BAYESIAN DIAGNOSTICS FOR TEST DESIGN AND ANALYSIS

R. M. Silva[1], Y. Guan[2], T. B. Swartz[✉3]

[1]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka

[2]Pelesys Learning Systems Inc, Canada

[3✉]Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby, V5A1S6, Canada, +1 778 782 4579, tswartz@sfu.ca

Highlights

- *This paper extends Classical Test Theory statistics to the Bayesian framework and admits inference*

Abstract

This paper attempts to bridge the gap between classical test theory and item response theory. It is demonstrated that the familiar and popular statistics used in classical test theory can be translated into a Bayesian framework where all of the advantages of the Bayesian paradigm can be realized. In particular, prior opinion can be introduced and inferences can be obtained using posterior distributions. In classical test theory, inferential decisions are based on the values of statistics that are calculated from the responses of subjects over various test questions. In the proposed approach, analogous “statistics” are constructed from the output of simulation from the posterior distribution. This leads to population- based inferences which focus on the properties of the test rather than the performance of specific subjects. The use of the JAGS programming language facilitates extensions to more complex scenarios involving the assessment of tests and questionnaires.

Article type

Full research paper

Article history

Received: March 17, 2017
Received in revised form: June 12, 2017
Accepted: June 14, 2017
Available on-line: July 17, 2017

Keywords

Classical test theory, Empirical Bayes, Item response theory, Markov chain Monte Carlo, JAGS programming language

Silva R. M., Guan Y., Swartz T. B. (2017) “Bayesian Diagnostics for Test Design and Analysis”, *Journal on Efficiency and Responsibility in Education and Science, Vol. 10, No. 2*, pp. 44-50, online ISSN 1803-1617, printed ISSN 2336-2375, doi: 10.7160/eriesj.2017.100202.

Introduction

The important problems of test/questionnaire design and analysis have historically been ap- proached from either the perspective of *classical test theory* (CTT) or *item response theory* (IRT). Both of these research areas have an extensive literature where numerous comparative studies have been carried out (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 com- parison of the two approaches, CTT is the original testing framework and essentially concerns 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, more complex models are considered where these models have components (i.e. parameters) that distinguish particular aspects of tests and are generalizable to a population of respondents. IRT relies more on statistical theory and is less accessible to some practioners. IRT has grown in many directions where various models have been proposed. Most 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 thepractitioner.

In this paper, we demonstrate how some of the very simple and still popular statistics of CTT can be directly translated into a Bayesian IRT framework. The advantage to the practitioner is that they may continue using familiar measures but simultaneously take advantage of the utility of the Bayesian paradigm. For example, they can introduce subjective prior opinion (if deemed necessary) and they can view their familiar measures from the perspective of populations (using posterior distributions). In addition, the use of the JAGS programming language (Plummer 2015) facilitates extensions to more complex scenarios involving the assessment of tests and questionnaires.

In Section 2, we provide the background for the typical testing framework involving dichotomous responses arising from test questions. In this context, some of the common statistics used in CTT are provided. This scenario is then imbedded into a Bayesian framework and it is demonstrated how the familiar testing measures can be easily translated into Bayesian diagnostics. Initially, a very simple prior distribution is introduced. In this section, we emphasize the ad- vantages of the proposed approach over the use of the familiar statistics used in CTT. We also demonstrate how missing data pose no difficulty.

In Section 3, we examine some real data taken from the aviation industry that consists of the results of multiple-choice questions given to pilots. We compare the traditional statistics with analogous Bayesian diagnostics. We also consider several extensions to the basic model introduced in Section 2. In particular, we introduce a more realistic prior which recognizes that some questions are more/less difficult for most respondents and that some respondents are stronger/weaker across most questions. The prior is also beneficial in that it reduces the effective dimensionality of the parametrization. We also indicate how the model can be extended to account for different instructors who have an effect on the performance of their students. Finally, we provide a discussion in Section 4 and a short conclusion in Section 5.

Materials and Methods

We consider test data presented in a $n \times k$ matrix $X = (x_{ij})$ where the n rows correspond to the respondents and the k columns refer to the test questions. The data are dichotomous (binary) where $x_{ij} = 1(0)$ specifies that the ith respondent provides a correct (incorrect) answer to the jth question. Therefore, the setup is applicable to true/false questions and to multiple-choice questions. For questions with ordinal grading, it is possible to introduce a threshold that corresponds to pass (fail) so that such questions can also be analyzed within the above framework.

In CTT, there are various statistics that have been proposed to assess the characteristics of test questions and the overall test. We now review three of these statistics. The first statistic, sometimes referred to as the *P-value*, is calculated on each of the *k* test questions. For the *j*th question, its P-value is defined as

$$p_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad (1)$$

and is the proportion of correct responses on the *j*th question. Typically, a question is not viewed as a "good" question if its P-value is either too close to 0 (the question is difficult) or too close to 1 (the question is easy). In such cases, there is little testing taking place since most respondents have the same result.

The second statistic that is referred to as the *discrimination index* is also calculated for each of the *k* test questions. For the *j*th question, its discrimination index is defined as

$$d_j = \frac{N_{Uj} - N_{Lj}}{n/2} \qquad (2)$$

where N_{Uj} is the number of `strong' students who answered the *j*th question correctly and $N_{Lj}$ is the number of `weak' students who answered the *j*th question correctly. The subscripts *U* and *L* denote `upper' and `lower' respectively. The strong and weak students are categorized into two groups according to their overall test score where the test score for the *i*th student is given by $x_{i.} = \sum_{j=1}^{k} x_{ij}$. When *n* is even and the order statistics *x*(*n/*2) and *x*(*n/*2+1) differ, then the two groups form a partition of the set of the *n* respondents. In other cases, slight adjustments are made in forming the two groups. The discrimination index lies in the interval (−1*,* 1) where large positive values are viewed as desirable (strong students do better on the question than weak students), values near zero indicate that the question does not differentiate between strong and weak students, and negative values are viewed as undesirable (weak students do better on the question than strong students).

The third statistic which is referred to as *Cronbach's alpha* is used to describe the *reliability* or *internal consistency* of the overall test. It is defined as

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} s_j^2}{s_X^2}\right) \qquad (3)$$

where

$$s_j^2 = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 / n$$

is the variance with respect to the *j*th question and

$$s_X^2 = \sum_{j=1}^{k} s_j^2 + 2\sum_{j_1 < j_2} \sum_{i=1}^{n} (x_{ij_1} - \bar{x}_{j_1})(x_{ij_2} - \bar{x}_{j_2}) / n$$

is the overall test variance. Cronbach's alpha is constrained to the interval $(-\infty, 1)$ where values near the upper limit are generally preferred (DeVellis 2012). However, we note that various criticisms have been made related to the above interpretation (Sijtsma 2009). For example, if for a given subject, the *k* questions all have the same response, then the questions are redundant, which is obviously not desirable. However, in this case, $\alpha = 1$.

Before introducing the Bayesian analogue corresponding to CTT, there are two points that we wish to emphasize. First, although IRT has overtaken CTT in various ways, the CTT statistics (1), (2) and (3) are still widely used in practice (see for example, Yuan et al. 2012, Brozova and Rydval 2014). Second, as forcibly argued in the IRT literature (e.g. Hambleton and Jones 1993), an important feature of the more complex IRT models is that *item* (question) performance is linked to respondent ability. In other words, the results on test questions vary according to the strength of the student. The models and methods introduced in this paper preserve the simplicity of the common CTT statistics yet allow for the interplay between item performance and ability.Our approach is based on simple Bernoulli models where x_{ij} ~ Bernoulli(θ_{ij}). The model stipulates that the probablity of a correct answer by the *i*th respondent to the *j*th question is given by

$$\text{Prob}(x_{ij} = 1) = \theta_{ij}. \qquad (4)$$

An immediate reaction to (4) 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 is available and parameters may "borrow" from one another such that the effective parameterization is reduced.

Under (4), the development of measures comparable to the statistics (1), (2) and (3) is straight- forward. Instead of calculating (1), (2) and (3) based on the data matrix *X*, the calculations are carried out on the parameter matrix $\Theta = (\theta_{ij})$. And herein lies a possible second reaction - the θ_{ij} 's are unknown. How can one calculate "statistics" based on Θ? The answer again relies on the Bayesian formulation. Under a simulation-based Bayesian approach, Θ's are generated from the posterior distribution, and each simulated sample gives rise to the analogous measures. An important added benefit is that we do not have a single observed statistic (*p, d,* α) as in CTT, but rather, we have a posterior distribution corresponding to our new measures and this facilitates the assessment of variability. These features and other features are emphasized in the real data example presented in Section 4.

There is another attractive aspect of the Bayesian formulation. Whereas the statistics (1), (2) and (3) refer to the observed *X* values, the Bayesian measures refer to theprobabilities associated with the questions and the respondents. And we suggest that this corresponds to the real problem of interest where the properties of the questions/respondents is more important to practitioners than the particular sample. 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. We also mention that there is great flexibility in the approach. Not only can the statistics (1), (2) and (3) be translated to Bayesian versions, we can do likewise with any CTT statistic. The only additional ingredient that is required for the Bayesian implementation is thespecification of a prior distribution on the parameters. Initially, we consider a somewhat unrealistic prior where we assume that the θ_{ij} are independent and identically distributed (iid) Uniform (0*,* 1) random variables. The Uniform distribution is sometimes referred to as a reference prior; it is flat and has the required domain $\theta_{ij} \in (0, 1)$.

Above, we alluded to simulation-based Bayesian software. Accordingly, we use the JAGS programming language which is relatively simple to use and avoids the need of special purpose Markov chain Monte Carlo code. JAGS is open source software (www.mcmc-jags.sourceforge.net) which is very similar to WinBUGS. Details on WinBUGS and an introduction to the Bayesian approach are given by Lunn et al. (2013).

Relationship of approach to IRT

Various models have been proposed in IRT. In a three-parameter logistic IRT model, we retain the notation above and express

$$\theta_{ij} = \frac{1-c_j}{1+\exp\left(-a_j\left(p_i-b_j\right)\right)} \qquad (5)$$

where p_i is the ability parameter for the ith respondent and a_j, b_j and c_j are characteristics of the jth test question.

The relationship (5) is known as an item response function (IRF). The IRF is an important feature of IRT and is typically plotted as a function of the ability p_i for estimated test characteristics $\hat{a}_j$, $\hat{b}_j$ and $\hat{c}_j$. One of the notable differences between our approach and IRT is that we allow more freedom in the θ_{ij} parameters since the θ_{ij} are assigned a prior probability distribution. In IRT, the functional relationship is fixed according to (5) or by some alternative IRT model. Accordingly, in our framework, measures such as the Bayesian P-value and the Bayesian discrimination are not constrained by functional relationships.

Missing data

The Bayesian model is appealing in its simplicity. Via the simulated parameters θ_{ij}, researchers are able to investigate questions involving both respondents and test questions.

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 $[A \mid B]$ denote the generic conditional density of A given B, the relevant posterior distribution in this case is

$$\begin{aligned}\left[\Theta, x_{\text{mis}} \middle| x_{\text{obs}}\right] &\propto \left[x_{\text{mis}}, x_{\text{obs}}\right] \\ &= \left[x_{\text{obs}}, x_{\text{mis}} \middle| \Theta\right]\left[\Theta\right].\end{aligned} \qquad (6)$$

The key observation from (6) is that $\left[x_{\text{obs}}, x_{\text{mis}} \middle| \Theta\right]\left[\Theta\right]$ 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. To handle missing data in JAGS, we need only code the unobserved data values with the NA symbol. We emphasize that this is incredibly easy to do.

Results

We consider the results of a multiple-choice exam given to pilots where there are n = 307 respondents (pilots) and k = 10 test questions. In the aviation industry, safety is of paramount importance, and therefore, the proportion of correct answers must be very high. We first calculate various CTT statistics. For this dataset the vector of P-values is

$$p = \left(0.925, 0.837, 0.990, 0.967, 0.971, 0.932, 0.977, 0.993, 0.896, 0.951\right)'.$$

The vector for the discrimination index is

$$d = \left(0.150, 0.326, 0.020, 0.065, 0.059, 0.137, 0.046, 0.013, 0.208, 0.098\right)'$$

which indicates that all questions are answered better by the stronger students than by the weaker students. Cronbach's alpha is α = 0.492 which (for many researchers) indicates that the test is reliable.

Since the P-value and discrimination index provide properties of the same test, they are sometimes interpreted jointly. In Table 1, we provide guidelines (Skoda, Doulik and Hajerova- Mullerova 2006) that have been proposed for a suitable test and have been endorsed by Brozova and Rydval (2014). Although practitioners may have alternative guidelines for a particular application, here we illustrate the utility of the proposed Bayesian with respect to the guidelines provided in Table 1.

We now present some results based on 1000 simulations from the posterior distribution. For each simulation, the Bayesian P-value, the discrimination index and Cronbach's alpha were calcu- lated. In Figure 1, we provide the joint distribution of the Bayesian P-value and the discrimination index for questions 1 and 2. In contrast to the single paired observations (p_1 = 0.925, d_1 = 0.612) and (p_2 = 0.837, d_2 = 0.788), Figure 1 highlights that there is variability associated with each measure and uncertainty is expressed via the posterior distribution. In each of the plots, we have provided bars according to the guidelines in Table 1 which allows us to assess the suitability of the test questions. We observe a difference between the properties of question 1 and question 2. For example, question 2 is more difficult (i.e. the cloud of points is slightly shifted to the left). We also observe that there is more variability in the discrimination index than in the P-value.

P-value	[0.20,0.30]	[0.30,0.70]	[0.70,0.80]
Discrimination	≥0.15	≥0.25	≥0.15

Table 1: Recommended values for the P-value and discrimination index for a test question (Skoda, Doulik and Hajerova-Mullerova 2006).

We also observe in Figure 1 that the generated P-values are smaller than the traditional CTT statistics p_1 = 0.925 and p_2 = 0.837. This is due to the unrealistic θ_{ij} ~ Uniform(0, 1) prior distribution which shrinks the posterior distribution of θ_{ij} towards 0.5. In a particular application, we may have specific knowledge concerning the θ_{ij} values, and this knowledge can be incorporated into the prior distribution. We illustrate this flexibility in Section 4.

In Figure 2, we provide a density plot of the posterior distribution of the Bayesian version of Cronbach's alpha. Again, the figure highlights that there is variability associated with the measure. One of the frequent discussion points concerning the use of Cronbach's alpha is that its interpretation is subject to the dimension of the $n \times k$ data matrix X. With the Bayesian version of Cronbach's alpha, the observed variability depends on the dimension of X. We note that the posterior mean 0.075 in Figure 2 differs from the traditional CTT statistics α = 0.492. In Section 4, we vary the prior and observe changes in the resultant posterior mean.

A more realistic prior

We now turn our attention to the development of a more realistic prior, one which recognizes that some questions are more/ less difficult for most respondents and that some respondents are stronger/weaker across most questions. The intention is to introduce a prior distribution that leads to Bayesian CTT statistics that are more in line with the traditional CTT statistics. This allows practitioners to use the same calibration scales with which they are comfortable.

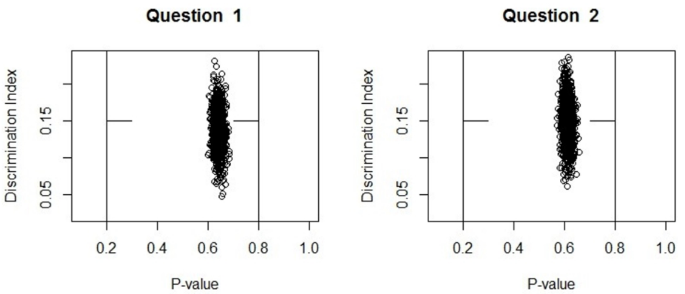

Figure 1: Posterior simulations of the Bayesian P-value and discrimination index for questions 1 and 2 using the iid uniform prior. Horizontal lines are drawn to delineate the recommendations from Table 1.

The suggested prior has the following assumed structure

$$[\Theta] = \prod_{ij}\left[\theta_{ij}\right]$$

where

$$[\Theta] \sim \text{truncated} - \text{Normal}\left(\mu_{ij}, \sigma_{ij}^2\right). \qquad (7)$$

In (7), the truncation corresponds to the interval (0, 1) and the parameters μ_{ij} and σ_{ij}^2 are specified according to an empirical Bayes procedure. The procedure first requires logistic regression involving the original data X where

$$\text{logit}\left(\theta_{ij} \middle| \beta_0, \alpha_i, \gamma_j\right) = \beta_0 + \alpha_i + \gamma_j. \qquad (8)$$

Logistic regression provides us with parameter estimates $\hat{\beta}_0, \hat{\alpha}_i$ and $\hat{\gamma}_j$. We then invert the logistic function and set

$$\mu_{ij} = \frac{\exp\left(\hat{\beta}_0 + \hat{\alpha}_i + \hat{\gamma}_j\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\alpha}_i + \hat{\gamma}_j\right)}$$

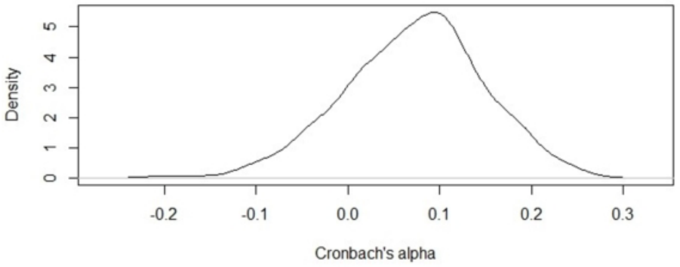

Figure 2: Posterior density plot of the Bayesian version of Cronbach's alpha using the iid uniform prior.

To set σ_{ij}^2, we make use of the Delta method applied to (8). After some calculations, this yields

$$\sigma_{ij}^2 = \frac{\exp\left(2\left(\hat{\beta}_0 + \hat{\alpha}_i + \hat{\gamma}_j\right)\right)\hat{\upsilon}}{\left(1 + \exp\left(\hat{\beta}_0 + \hat{\alpha}_i + \hat{\gamma}_j\right)\right)^4}$$

where $\hat{v}$ is the sum of the entries in the variance-covariance matrix corresponding to the parameter estimates.

Whereas the calculation of μ_{ij} and σ_{ij}^2 may appear daunting for some practitioners, we note that the *predict* function can be used on a *glm* object in R to provide the values. This is most convenient when running the *rjags* package since it provides an interface from R to the JAGS library. In the Appendix, we see that the empirical Bayes procedure requires only three statements of code.

To check the impact of the empirical Bayes prior specification (7), we repeat the Bayesian analysis on the aviation dataset. Recall for question 1, the CTT P-value was 0.925 and the posterior mean of the Bayesian P-value was 0.642. With the new prior that takes into account student ability and test difficulty, the posterior mean of the Bayesian P-value is 0.912. We therefore see that the new value has moved towards the CTT value. Similarly, with Cronbach's alpha, the CTT value was 0.492, the posterior mean of the Bayesian α was 0.075, and the posterior mean of the Bayesian α based on the empirical Bayes prior specification (7) is 0.201.

In Figure 3, we provide the joint distribution of the Bayesian P-value and the discrimination index for questions 1 and 2 based on the empirical Bayes prior of Section 4.2. The distribution of values are more in line with the CTT diagnostics. In Figure 4, we provide a density plot of the posterior distribution of the Bayesian version of Cronbach's alpha based on the empirical Bayes prior of Section 4.2. Again, the distribution of values are more in line with the CTT diagnostic. We repeat that a main advantage of the empirical Bayes procedure is that it takes into account the difficulty of questions and the strength of the respondent.

The prior specification in (7) provides only a template of what can be done. For example, one could introduce alternative distributions. One could also introduce more knowledge about students and test questions by modifying the truncated-Normal distribution. In the Appendix, we see that the specification of the prior in JAGS is straightforward (e.g. one line involving the dnorm function).

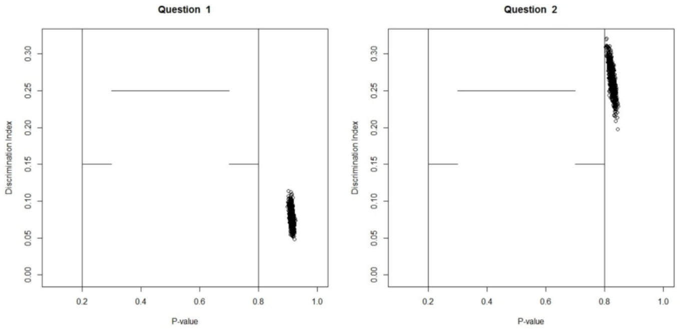

Figure 3: Posterior simulations of the Bayesian P-value and discrimination index for questions 1 and 2 using the empirical Bayes prior of Section 4.2. Horizontal lines are drawn to delineate the recommendations from Table 1.

Generalizing with respect to instructors

We now demonstrate that the Bayesian framework provides advantages that are not available in the classical CTT framework.

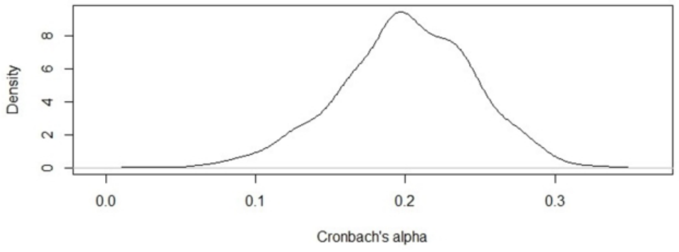

Figure 4: Posterior density plot of the Bayesian version of Cronbach's alpha using the empirical Bayes prior of Section 4.2.

A possible application is 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} = 1(0) denotes that the i student has a correct (incorrect) response to the jth question and that this student received instruction on this question by instructor l. We similarly extend the notation for the parameters leading to terms θ_{ijl}. The above setup is also applicable to other situations.

For example, a comparison of different groups of students may be of interest where the groups are designated by the index *l*.

Using either the simple uniform prior or the more realistic prior given by (7) and (8), posterior realizations of θ_{ijl} are generated as before. Let $S_l = \{\theta_{ijm}: m = l\}$ and let *nl* be the number of terms in the set S_l. Then an analysis of instructors in the spirit of the CTT Bayesian framework can be based by calculating

$$\bar{\theta}_{..l} = \frac{1}{n_l} \sum\nolimits_{S_l} \theta_{ijl} \tag{9}$$

which can be interpreted as the average probability of a correct answer for instructor *l*. One can compare the $\bar{\theta}_{..l}$ values, $l = 1,\dots, L$, and assess their relative magnitudes by also calculating their corresponding posterior standard deviations.

Discussion

The two main approaches to questionnaire design and analysis are IRT and CTT. Methods based on IRT require the specification of statistical models and permit the inferential benefits associated with the models. IRT is the dominant approach used in major educational testing initiatives (An and Yung 2014) and IRT software is now widely accessible including popular statistical packages such as SAS (Choi 2017). Much recent research has been carried out under the IRT umbrella and there are now many IRT models that can be considered for a given application (Cai et al. 2016).

However, despite the popularity of IRT, there are two main drawbacks involving IRT. First, sometimes the existing statistical models do not adequately characterize the special features of an application and the models need to be modified (if possible) to account for these features. In comparison to CTT, Hambleton and Jones (1993) describe the assumptions related to IRT as `strong'. Second, the sophistication of the IRT models in terms of model fitting and interpretation is sometimes beyond the technical scope of practitioners. For example, even the simple IRF given in (5) often poses a challenge for a non-technical audience.

On the other hand, CTT approaches consist of few assumptions and are easily adopted by practitioners. These appealing features have led to the continuation of the use of CTT despite the lack of inferential capabilities under CTT. For example, in clinical psychology when there are fewer than 20 test items, Jabrayilov, Emons and Sijtsma (2016) recommend CTT over IRT for detecting change in individuals. In discussing CTT, Hambleton and Jones (1993) write that the dependence of the methodology on the particular test and examinees `limit the utility of the person and item statistics in practical test development work and complicate any analyses'.

The methods proposed in this paper allow practitioners to work under the familiar CTT approach, yet benefit from inferential capabilities. This is accomplished by imbedding the CTT structure within a Bayesian framework. The inferential component is accomplished via simulation from posterior distributions where simulated values provide population-level descriptions of questionnaires.

However, the greatest advantage of the proposed approach is its flexibility. We have seen that we can vary the prior to take into account subjective beliefs concerning students and test questions. In addition, the flexibility of applications is facilitated through the availability of the simulated θ_{ij} values (something that is not immediately available in IRT). For example, we have shown in Section 3 how the introduction of a new subscript can extend an investigation to take into account the effect of instructors. As another example, suppose that a researcher is interested in the performance of students on test questions 6, 7 and 8. Then, for the *i*th student, the researcher needs only keep track of the simulated outcomes $T_i = \theta_{i6} + \theta_{i7} + \theta_{i8}$. Essentially, with the θ_{ij} values, the researcher can investigate any aspect of interest regarding students and test questions.

Finally, we have used an empirical Bayes procedure based on fitting a logistic regression model according to (8). Nothing prevents us from using a similar procedure based on an alternative parametrization. For example, we could fit a logistic regression model according to three-parameter IRF (5). This would further tighten the relationship between our Bayesian CTT approach and IRT.

Conclusion

We have made the case that the approach developed in this paper may help bridge the gap between CTT and IRT, by retaining the simplicity of CTT and by providing the inferential advantages of IRT. In particular, when compared to traditional CTT, the proposed approach does not rely on the interpretation of summary statistics. Rather, variability can be assessed via posterior distributions.

Acknowledgements

Swartz has been partially supported by grants from the Natural Sciences and Engineering Re- search Council of Canada (NSERC). The authors thank three anonymous reviewers for their comments.

References

An, X. and Yung, Y.-F. (2014). ‘Item response theory: what it is and how you can use the IRT procedure to apply it’. *SAS Institute Inc., Cary NC, Paper SAS364-2014*. Accessed online June 5, 2017 at https://pdfs.semanticscholar.org/d85a/7025441f5685b287b53234ce6456dcd40192.pdf

Brozova, H. and Rydval J. (2014). ‘Analysis of exam results of the subject ”Applied Mathematics for It”’. *Journal on Efficiency and Responsibility in Education and Science*, Vol. 7, No. 3-4, pp. 59-65. https://doi.org/10.7160/eriesj.2014.070303

Cai, L., Choi, K., Hansen, M. and Harrell, L. (2016). ‘Item response theory’. *Annual Review of Statistics and Its Application*, Vol. 3, pp. 297-321. https://doi.org/10.1146/annurev-statistics-041715-033702

Choi, J. (2017). ‘A review of PROC IRT in SAS’. *Journal of Educational and Behavioral Statistics*, Vol. 42, No. 2, pp. 195-205. https://doi.org/10.3102/1076998616664568

DeVellis, R.F. (2012). *Scale Development: Theory and Applications, Third Edition*, Applied Social Methods Research Series, Editors L. Bickman and D.J. Rog, Sage, Los Angeles.

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. https://doi.org/10.1177/0013164498058003001

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. Accessed online June 5, 2017 at http://iassr2.org/rs/020101.pdf

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. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Jabrayilov, R., Emons, W.H.M. and Sijtsma, K. (2016). 'Comparison of classical test theory and item response theory in individual change assessment". *Applied Psychological Measurement*, Vol. 40, No. 8, pp. 559-572. https://doi.org/10.1177/0146621616664046

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. https://doi.org/10.1177/0013164414559071

Levy, R. and Mislevy, R.J. (2016). *Bayesian Psychometric Modeling*, Chapman & Hall/CRC Statistics in the Social and Behavioral Science Series, Boca Raton.

Lunn, D., Jackson, C., Best, N., Thomas, A. and Spiegelhalter, D. (2013). *The BUGS Book: A Practical Introduction to Bayesian Analysis*, Chapman & Hall/CRC Texts in Statistical Science Series, Boca Raton.

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. https://doi.org/10.1177/0013164415576958

Sijtsma, K. (2009). 'On the use, misuse, and the very limited usefulness of Cronbach's alpha'. *Psy- chometrika*, Vol. 74, No. 1, pp. 107-120. https://doi.org/10.1007/s11336-008-9101-0

Skoda, J., Doulik, P. and Hajerova-Mullerova, L. (2006). '*Zasady spravne tvorby, pouziti a hodnoceni didaktickych testu v pripave budoucich ucitelu*'. Accessed online June 5, 2017 at http://cvicebnice. ujep.cz/cvicebnice/FRVS1973F5d

Swartz, T.B. (2011). 'Bayesian clustering with priors on partitions'. *Statistica Neerlandica*, Vol. 65, No. 4, pp. 371-386. https://doi.org/10.1111/j.1467-9574.2011.00490.x

Yuan, W., Deng, C., Zhu, H. and Li, J. (2012). 'The statistical analysis and evaluation of exam- ination results of materials research methods course'. *Creative Education*, Vol. 3, pp. 162-164. https://doi.org/10.4236/ce.2012.37B042

Appendix

Here we provide the JAGS code used in the analysis in Section 4.2. We see that the code is straightforward and is easily adaptable to more complex testing problems.

```
# The following code reads in a test matrix and obtains
# posterior means of various test parameters using Just
# Another Gibbs Sampler (JAGS) through the R library
# 'rjags'. Here we assume the realistic independent
# truncated normal prior.

sink(file.path(tempdir(),"model.txt"))
cat("
     model
{
for (i in 1:n)
{
for( j in 1:k)
{
x[i,j]      ~ dbern(theta[i,j])
theta[i,j] ~ dnorm(mu_ij[i,j],1/pow(se.fit[i,j],2))T(0,1)
}
}
for(j in 1:k)
{
theta_dotj [ j ] <- sum(theta[, j ])
}
for( i in 1:n)
{
thetai_dot [ i ] <- sum(theta[ i, ])
}
for( j in 1:k)
{
Pvalue[ j ] <- theta_dotj [ j ] / n
}
thetai_dotbar <- mean(thetai_dot[]) mid <- (n+1)/2
Index <- rank(thetai_dot[])
for(i in  1:n)
{
for(j in 1:k)
{
G[i,j]    <-  step(Index[i]  -  mid)*theta[i,  j]  G_dash[i,j]  <-
step(mid - Index[i])*theta[i, j]
}
}
for(j in 1:k)
{
Nu[j] <- sum(G[, j])
Nl[j] <- sum(G_dash[, j])
discrim[j] <- 2*(Nu[j] -  Nl[j])/n
}
for(i in  1:n)
{
for(j in 1:k)
{
tmp[i,j] <- theta[i,j] - Pvalue[j]
}
}
covari[1:k,1:k] <- t(tmp[,]) %*% tmp[,]
for(j in 1:k)
{
covi[j] <- sum(covari[j,])
}
sum.cov <- sum(covi[1:k]) for(j in 1:k)
{
vari[j] <- covari[j,j]
}
sum.var <- sum(vari[1:k])

# To calculate kron = k*(1 - sum(diag(cov[,]))/sum(cov[,]))/(k-

1) eps <- pow(10,-50)
A <- (sum.var)/(sum.cov+eps) kron <- k*(1 -A)/(k-1)
}
        ", fill = TRUE) sink()
data0 <- read.csv("data123.csv", header=T,row.names =
"UserID") x <- data0[complete.cases(data0),]
long_format = matrix(ncol=3, nrow=nrow(x)*ncol(x))
for(i in1:nrow(x))
{
for(j in 1:ncol(x))
{
k = j + (i-1)*ncol(x)
long_format[k,1] = x[i,j] long_format[k,2] = row.names(x)[i]
long_format[k,3] = names(x)[j]
}
}
long_format = data.frame(long_format)
names(long_format) = c("Correct","Respondent","Question")
long_format$Respondent = as.character(sort(as.
numeric(levels(long_format$Respondent))))
# Empirical Bayes procedure
mod = glm(Correct ~ Respondent + Question,data=long_
format,family="binomial")
mu_ij = matrix(predict(mod,type="response",data=long_forma
t),ncol=ncol(x),nrow=nrow(x), byrow=TRUE)
se.fit_ij = matrix(predict(mod,type="response",data=long_
format,se.fit=TRUE)$se.fit, ncol=ncol(x),nrow=nrow(x),byrow
=TRUE)
n <- nrow(x) k <- ncol(x)
linedata <- list("n" = n, "k"=k, "x" = x, "mu_ij"=mu_ij,
"se.fit_ij"=se.fit_ij) parameters <- c("Pvalue","discrim","theta
","kron")
# We call the model above into JAGS
mult.sim <- jags.Model(file = file.path(tempdir(),"model.txt"),
                  data = linedata,
                  inits = NULL, n.chains = 1,
                  n.adapt = 1000)
# We update the MCMC chains 1000 times for burn-in
update(mult.sim, n.iter = 1000)
# Sampling phase
mcmc.out <- coda.samples(mult.sim,
             variable.names = parameters,
             thin = 1,
             n.iter = 1000)
# To get the output
output <- as.data.frame(as.matrix(mcmc.out, chains = TRUE))
```