Fit-to-Model Statistics for Evaluating Quality of Bayesian Student Ability Estimation

Ling Tan
Australian Council *for* Educational Research
19 Prospect Hill Rd,
Camberwell, VIC, Australia 3124
ling.tan@acer.edu.au

ABSTRACT

Bayesian ability estimation is a statistical inferential framework constructed from a measurement model and a prior knowledge model. It is attractive in practice because Bayesian estimation methods offer an elegant way to incorporate appropriate knowledge on target ability distribution in order to improve the accuracy of ability estimation, when there are uncertainties or errors in observable data. One hurdle for applying Bayesian-based methods is evaluating the validity of Bayesian ability estimates at individual-level. This study investigated a class of fit-to-model statistics for quantifying the evidence used in learning Bayesian estimates. The relationship between fit-to-model statistics and root mean square error of Bayesian ability estimation was demonstrated with simulation.

Keywords

Bayesian ability estimation, student modeling, evaluation.

1. INTRODUCTION

Bayesian ability estimation methods have been widely applied in educational testing [2]. Despite its popularity, Bayesian ability estimation is not the standard estimation method of individual ability estimation for several reasons. One criticism of the Bayesian approach is that one can arrive at significantly different answers if different prior distributions are used when analysing the same evidence [2]. In general, two students with the same raw score may get two different Bayesian score if they have different prior distribution. The student having a higher average prior will get a higher Bayesian score than the student having a lower average prior. This criticism invites the methodology of evaluating alternative prior distributions, when more than one prior distribution is available. Another closely related criticism is that Bayesian scores may be biased towards the mean of prior distribution [2]. Again, this points to the necessity of assessing the weight of prior knowledge against evidence from empirical data. This study investigated a class of fit-to-model statistics for quantifying the evidence used in learning Bayesian estimates. The relationship between fit-to-model statistics and root mean square error of Bayesian ability estimation was demonstrated with simulation. In this study, latent ability and evidence are assumed to be uni-dimensional. It means that tasks (or items) come from the same domain, and latent ability can be measured on the same scale as tasks.

2. METHODOLOGY

In the context of individual ability estimation, Bayesian ability estimation can be stated as following. Given a student responses $\mathbf{x}=\{x_1, x_2, \ldots, x_L\}$, and a prior ability θ, the posterior distribution of this student ability is written as,

$$P(\theta|\mathbf{x}) = \frac{P(\mathbf{x}|\theta)P(\theta)}{P(\mathbf{x})}$$

$P(\mathbf{x}|\theta)$ represents the causal relationship between latent ability θ and response vector $\mathbf{x}$. An important assumption in $P(\mathbf{x}|\theta)$ is that no dependence is among item responses given θ. This study uses Expected A Posteriori (EAP) proposed by Bock & Mislevy [1] as Bayesian estimation method. The Expected A Posteriori (EAP) score is the sum of all possible products of $P(\theta|\mathbf{x})$ and θ.

The following section presents a class of fit-to-model statistics for evaluating EAP ability estimation. In this study, the methodology to evaluate evidence is based upon evaluating the conformality of empirical data to an ideal of conjoint measurement. Rasch measurement model is one instance of conjoint measurement. Measurements using this methodology are known as fit-to-model (or model-fit) statistics. These statistics may be based on a residual-based index measuring the distance between observed reponses and the expectation of Rasch-type measurement model. This class of fit-to-model statistics is based on substantial measurement theory. Specifically, these fit-to-model statistics allow one to assess the non-crossing properties of person response functions, which are characterised by P_i. When person response functions are parrallel (or non-crossing), the invariance of person order is maintained. In other words, the order of individual abilities is the same across item difficulty scale. In Rasch-type measurement model, both person order and item order are invariant by definition. Therefore, checking the comformality to Rasch model is effectively assessing the quality of evidence.

A simple residual statistic is the squared standardised residual. A mean squared standardised residual is the squared standardised residual divided by the degree of freedom. The mean squared standardised residual fit statistic (MNSQ) [3] for an individual with latent ability estimate $\hat{\theta}$ and observed responses **x** of the length *L* is represented as,

$$MNSQ = \frac{1}{L-1}\sum_{i=1}^{L}\frac{(x_i - E_i)^2}{Var_i}$$

3. SIMULATION STUDY

The accuracy of estimated Bayesian abilities was compared with the true abilities. The evaluation of ability estimation is done by using Root Mean Square Error (RMSE), i.e. $RMSE(\hat{\theta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta)^2}$, where $\hat{\theta}_r$ is the estimated ability at r^{th} replication, and R is the number of replications. Simulation was performed with the help of ConQuest software [4]. The quadrature points used in marginal probability estimation are 15.

This experiment aims to show the relationship between the MNSQ fit-to-model statistic and the accuracy of Bayesian ability estimates. For each replication, a test of 30 items was generated in normal distribution, $N(0,1)$. A sample of 5,000 response data was generated from a norm ability distribution $N(0,1)$, and the 30-item test. This data set consists of 16% cheating examinees at low ability range (i.e. $\theta \leq -1$), and 16% careless examinees at high ability range $\theta \geq 1$. The cheating responses were created by imputing correct responses to the most difficulty items (i.e. $\delta > 1$), and the careless responses were created by imputing incorrect responses to the easiest items (i.e. $\delta < -1$). Another data set of 5,000 data was generated from the same test without aberrant responses, and this data set was used to set a baseline benchmark. The RMSE was calculated with a replication of 20.

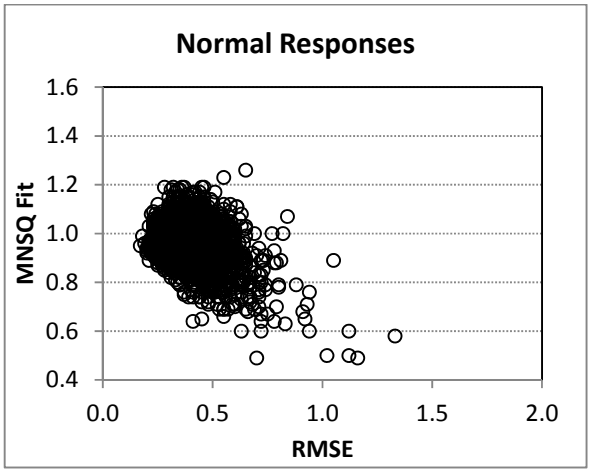

Figure 1a

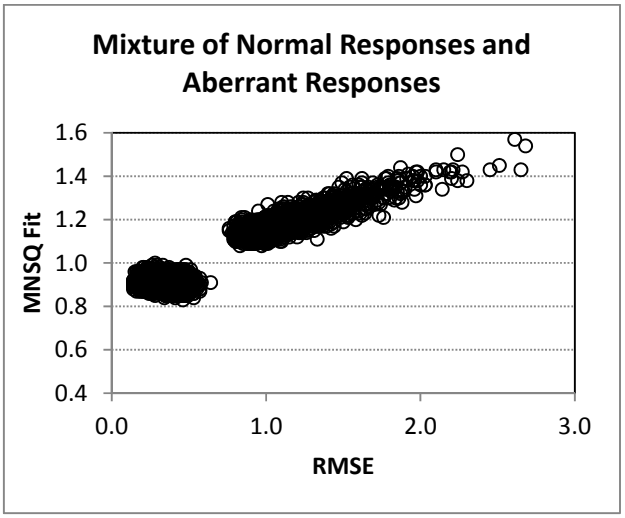

Figure 1b

The Figure 1a shows the relationship between MNSQ fit and RMSE for normal responses, and Figure 1b shows the relationship between MNSQ fit and RMSE for a mixture of normal responses and aberrant responses (i.e. 32%). Both MNSQ and weighted MNSQ fit statistics have an expectation of 1 and variance of $2/L$ [12]. Thus, MNSQ fit values are expected to be centred at 1 and in the range of (0.5, 1.5) for a test with 30 items. MNSQ values less than 0.5 are considered as over-fit, and MNSQ values greater than 1.5 are considered as under-fit. For the normal responses, MNSQs are centred on 1, and they are mostly clustered in the range of (0.6, 1.2), and RMSEs are mostly scattered in the range of (0.2, 0.7). Thus, MNSQ model-fit statistics for normal responses are in a reasonable range. For the mixed data set, responses were scattered in two distinct clusters. The cluster located at the bottom-left has the RMSE in the range of (0.2, 0.65), which is similar to the RMSE in the baseline figure. The bottom-left cluster has the MNSQ in the range of (0.8, 1.0), which is in the range of reasonably good fit. The top-right cluster has the MNSQ in the range of (1.05, 1.4) and the RMSE greater than 0.7. It appears that the MNSQ fit statistic is reasonably sensitive to large RMSE, for at least this experiment.

REFERENCES

[1] Bock, R.D. & Mislevy, R.J (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431-444.

[2] Hambleton, R.K. & Swaminathan,H. (1985). *Item response theory principles and applications*. Boston, MA: Kluwer-Nijhoff Publishing.

[3] Wright, B.D. & Stone, M.H. (1979), *Best Test Design*, Chicago, MESA Press.

[4] Wu, M. L., Adams, R. J., Wilson, M. R., Haldane, S.A. (2007). ACER ConQuest Version 2: Generalised item response modelling software [computer program]. Camberwell: Australian Council *for* Educational Research.