Bayesian Item Response Models For Complex Survey Data

Jean-Paul Fox[1]

[1] Twente University, Faculty of Behavioural Sciences, Enschede, The Netherlands

Abstract: IRT methods have become an important tool in analyzing large-scale survey data. The application of the common IRT models raises several issues like the implicit assumption of conditionally independent observations, handling collateral information, and dealing with misreporting. It is shown that the Bayesian IRT approach leads to a very flexible modeling framework for analyzing large-scale survey data. The Bayesian IRT models are extended to provide a better fit to the data and to extract richer information from the survey data. A variety of extensions will be discussed.

Keywords: Bayesian; Complex Surveys; IRT; Hierarchical Models

1 Introduction

The common item response theory (IRT) methods (Lord and Novick, 1968) are standard tools for the analysis of large-scale survey data. For example, in educational survey research, the National Assessment of Educational Progress (NAEP) is primary focused on scaling the performances of a sample of students in a subject area (e.g., mathematics, reading, science) on a single common scale, and measuring change in educational performance over time. Further, the Organization for Economic Cooperation and Development (OECD) organizes the Program for International Student Assessment (PISA) that is focused on measuring and comparing the abilities in reading, mathematics, and science of 15-year-old pupils over 32 countries in 2000. Another example is the large international survey Trends in International Mathematics and Science Study (TIMSS) conducted by the International Association for the Evaluation of Educational Achievement (IEA) also to measure trends in students' mathematics and science performances.

IRT methods provide a set of techniques for estimating individual ability (e.g., attitude, behavior, performance) levels and item characteristics from observed discrete multivariate response data. The ability levels cannot be observed directly but are measured via a questionnaire or test. Item response theory (IRT) is, in particular, useful for large-scale survey response data where (1) the observations often have an ordinal character, (2) the sampling designs are complex with individuals responding to different sets

(booklets) of questions, (3) booklet effects are present (the performance on items depends on an underlying latent variable but also on the positioning of the items in a test), and (4) missing data occur. The essential idea of IRT is that the effects of the persons and the items on the response data are modeled by separate sets of parameters. The person parameters are usually referred to as the latent variables, and the item parameters are usually labeled item difficulties or thresholds, item discrimination parameters and guessing parameters.

The common IRT models are not directly applicable to analysing large-scale survey data for comparative research. There are several measurement issues connected to survey research that need to be addressed since ignoring them may lead to inferential errors. Further, there is often a wide variety of additional information available besides the observed response data. More accurate inferences can be made when the different sources of information can be combined.

Three topics will be considered. First, the multistage sampling design since respondents are nested in classrooms, classrooms in schools, schools within countries and so on. In a Bayesian modeling approach, a hierarchical population distribution for the respondents is easily specified that accounts for the fact that respondents are nested within clusters. Common IRT models assume a priori independence between individual abilities but homogeneity of results of individuals in the same school is to be expected since pupils in the same school share common experiences. Second, collateral information can be used when response times are observed besides the response patterns. Response times on test items are easily collected in modern computerized testing. When collecting both (binary) responses and (continuous) response times on test items, it is possible to measure the accuracy and speed of respondents. The observed response times can be informative with respect to the latent individual abilities. Third, the collection of data through surveys on personal and sensitive issues may lead to answer refusals and false responses, making inferences difficult. Respondents often have a tendency to agree rather than disagree (acquiescence) and a tendency to give social desirable answers (social desirability). A multivariate randomized response sampling design can be used to improve the quality of the survey data. It is shown that a Bayesian IRT model is easily adjusted to handle the multivariate randomized (item) response data.

2 Bayesian IRT Models for Binary Response Data

An IRT model for binary response data defines the probability of a correct or positive response to item k ($k = 1, \ldots, K$) for individual i ($i = 1, \ldots, n$) given the item characteristics, denoted as $\boldsymbol{\xi}_k = (a_k, b_k)^t$, and the individual ability level, θ_i. The well known probit version of the two-parameter IRT model is also known as the normal ogive model where the probability of

success is defined via a cumulative normal distribution,

$$P(Y_{ik} = 1 \mid \theta_i, a_k, b_k) = \Phi(a_k\theta_i - b_k) = \int_{-\infty}^{a_k\theta_i - b_k} \varphi(z)dz, \tag{1}$$

where $\Phi(.)$ and $\varphi(.)$ is the cumulative normal distribution function and the normal density function, respectively. The a_k is referred to as the discrimination parameter and the b_k as the item difficulty parameter.
The Bayesian approach towards IRT modeling starts with the specification of prior distributions. In most cases there is not much information about the values of the item parameters. Without a priori knowledge to distinguish the item parameters it is reasonable to assume a common distribution for them.
An intuitive assumption of an IRT model is that the higher a respondent's ability level the more likely it is the respondent scores well on each item. This so-called monotonicity assumption implies that $P(Y_{ik} = 1 \mid \theta_i)$ is nondecreasing in θ_i, for binary response data, which is satisfied when the discrimination parameter is restricted to be positive. A common dependency structure of item parameters is specified via a hierarchical structured prior. A multivariate normal prior distributed prior is assumed for the item parameters. It follows that,

$$(a_k, b_k)^t \sim \mathcal{N}\left(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi\right) I_{\mathcal{A}_k}(a_k), \tag{2}$$

where the set $\mathcal{A}_k = \{a_k \in \mathcal{R}, a_k > 0\}$ with hyper prior parameters

$$\boldsymbol{\Sigma}_\xi \sim \mathcal{IW}(\nu, \boldsymbol{\Sigma}_0) \tag{3}$$

$$\boldsymbol{\mu}_\xi \mid \boldsymbol{\Sigma}_\xi \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_\xi/K_0), \tag{4}$$

for $k = 1, \ldots, K$. The truncated multivariate Normal distribution in Equation (2) is the exchangeable prior for the set of K item parameters $\boldsymbol{\xi}_k$. The joint hyper prior distribution for $(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)$ is a Normal inverse Wishart distribution, denoted as $\mathcal{IW}$, with parameters $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0/K_0; \nu, \boldsymbol{\Sigma}_0)$ where K_0 denotes the number of prior measurements, and ν and $\boldsymbol{\Sigma}_0$ describe the degrees of freedom and scale matrix of the inverse-Wishart distribution. These parameters are usually fixed at specified values. A proper vague prior is specified with $\boldsymbol{\mu}_0 = \mathbf{0}$, $\nu = 2$, a diagonal scale matrix $\boldsymbol{\Sigma}_0$ with elements 100 and K_0 a small number.
In general, the respondents are are assumed to be sampled independently and identical distributed from a large population. So, an independent prior distribution is specified for the ability parameter,

$$\theta_i \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2), \tag{5}$$

for $i = 1, \ldots, n$. A Normal inverse Gamma prior is the conjugate prior for the Normal distribution with unknown mean and variance. Therefore, a

hyper prior distribution is specified as,

$$\sigma_\theta^2 \sim \mathcal{IG}(g_1, g_2) \tag{6}$$

$$\mu_\theta \mid \sigma_\theta^2 \sim \mathcal{N}(\mu_0, \sigma_\theta^2/n_0), \tag{7}$$

where g_1 and g_2 are the parameters of the inverse Gamma distribution denoted as $\mathcal{IG}$ and n_0 presents the number of prior measurements.

3 Heterogeneity of the Respondent Population

Educational survey research is often concerned with exploring differences within and between schools. The objective is to investigate the relationship between explanatory and outcome factors. This involves choosing an outcome variable, such as student's ability, and studying differences among schools after adjusting for relevant background variables. A general acceptable statistical model in the assessment requires the deployment of multilevel analysis techniques. The student's ability is considered to be an outcome variable of the multilevel regression model. This outcome variable is not directly observable but is known to be a latent variable. The idea is to integrate the IRT model for measuring the individual abilities with a (structural) multilevel model that explains differences at different levels of abilities. The IRT measurement model defines the relationship between the ability and the corresponding observed response data. The structural multilevel model describes the nested structure of individual abilities in the population.

The respondents at level-1 are nested in clusters and indexed $i = 1, \ldots, n_j$ for $j = 1, \ldots, J$ clusters. Let level-1 respondent-specific covariates be denoted by $\mathbf{x}_{ij}$. The level-1 prior distribution for the ability parameter θ_i is specified as

$$\theta_{ij} \mid \boldsymbol{\beta}_j \quad \sim \quad \mathcal{N}\left(\mathbf{x}_{ij}^t \boldsymbol{\beta}_j, \sigma_\theta^2\right), \tag{8}$$

and the level-2 covariates are denoted as $\mathbf{w}_{qj}$ for $q = 0, \ldots, Q$, such that the level-2 prior is specified as

$$\boldsymbol{\beta}_j \quad \sim \quad \mathcal{N}\left(\mathbf{w}_j \boldsymbol{\gamma}, \mathbf{T}\right), \tag{9}$$

An inverse-gamma prior distribution and an inverse-Wishart prior distribution is specified for the variance components σ_θ^2 and $\mathbf{T}$ respectively. The extension to more levels is easily made. The IRT measurement model with a multilevel population model for the ability parameters is called a multilevel IRT model (MLIRT). An MCMC algorithm can be used to concurrently estimate all model parameters (e.g., Fox, 2007; Fox and Glas, 2001).

Several advantages can be given of the MLIRT modeling framework. The multilevel population model parameters are estimated from the item response data without having to condition on estimated ability parameters. In

empirical multilevel studies, estimated ability parameters are often considered to be measured without an error and treated as an observed outcome variable. Ignoring the uncertainty regarding the estimated abilities may lead to biased parameter estimates and the statistical inference may be misleading. The modeling framework allows the incorporation of explanatory variables at different levels of hierarchy. The inclusion of explanatory information can be important in various situations. The use of explanatory information may lead to more accurate item parameter estimates. Another related advantage of the model is that it can handle incomplete data in a very flexible way.

4 The Use of Response Times as Collateral Information

Bassili and Fletcher (1991) introduced a methodology for measuring accurately response time within the context of a telephone survey. They showed how response times can be measured precisely and reliably and that the data from such measurement offer insight into the processes underlying survey responses applied to most types of computer-assisted telephone interviewing (CATI). Nowadays, computer-based testing has become a popular mode for retrieving information and response times can be collected automatically. In educational research, when the test takers' responses as well as their response times on the items are recorded, the relationship between response times and response accuracies can be explored. This relationship is complex at different hierarchical levels and it takes the form of a tradeoff between speed and accuracy at the level of a fixed person but may become a positive correlation for a population of test takers. The response times can provide information about the items' characteristics and the individual response process. More specific, they can be used to identify bad items using, for example, the average response times as an indicator of question problems, and they may serve as indicators of uncertainty and response error. This way, the response times contain information about the item and person characteristics.

A log-normal distribution is used taking account of the natural lower-bound at zero to model the response times. Each respondent complete the items at a certain level of speed denoted as ζ_i. The time needed to complete an item k also depends on item characteristic parameters. They are denoted as ϕ_k and λ_k, and can be seen as a discrimination and time-intensity parameter, respectively. The log of the response time, $\log T_{ik}$, is normal distributed with mean $-\phi_k\zeta_i + \lambda_k$ and variance σ_t^2. It follows that

$$P\big(t_{ik} \leq t'_{ik}\big)\big) \quad = \quad \Phi\left(\left(\log t'_{ik} - \left(\lambda_k - \phi_k\zeta_i\right)\right)/\sigma_t\right), t'_{ik} > 0. \qquad (10)$$

Hence, increasing the time intensity λ_k leads to a positive shift of the location of the time distribution on the item. Likewise, an increase in the speed parameter ζ_i leads to a negative shift.

Assume the normal ogive response model, Equation (1), for the response data for measuring the ability levels. At the individual level, a bivariate normal distribution is defined for the ability and speed parameter of the test taker,

$$\begin{pmatrix} \theta_i \\ \zeta_i \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} \mu_\theta \\ \mu_\zeta \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & \rho \\ \rho & \sigma_\zeta^2 \end{pmatrix}\right), \tag{11}$$

where parameter ρ denotes the covariance between the person parameters. This population distribution reflects the individual speed and ability levels in the population of test takers. This conjoint Bayesian IRT model constitutes a two-parameter normal ogive, Equation (1), for the multivariate response patterns and a normal model for the multivariate log transformed response times, Equation (10), and, at a lower level, a multivariate normal model is specified for the underlying ability and speed parameters, Equation (11). The model enables the simultaneous analysis of speed and accuracy given response times and patterns. More about MCMC methods for parameter estimation can be found in Fox, Klein Entink, van der Linden (2007), and van der Linden (2007) that also includes a detailed description of the model.

5 Asking Sensitive Questions

Survey researchers that are dealing with sensitive topics are often confronted with misreporting of respondents leading to biased estimates. The sensitive questions asked in the survey may lead to social desirable response behavior where respondents edit the information they report to avoid embarrassing themselves. Sensitive questions can also be seen as intrusive by the respondents or raise concerns about the possible repercussions of disclosing the information. The extent of misreporting depends on the design of the survey and whether the respondent has anything embarrassing to report. Asking sensitive questions in survey research usually affects the response rates, the item nonresponse rates, and the response accuracy.
Self-administration, collecting the data in private, and confidentiality assurances are several design features that positively influences the accuracy of reports on sensitive topics. Warner (1965) introduced the randomized response technique for improving the accuracy of estimates from survey data. A popular variation on Warner's method is the unrelated-question method of Greenberg, Abu-Ela, Simmons, and Horvitz (1969) where the essential idea is that the interviewer is unaware whether the respondent is answering the sensitive question or the non-sensitive question. A randomizing device (dice, coin) is used such that with probability p_1 a respondent is confronted with the sensitive question. The univariate randomized response technique enables the computation of (aggregated) estimated proportions without revealing the significance of the individual answers.

Fox (2005) introduced a multivariate randomized response technique with the two-parameter normal ogive model in Equation (1) as the response model for the randomly selected question. Let Y_{ik} and $\tilde{Y}_{ik}$ denotes the observed randomized response and the latent response to the sensitive question, respectively, of respondent i to item k. The probability of observing a positive randomized response equals,

$$\begin{aligned} P\left(Y_{ik}=1 \mid \theta_i, \boldsymbol{\xi}_k\right) &= p_1 P\left(\tilde{Y}_{ik}=1 \mid \theta_i, \boldsymbol{\xi}_k\right)+(1-p_1)p_{2,i} \\ &= p_1 \Phi\left(a_k\theta_i - b_k\right)+(1-p_1)p_{2,i}, \end{aligned} \tag{12}$$

where $p_{2,i}$ denotes the known probability of a positive response to the non-sensitive question. In an alternative method the response to the non-sensitive question is simulated via a randomizing device that determines the respondent's answer and $p_{2,i}$ is defined by the properties of the randomizing device.

The multivariate randomized response model makes it possible to measure the underlying sensitive behavior θ_i of the respondents. At a lower level, the underlying sensitive behavior of the respondent can be related to other respondent or group characteristics. Therefore, assume the structural multilevel model for θ_{ij}, Equation (8) and (9). The likelihood of interest of $\boldsymbol{\Omega}=(\boldsymbol{\xi}, \sigma_\theta^2, \boldsymbol{\gamma}, \boldsymbol{T})$ given the randomized response data can be expressed as

$$\begin{aligned} p\left(\boldsymbol{y} \mid \boldsymbol{\Omega}\right) = \prod_{j=1}^{J} \Bigg[\int \Bigg[\prod_{i=1|j}^{n_j} \int \prod_{k=1}^{K} \Big[& \left[p_1 \Phi\left(a_k\theta_{ij} - b_k\right)+(1-p_1)p_{2,i}\right]^{Y_{ijk}} \\ & \left[p_1\left(1-\Phi\left(a_k\theta_{ij} - b_k\right)\right)+(1-p_1)\left(1-p_{2,i}\right)\right]^{1-Y_{ijk}} \Big] \\ & p\left(\theta_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{\beta}_j, \sigma_\theta^2\right) d\theta_{ij} \Bigg] p\left(\boldsymbol{\beta}_j \mid \boldsymbol{w}_j, \boldsymbol{\gamma}, \boldsymbol{T}\right) \Bigg] d\boldsymbol{\beta}_j. \end{aligned}$$

MCMC methods makes it possible to estimate simultaneously all parameters (Fox, 2005).

6 Conclusions

The Bayesian IRT framework provides a set of powerful tools for the analysis of large-scale complex survey data. The Bayesian IRT model can be extended in different ways to handle measurement issues involved in large-scale survey research. It is shown that the population distribution of the respondents is easily extended to take account of a nested structure. Additional information can be incorporated via prior specifications. The framework can also be extended to a multivariate framework with different link functions to relate multivariate discrete and/or continuous observations with multiple underlying latent variables. This makes it possible to conduct

a simultaneous analysis of multiple tests each measuring a different latent variable with an underlying correlation structure. The framework can handle different complex sampling strategies to collect reliable response data which includes the randomized response sampling design.
MCMC methods can be used to estimate simultaneously the Bayesian IRT model parameters. The MCMC estimation methods make it possible to add additional complexity in a straightforward way. This includes the specification of different priors, constraints on parameters, and different distributional assumptions. The powerful estimation methods can also be used for the computation of a Deviance Information Criteria that can be used for comparing the fit of different Bayesian IRT models.

References

Bassili, J.N., and Fletcher, J.F. (1991). Response-time measurement in survey research. A method for CATI and a new look at nonattitudes. *The Public Opinion Quarterly*, **55**, 331-346.

Greenberg, B.G., Abul-Ela, A., Simmons, W.R., and Horvitz, D.G. (1969). The Unrelated Question Randomized Response Model: Theoretical Framework. *Journal of the American Statistical Association*, **64**, 520-539.

Fox, J.-P. (2005). Randomized item response theory models. *Journal of Educational and Behavioral Statistics*, **30**, 189-212.

Fox, J.-P. (2007). Multilevel IRT modeling in practice. *Journal of Statistical Software*, **20**, Issue 5.

Fox, J.-P., and Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, **66**, 269-286.

Fox, J.-P., Klein Entink, R.H., and van der Linden, W.J. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software*, **20**, Issue 7.

Lord, F.M., and Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

van der Linden, W.J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, **72**, 287-308.

Warner, S.L. (1965). Randomized Response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **60**, 63-69.