correct range, $x_{ij} = 1, 2, 3, 4, 5$ and its probability mass function exhibits concavity with the presence of an interior mode provided that $\theta_{ij} \neq 0, 1$. With test scores, we would naturally expect decreasing probabilities as we move away from the mode towards more extreme scores in the tails. In the sampling model (4), θ_{ij} represents a pilot/task characteristic. For $\theta_{i_1,j} > \theta_{i_2,j}$, pilot i_1 has an expected stronger performance than pilot i_2 on task j. Similarly, for $\theta_{i,j_1} > \theta_{i,j_2}$, it is expected that task j_1 is easier than task j_2 for pilot i. Following (2), this leads to the posterior density

$$\pi(\theta \mid x) \propto \prod_{i=1}^{n} \prod_{j=1}^{k} \ \theta_{ij}^{x_{ij}-1} (1 - \theta_{ij})^{5-x_{ij}} \ \pi(\theta) \ . \tag{5}$$

The next challenge is the specification of the prior density $\pi(\theta)$ in (5). Our suggested prior density has the following structure

$$\pi(\theta) \;\; = \;\; \prod_{i,j} \pi(\theta_{ij})$$

where the θ_{ij} are conditionally independent with

$$\theta_{ij} \;\; \sim \;\; \text{truncated-Normal}(\mu_{ij}, \sigma^2_{ij}) \; . \tag{6}$$

In (6), there are various potential specifications for the hyperparameters μ_{ij} and σ_{ij}. A simple choice is $\mu_{ij} = k_1$ (say $k_1 = 0.5$), and $\sigma_{ij} = k_2$ (large, say $k_2 = 20$). This nearly corresponds to a flat prior over the interval $(0, 1)$. However, such a prior does not take into account the physical knowledge associated with test analysis. For example, some questions are typically more difficult than other questions and we would like to take this prior knowledge into account. A drawback of a flat prior (i.e. Uniform(0,1)) is that the posterior mean of θ_{ij} is pulled towards the prior mean 0.5 to the extent that our inferences depart markedly from CTT inferences. This is illustrated in Silva, Guan and Swartz (2017).

Another possibility for the specification of the hyperparameters μ_{ij} and σ_{ij} in (6) based on an empirical Bayes (EB) procedure. One EB procedure requires fitting (4) with a generalized linear model (glm) where we introduce the parametrization

$$\text{logit}(\theta_{ij} \mid \beta_0, \alpha_i, \gamma_j) = \beta_0 + \alpha_i + \gamma_j \; . \tag{7}$$

The logistic relationship (7) states that good pilots (i.e. pilots i where α_i is large) tend to perform better across all tasks and that simple tasks (i.e. tasks j where γ_j is large) tend to be easier across all pilots. A difficulty with this parametrization is that the prior too closely resembles the likelihood, and we do not observe the desired shrinkage in parameter estimates which is discussed later in the following example.

Our preferred choice for the prior specification of μ_{ij} in (6) is similar to (7) but is based on fitting the simpler glm

$$\text{logit}(\theta_{ij} \mid \beta_0, \gamma_j) = \beta_0 + \gamma_j \tag{8}$$

which states that simple tasks (i.e. tasks j where γ_j is large) tend to be easier across all pilots. In addition to enabling shrinkage, an additional advantage of (8) over (7) is that we may fit (8) using the same test questions from other pilots to obtain our prior. In that case, we have a "true" prior rather than an EB procedure. Using (8), the fitted glm provides parameter estimates $\hat{\beta}_0$ and $\hat{\gamma}_j$. Since μ_{ij} denotes the mean $\mathrm{E}(\theta_{ij})$, we therefore invert the logistic function (8) and set

$$\mu_{ij} = \frac{\exp(\hat{\beta}_0 + \hat{\gamma}_j)}{1 + \exp(\hat{\beta}_0 + \hat{\gamma}_j)} \ .$$

We treat $\sigma_{ij} = k$ as a tuning parameter where larger (smaller) values of k provide relatively less (more) weight on the prior structure and more (less) weight on the data. In this example, we choose $\sigma_{ij} = 0.2$ which allows both likelihood and prior to impact the posterior. The calculation of the estimates μ_{ij} are immediately available from the *predict* function which can be used on a *glm* object in R.

Following the model development described above, the data generation procedure begins by setting the underlying parameters in (7) according to $\beta_0 = 0$, $\alpha_i = (i-1)/n$ and $\gamma_j = (j-1)/k$ for $i = 1, \ldots, n$ and $j = 1, \ldots, k$. With these prescribed values, the θ_{ij} terms are determined via (7) and we generate data x_{ij} according to (4). Under these settings, the test scores x_{ij} tend to increase with increasing i and increasing j. The values of β_0, α_i and γ_j were chosen so that pilots tend to have high test scores as is typically the case in practice. For example, with our dataset, pilot #1 scored 69 out of 100 and pilot $n = 80$ scored 88 out of 100.

The first thing that we wish to check is the estimation procedure. In Figure 1, we plot the posterior means of the test scores of pilots $\mathrm{E}(T_i \mid x) = \sum_{j=1}^{k}(1 + 4\mathrm{E}(\theta_{ij} \mid x))$ verus the associated quantity $\mathrm{E}(T_i \mid \theta_i) = \sum_{j=1}^{k}(1+4\theta_{ij})$ where the θ_{ij} are the true underlying values obtained from (7) based on the specified settings of β_0, α_i and γ_j in the data generating mechanism. The agreement in the pairs suggest that the model and the associated Markov chain procedure are producing sensible results. The dispersion of points about the line $y = x$ describes the natural variation based on randomly generated data. We also see that the pairs of points are generally increasing with i and this is in keeping with the fixed settings of α_i.

To get a sense of the utility of the approach in this nonstandard problem, we investigate the average performance of pilots by looking at posterior quantities associated with $\theta_{i\cdot} = \sum_{j=1}^{k} \theta_{ij}$. The quantity k+4$\theta_{i\cdot}$ can be interpreted as the expected total score on the test (consisting of k questions) which we would expect to be attained by the ith pilot. In Figure 2, we plot 90% posterior bands of this quantity for the $n = 80$ pilots in the study. The posterior intervals were estimated from $M = 10,000$ runs of a Markov chain obtained from JAGS where the tuning parameter in (6) was again set at $\sigma_{ij} = 0.2$. The posterior intervals in Figure 2 based on the

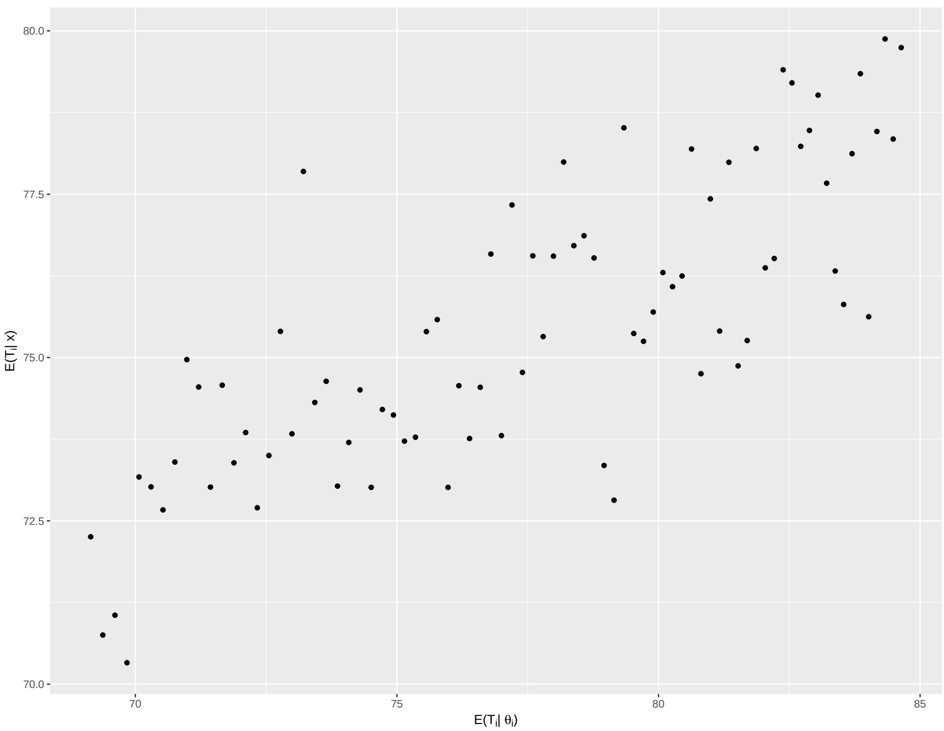

Figure 1: Based on the Markov chain output, the posterior means of total pilot scores $\mathrm{E}(T_i \mid x)$ are plotted against their expected underlying values $\mathrm{E}(T_i \mid \theta_i)$.

empirical Bayes procedure are appealing as they allow us to differentiate between pilots. Specifically, we observe performance differences between pilot pairs corresponding to small and large i. We also observe a shrinkage effect where expected test scores are smaller (larger) for those pilots who attained high (low) test scores. As an example of a pilot with a low test score, pilot 22 had an actual test score of 71/100 and an expected test score of $k + 4\theta_{22\cdot} = 73.12$. As an example of a pilot with a high test score, pilot 49 had an actual test score of 85/100 and an expected test score of $k + 4\theta_{49\cdot} = 78.51$. Doing exceedingly well (poorly) on a test may be perceived as being partially lucky (unlucky). This is a standard regression to the mean effect. Note that by increasing (decreasing) the tuning parameter σ_{ij} we can decrease (increase) the shrinkage effect. To illustrate the effect of the tuning parameter $\sigma_{ij} = k$, we repeat the above calculations for pilot 49. With $\sigma_{ij} = 0.22$, the expected test score $k + 4\theta_{49\cdot}$ is 78.46. With $\sigma_{ij} = 0.18$, the expected test score $k + 4\theta_{49\cdot}$ is 78.64.

We next investigate the comparative difficulty of the $k = 20$ tasks undertaken by the pilots as described by $\bar{\theta}_{\cdot j} = \sum_{i=1}^{n} \theta_{ij}/n$, $j = 1, \ldots, 20$. In Figure 3, we plot the associated 90% posterior bands. We observe that there is a general decrease in the level of difficulty of test questions (i.e.

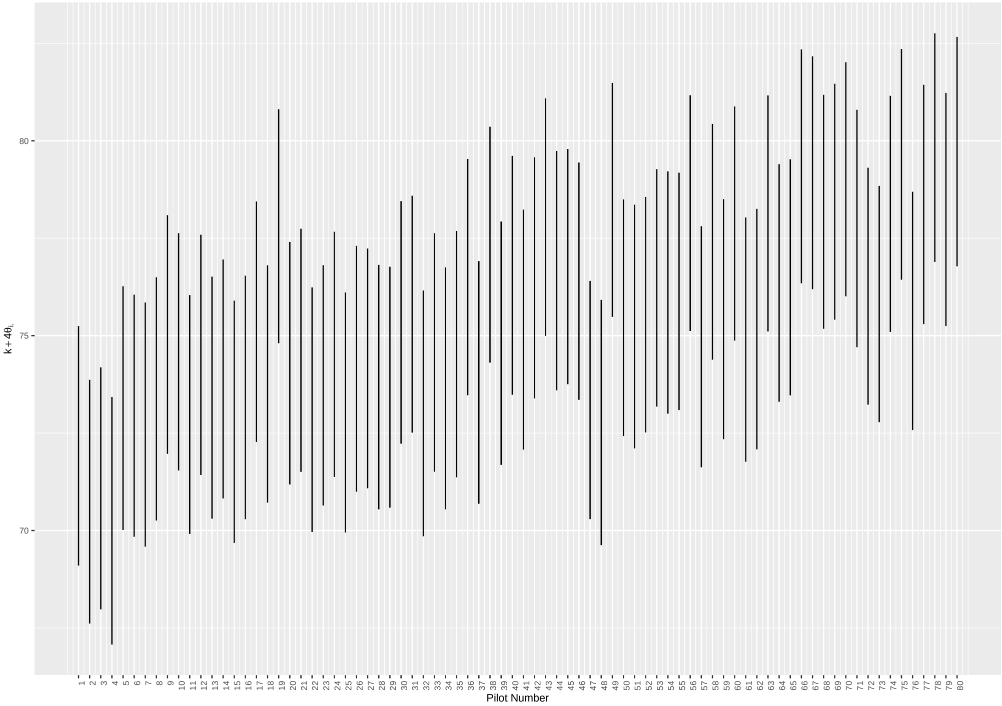

Figure 2: Posterior intervals (90%) of the expected posterior test scores $k + 4\theta_{i\cdot}$ for the $n = 80$ pilots.

increasing $\bar{\theta}_{\cdot j}$). This is in accordance with the data generating mechanism.

We may also be interested in the redundancy of tasks. The problem of survey fatigue is well-known in the survey literature where it is not desirable to have two questions that address essentially the same problem for which the responses are similar. We therefore consider the correlation parameter

$$r_{j_1,j_2} = \frac{\sum_{i=1}^{n}(\theta_{i,j_1} - \bar{\theta}_{\cdot j_1})(\theta_{i,j_2} - \bar{\theta}_{\cdot j_2})}{\sqrt{\sum_{i=1}^{n}(\theta_{i,j_1} - \bar{\theta}_{\cdot j_1})^2}\sqrt{\sum_{i=1}^{n}(\theta_{i,j_2} - \bar{\theta}_{\cdot j_2})^2}}$$

which compares the similarity of tasks j_1 and j_2. With $k = 20$ tasks, there are $\binom{k}{2} = 190$ correlations of interest. In each iteration of the Markov chain, r_{j_1,j_2} can be calculated and these values are averaged over the M iterations to provide a posterior mean of r_{j_1,j_2}. In Figure 4, we provide a scatterplot corresponding to the posterior output of $\theta_{i,1}$ and $\theta_{i,2}$ based on a single Monte Carlo iteration after burn-in. The plot consisting of $n = 80$ points examines the comparative difficulty of tasks #1 and #2. Although these two questions (tasks #1 and #2) are similar in difficulty, we recall that the data were generated independently with respect to pilots and

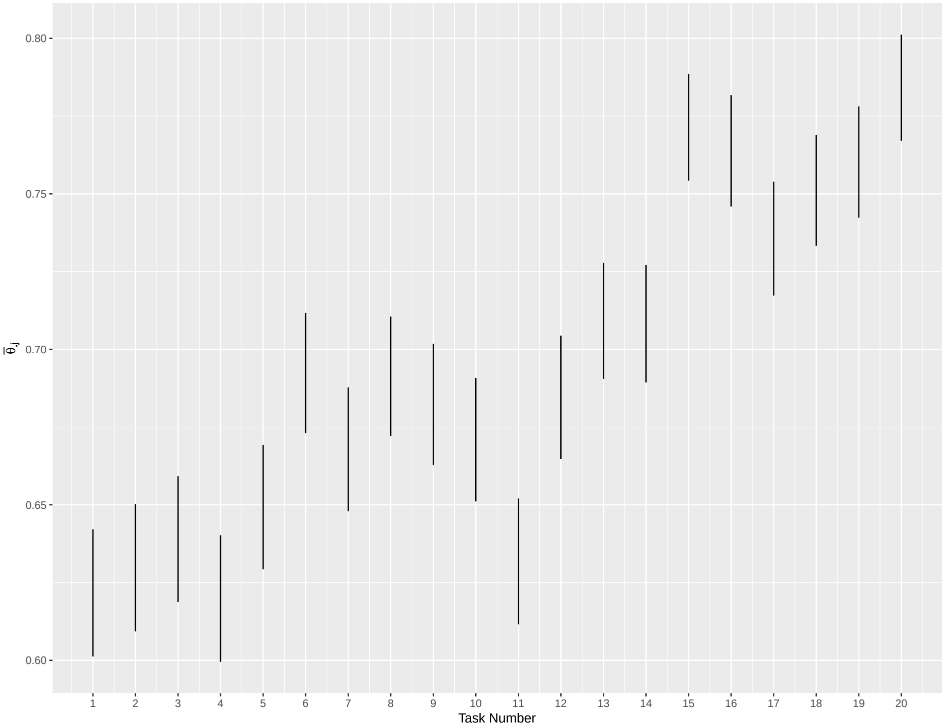

Figure 3: Posterior intervals (90%) for the task difficulty parameter $\bar{\theta}_{\cdot j}$ for the $j = 1, \ldots, 20$ tasks in the study.

tasks. Therefore, we should not expect a correlation between these two tasks, and we observe the expected posterior mean $\mathrm{E}(r_{1,2} \mid x) = -0.025$ which is indeed small. With real data, if the correlation between two tasks is strong, this suggests that the two tasks are redundant, and in the interest of efficiency and survey-fatigue, perhaps one of the tasks could be removed from the test.

4 EXTENSION 2: ASSESSMENT OF INSTRUCTORS

In this nonstandard application in test analysis, we are interested in the assessment of instructors. For example, we may have L instructors who are each responsible for a cohort of students. In this case, every observation x_{ij} has an added subscript such that x_{ijl} denotes the score by the ith student on the jth question, and this student received instruction on this question by instructor l. We similarly extend the notation for the parameters leading to terms θ_{ijl}. In the assessment of instructors, it would be difficult to conclude a causal relationship due to instructors if the students were not randomized amongst instructors. For example, there could be an underlying confounding

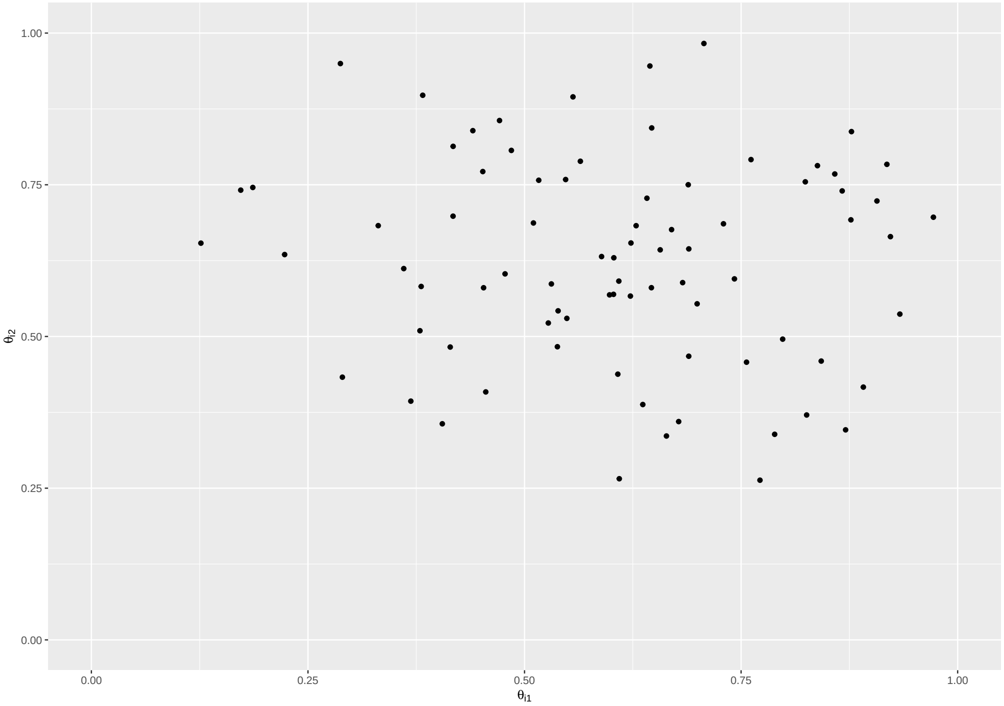

Figure 4: Scatterplot of the posterior output of $\theta_{i,1}$ and $\theta_{i,2}$ based on a single iteration from the Markov chain.

variable that is associated with instructor differences. The above setup is also applicable to other situations. For example, a comparison of different groups of students may be of interest.

We consider a complex design involving $n = 12$ students, $k = 11$ test questions on a five-point ordinal scale (1-5) and $L = 12$ instructors. As in Section 3, the questions are tasks so that students may have more than one response to a specific task with each response associated with a specific instructor. The experimental design is summarized in Table 1. The discussion in Section 2 on missing data is particularly relevant in this example as we have many cases of missing data due to students not having received instruction by a given instructor (e.g. $x_{1,j,3}$ for all $j = 1, \ldots, k$).

The statistical model which we consider is similar to Section 3 where we now define

$$x_{ijl} \sim 1 + \text{Binomial}(4, \theta_{ijl}) \tag{9}$$

which leads to the posterior density

$$\pi(\theta \mid x) \propto \prod_{i=1}^{n} \prod_{j=1}^{k} \prod_{l=1}^{L} \theta_{ijl}^{x_{ijl}-1} (1 - \theta_{ijl})^{5-x_{ijl}} \; \pi(\theta) \tag{10}$$

where θ is the parameter space. Also, following Section 3, we specify the prior density $\pi(\theta)$ in

	Student											
	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}
Instructor_1	1	1	1	1	1	0	0	0	0	0	0	0
Instructor_2	1	1	0	0	0	0	0	0	1	1	1	1
Instructor_3	0	0	0	0	0	1	1	1	1	1	1	1
Instructor_4	1	0	0	0	0	0	1	1	1	1	1	1
Instructor_5	1	1	1	0	0	0	0	0	1	1	1	1
Instructor_6	1	1	1	1	0	0	0	0	1	1	0	0
Instructor_7	1	0	0	0	1	1	1	0	0	0	0	1
Instructor_8	1	1	0	0	0	1	1	0	0	0	0	0
Instructor_9	0	0	0	0	0	0	0	0	1	1	1	1
Instructor_{10}	1	1	0	0	0	0	0	0	0	0	1	1
Instructor_{11}	1	1	0	0	0	0	0	1	1	0	0	1
Instructor_{12}	1	1	0	0	0	0	1	1	1	0	0	0

Table 1: An entry of 1(0) in the (i,j)th cell indicates that Instructor i provided (did not provide) instruction to student j on all $k = 11$ tasks.

(10) according to

$$\pi(\theta) \;=\; \prod_{i,j,l} \pi(\theta_{ijl})$$

where the θ_{ijl} are conditionally independent with

$$\theta_{ijl} \;\sim\; \text{truncated-Normal}(\mu_{ijl}, \sigma^2_{ijl}) \; . \tag{11}$$

In (11), the truncation again corresponds to the interval $(0, 1)$ and the parameters μ_{ijl} and σ_{ijl} are specified according to an EB procedure. The EB procedure first requires fitting (9) with a glm where we define

$$\text{logit}(\theta_{ijl} \mid \beta_0, \gamma_j, \lambda_l) = \beta_0 + \gamma_j + \lambda_l \; . \tag{12}$$

The difference between (12) and (8) involves the inclusion of the λ_l term in (12) which accounts for the instructor effect. Following the development in Section 3, the estimates $\hat{\beta}_0$, $\hat{\gamma}_j$ and $\hat{\lambda}_l$ are used to specify μ_{ijl} in (11). We again treat $\sigma_{ijl} = k$ as a tuning parameter.

Our interest in instructors may be expressed by the rating index $\bar{\theta}_{\cdot\cdot l} = \frac{1}{n}\frac{1}{k}\sum_{i=1}^{n}\sum_{j=1}^{k}\theta_{ijl}$ for instructors $l = 1, \ldots, L$ which can be interpreted as the average probability of a correct answer based on instruction from instructor l. Posterior realizations of the parameters θ_{ijl} are generated as before, and these are used to obtain more complex quantities of interest such as $\bar{\theta}_{\cdot\cdot l}$.

Using our data, in Figure 5, we plot 90% confidence bands for the instructor rating index $\bar{\theta}_{\cdot\cdot l}$ for the $L = 12$ instructors. According to the plot, we observe that Instructor 1 is the "best" instructor and that Instructor 9 is the "worst" Instructor. These inferences are consistent with the observed data where Instructor 1 had 5 students whose observed average score was 39.8/55, and Instructor 9 had 4 students whose observed average score was 30.0/55. The posterior mean scores corresponding to these instructors $(1/n)\sum_{i=1}^{n}\sum_{j=1}^{k}(4\theta_{ijl}+1)$ were 38.68 ($l = 1$) and 30.45 ($l = 9$) which demonstrates a desirable shrinkage effect from observed scores to posterior mean scores.

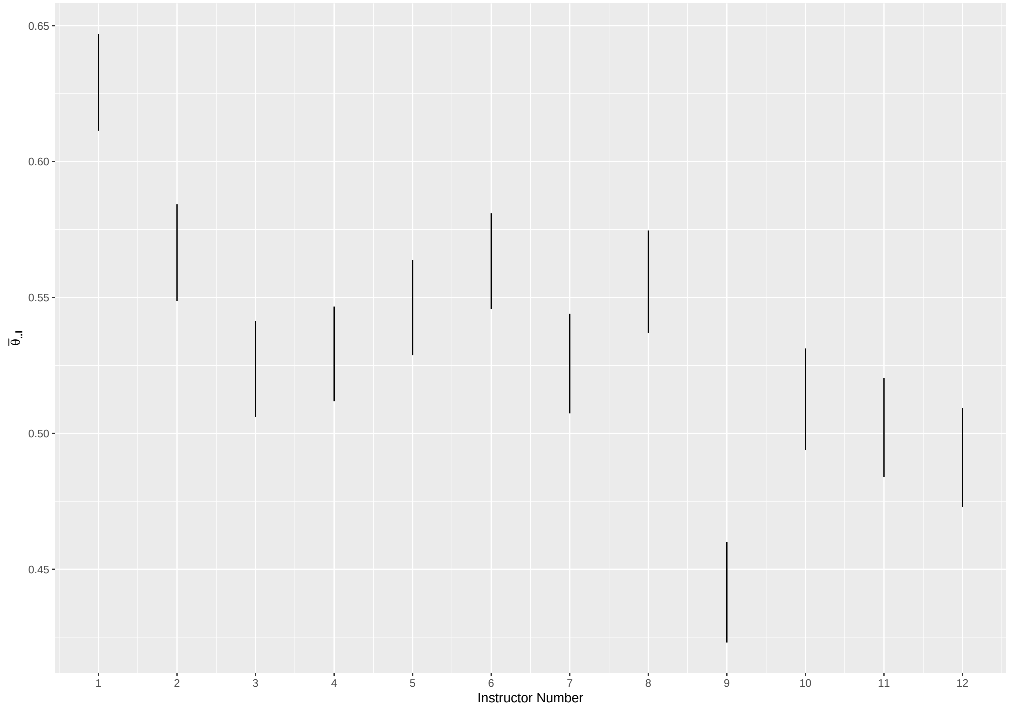

Figure 5: Posterior intervals (90%) for the instructor rating index $\bar{\theta}_{\cdot\cdot l}$ for the $l = 1, \ldots, 12$ instructors in the study.

The approach developed here may be adapted to other inferential questions of interest. For example, suppose pilot i had received instruction from instructor l_1 but we are interested in the counterfactual situation of how the pilot might have performed under instruction from instructor l_2 when no instruction was actually received by pilot i from instructor l_2. The average test score of pilot i under instruction from instructor l is given by $T_{i\cdot l} = \sum_{j=1}^{k}(4\theta_{ijl}+1)$. Therefore, the question of interest is addressed by comparing $T_{i\cdot l_1}$ with $T_{i\cdot l_2}$. When $i = 1$, $l_1 = 1$ and $l_2 = 9$, we have posterior means and posterior standard deviations of $T_{i\cdot l}$ given by 41.03 (1.73) for l_1 and 30.52 (2.36) for l_2. It is apparent that instructor $l_1 = 1$ is much more effective than instructor

$l_2 = 9$ with respect to instruction given to pilot $i = 1$. Notably, it is not clear how a CTT or IRT approach could be used to investigate this complex query.

5 DISCUSSION

We have extended the methods of Silva, Guan and Swartz (2017) to handle nonstandard problems in test analysis. We see that it is not difficult to modify the basic Bayesian model where inference is facilitated using posterior realizations from a Markov chain. The two nonstandard problems that are considered in this paper are real problems of interest in the aviation industry.

One of the features of the proposed approach is that the parameters that we have studied in this paper $\theta_{i\cdot}$, $\bar{\theta}_{\cdot j}$ and r_{j_1,j_2} (see Section 3) are direct analogs of the popular statistics which one would naturally study in the CTT framework (i.e. $x_{i\cdot}$, $\bar{x}_{\cdot j}$ and $r_{x_{i_1},x_{i_2}}$). In Section 3, we have studied data arising on the scale 1,2,3,4,5 which differs from the typical 0/1 data arising from multiple choice exams and true/false questions. In Section 4, our investigation was focused on the parameter $\theta_{\cdot\cdot l}$ whose analogous CTT statistic $x_{\cdot\cdot l}$ may not even be calculable due to missing data. An appeal of CTT over the years is that CTT statistics are intuitive and are often readily adaptive to various test designs. An advantage with our simulation-based Bayesian approach is that we have inferential capabilities to study the θ terms whereas inference does not form a part of CTT analyses. Another advantage of the Bayesian approach is that one may be able to incorporate prior knowledge. For example, we have emphasized that it may be reasonable to assume that good students generally do better than weaker students across all questions. It may also be possible to use results from previous tests to inform prior opinion.

We believe that the basic approach that we have illustrated here may be modified to suit other nonstandard problems of interest in test analysis. For example, suppose that the ith student has taken tests on multiple occasions. It would be a simple matter to treat the student as a different subject according to the times $t_1, \ldots, t_m$ that tests were written. Accordingly, the student would have performance measures $\theta_{i\cdot}^{(t_1)}, \ldots, \theta_{i\cdot}^{(t_m)}$. Posterior estimates and associated intervals for these performance measures could then be plotted against time to assess improvement.

Data and code developed in this paper (both R and JAGS) are available from the authors upon request.

6 REFERENCES

Fan, X. (1998). ‘Item response theory and classical test theory: an empirical comparison of their item/person statistics’. *Educational and Psychological Measurement*, Vol. 58, No. 3, pp. 357-381.

Fox, J-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*, Statistics for Social and Behavioral Sciences Series, Editors S.E. Fienberg and W.J. van der Linden, Springer, New York.

Guler, N., Uyanik, G.K. and Teker, G.T. (2014). ‘Comparison of classical test theory and item response theory in terms of item parameters’. *European Journal of Research on Education*, Vol. 2, No. 1, pp. 1-6.

Hambleton, R.K. and Jones, R.W. (1993). ‘Comparison of classical test theory and item response theory and their application to test development’. *Educational Measurement: Issues and Practice*, Vol. 12, No. 3, pp. 38-47.

Kohli, N., Koran, J. and Henn, L. (2015). ‘Relationships among classical test theory and item response theory frameworks via factor analytic models’. *Educational and Psychological Measurement*, Vol. 75, No. 3, pp. 389-405.

Plummer, M. (2015). *JAGS Version 4.0 User Manual*, Accessed online June 5, 2017 at http://www.uvm.edu/∼bbeckage/Teaching/DataAnalysis/Manuals/manual.jags.pdf

Raykov, T. and Marcoulides, G.A. (2016). ‘On the relationship between classical test theory and item response theory: from one to the other and back’. *Educational and Psychological Measurement*, Vol. 76, No. 2, pp. 325-338.

Silva, R., Guan, Y. and Swartz, T.B. (2017). ‘Bayesian diagnostics for test design and analysis’. *Journal on Efficiency and Responsibility in Education and Science*, 10, 44-50.

Swartz, T.B. (2011). ‘Bayesian clustering with priors on partitions’. *Statistica Neerlandica*, Vol. 65, No. 4, pp. 371-386.