Statistics
in Medicine

Tutorial in Biostatistics

Received 17 January 2011, Accepted 1 November 2011 Published online 23 February 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4475

Using R and WinBUGS to fit a generalized partial credit model for developing and evaluating patient-reported outcomes assessments

Yuelin Li[a,b*†] **and Ray Baser**[b]

The US Food and Drug Administration recently announced the final guidelines on the development and validation of patient-reported outcomes (PROs) assessments in drug labeling and clinical trials. This guidance paper may boost the demand for new PRO survey questionnaires. Henceforth, biostatisticians may encounter psychometric methods more frequently, particularly item response theory (IRT) models to guide the shortening of a PRO assessment instrument. This article aims to provide an introduction on the theory and practical analytic skills in fitting a generalized partial credit model (GPCM) in IRT. GPCM theory is explained first, with special attention to a clearer exposition of the formal mathematics than what is typically available in the psychometric literature. Then, a worked example is presented, using self-reported responses taken from the international personality item pool. The worked example contains step-by-step guides on using the statistical languages R and WinBUGS in fitting the GPCM. Finally, the Fisher information function of the GPCM model is derived and used to evaluate, as an illustrative example, the usefulness of assessment items by their information contents. This article aims to encourage biostatisticians to apply IRT models in the re-analysis of existing data and in future research. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: item response theory; generalized partial credit model; Rasch model

1. Introduction

The creation and validation of a new PRO questionnaire typically requires a team of experts, among them behavioral scientists who analyze the questionnaire data. However, the analytic tasks may become the responsibility of the biostatistician if a behavioral scientist is not available or if the behavioral scientist has limited training in the more sophisticated item response theory (IRT) methods [1–5]. Recent publications on health-related quality of life (HRQOL) assessments indicate the increasing use of IRT as the primary statistical method in developing and evaluating PRO instruments [6–8]. IRT models have also been used to address novel research questions in medicine and bioinformatics (e.g., [9] and [10]). Biostatisticians can easily reach new research territories by incorporating IRT to their data-analytic repertoire. This tutorial aims to provide a step-by-step guide in carrying out basic IRT analyses using freely available software programs R [11] and WinBUGS [12].

In December 2009, the US Food and Drug Administration (FDA) announced the final guidance on using PROs as part of clinical trials and for drug labeling [13], 3 years after the announcement of a draft guidance [14]. In 2005, the European Medicines Agency published an European guidance document in the evaluation of medical products in cancer by PROs [15]. The FDA guidelines are the regulatory agency's attempt to standardize the procedures in the creation, refinement, validation, and clinical use of

[a]*Department of Psychiatry & Behavioral Sciences, Memorial Sloan-Kettering Cancer Center, 641 Lexington Ave. 7th Floor, New York, NY 10022, USA*
[b]*Department of Epidemiology & Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA*
**Correspondence to: Yuelin Li, Department of Psychiatry & Behavioral Sciences, Memorial Sloan-Kettering Cancer Center, 641 Lexington Ave. 7th Floor, New York, NY 10022, USA.*
†*E-mail: liy12@mskcc.org*

Contract/grant sponsor: 2007 Prevention Control and Population Research Program (PCPR) Goldstein Award (Li); NIH Training grant T32CA009461; NIH CTSC Award to Weill Cornell Medical College, NIH UL1-RR024996

psychometric instruments in clinical trials and drug labeling. These guidelines aim to provide practical advice to researchers, including how to define the PRO domain(s) to be measured (e.g., Section III.C on conceptualizing the PRO constructs of interest), how to write survey items, how to decide what response format is most appropriate, and how to evaluate patient understanding. As part of 'item analysis', survey items may be deleted or modified in response to patient understanding and preliminary data analysis. This tutorial focuses primarily on item analysis from a Bayesian IRT perspective. Item analysis is typically iterative—several revisions may be required before the draft survey instrument is finally deemed valid (as per Section III.E), reliable, and responsive to changes in well-being, just to name a few of FDA's recommendations. However, the FDA guidelines offer no explicit recommendations on how to carry out these analyses, although the conceptual diagram hints at a factor analysis framework [13, Figure 4]. Most of our biostatistician colleagues are familiar with factor analysis but not with IRT. This motivated our writing of this article to cover item analysis using IRT.

Item response theory modeling is also useful in its own right. For example, IRT has been applied in analyzing inter-rater agreement data in rating the severity of hip fractures [9], and in microarray gene expression analysis to identify clusters of genes that are related to drug response in acute leukemias [10], Extensions of the classical Rasch model (RM) [16] have also been applied in identifying clusters of students with discrete levels of latent academic achievements. More generally, the RM is closely related to the conditional logit model [17] and the conditional logistic regression model for binary matched pairs [18, Chapter 10]. Despite their versatility, IRT models have yet to gain wider use in biostatistics. This is in part because the command syntax of popular IRT software programs can be arcane for new users (for a list of packages, see [2]). The occasional user of IRT may be hesitant in investing the time and effort in learning it. We hope to facilitate the use of IRT models in this tutorial—a distillation of the cited sources into a practical guide using freely available statistical programming languages so that the readers can immediately apply these analytic skills in their own research.

Our primary goal is to guide the readers in applying their Bayesian analytic skills to a previously unfamiliar area of statistics. (Thus, we provide details on how an IRT model is derived.) We also hope that this article is equally useful to statisticians who are quite familiar with IRT and/or psychometrics but are new to a Bayesian analytic approach to IRT modeling. (Thus, we provide details on Bayesian computation.) The overall plan is to provide enough mathematics in both IRT model derivations and Bayesian computing so that they can be quickly deployed in practice. Muraki [19] provided a worked example on the generalized partial credit model (GPCM). The GPCM is among several commonly used models in analyzing items with polytomous response categories [6]. This article is not about choosing one model out of alternative IRT models. Interested readers can find them elsewhere [20, 21]. The deviance information criterion by Spiegelhalter *et al.* [22], which is calculated as part of default output from R2WinBUGS, is useful in model selection. What we lack in breadth, we hope to compensate for in depth.

We organize this paper as follows. Section 2 covers the theories behind widely used IRT models, including the RM [16] for binary responses and the partial credit model (PCM) [23] and the GPCM [19] for polytomous item responses. In Section 3, we develop the GPCM model from a Bayesian perspective. We do not go into the details on how to manually carry out Gibbs sampling in IRT but provide a list of references for interested readers. Section 4 translates the GPCM model mathematics into WinBUGS syntax. We assume that the readers have gotten to the point of successfully installing R, the R2WinBUGS package in R, and WinBUGS on a computer platform of their choice. In Section 5, we illustrate on how to diagnose the convergence of iterative sampling. Sections 3–5 are the main focus of this paper. They cover, in detail, how to fit the GPCM using R and WinBUGS. Section 6 focuses on the practical aspects of item analysis, on how to decide which questionnaire items should be modified or deleted. Our overall pedagogical plan is to provide enough mathematical rigor on IRT so that readers can acquire a working knowledge of IRT without the need to review the vast psychometric literature spanning several decades. Finally, we discuss in Section 7 how these steps can be used to address the statistical considerations outlined in the FDA guidelines.

2. Models of item response data

2.1. Rasch model for dichotomized items

One of the simplest IRT models is the RM for dichotomized response data, developed by the Danish mathematician Georg Rasch [16]. The RM handles data coded as 'correct'/'incorrect' or 'yes'/'no' with

a value of 1 coding a correct answer or a 'yes' response. The log odds of answering an item correctly is a function of two parameters:

$$\ln\left[\frac{\Pr(y_{ij}=1|\theta_i,\beta_j)}{1-\Pr(y_{ij}=1|\theta_i,\beta_j)}\right]=\theta_i-\beta_j, \tag{1}$$

where $\Pr(y_{ij}=1|\,\theta_i,\beta_j)$ represents the probability of person i scoring a 1 vs. 0 on item j. The interpretation of this model is made clearer if we let θ_i represent person i's innate 'ability' and β_j represent item j's 'difficulty'. If a person's ability matches the difficulty of an item, then he or she has a 50/50 chance of answering the item correctly (assuming no guessing). This interpretation makes intuitive sense in an educational testing setting. Equation (1) can be unpacked by applying the inverse logit:

$$\begin{aligned}\Pr(y_{ij}=1|\theta_i,\beta_j)&=\text{logit}^{-1}(\theta_i-\beta_j)\\&=\frac{\exp(\theta_i-\beta_j)}{1+\exp(\theta_i-\beta_j)} \qquad (2)\\&=\frac{1}{1+\exp(-(\theta_i-\beta_j))}. \qquad (3)\end{aligned}$$

Equation (2) is more commonly used in the literature than the simpler Equation (3) [24].

At approximately the same time in the US, Birnbaum [25] and Lord [3] developed similar models independent of the RM [2, Chapter 1]. Their models contain an additional slope parameter α_j addressing the discriminating power of test items, thus the name two-parameter logistic model.

$$\Pr(y_{ij}=1|\theta_i,\beta_j,\alpha_j)=\frac{\exp[\alpha_j(\theta_i-\beta_j)]}{1+\exp[\alpha_j(\theta_i-\beta_j)]}. \tag{4}$$

Note that the difference between person i's 'ability' level and item j's 'difficulty' level is weighted by α_j. Thus, items with relatively large α_j parameter values provide more precise measurement of θ_i, particularly for values of θ_i near the β_j value. The α_j parameter is interpreted as an index of how strong of an indicator item j is of the latent variable θ. We can view the classical RM in Equation (2) as a special case of the two-parameter logistic model in Equation (4) with fixed $\alpha_j=1$ for all items.

2.2. *Partial credit model and generalized partial credit model for polytomous items*

The RM and two-parameter logistic models are not designed to handle multiple response categories or a mixture of yes/no and multiple responses. Masters [23] proposed a PCM to handle polytomous items by extending the dichotomous RM to K response categories. Master's PCM treats polytomous responses as ordered performance levels, assuming that the probability of selecting the kth category over the $[k-1]$th category is governed by the dichotomous RM. For example, $K=4$ if the responses are 'Strongly Disagree', 'Disagree', 'Agree', and 'Strongly agree' and are scored as 0, 1, 2, and 3, respectively. A person who chooses 'Agree' is considered to have chosen 'Disagree' over 'Strongly disagree' and 'Agree' over 'Disagree' but to have failed to choose 'Strongly agree' over 'Agree'.

In the following discussion, we follow Masters' derivation of the PCM [23, p. 158]. For each of the successive response categories, the probability of endorsing response category k over $k-1$ for item j follows a conditional probability governed by the RM:

$$\Phi_{jk}=\frac{\pi_{jk}}{\pi_{j[k-1]}+\pi_{jk}}=\frac{\exp(\theta_i-\beta_{jk})}{1+\exp(\theta_i-\beta_{jk})}.$$

Thus, $\pi_{jk}=[(\pi_{j[k-1]}+\pi_{jk})\exp(\theta_i-\beta_{jk})]/[1+\exp(\theta_i-\beta_{jk})]$. We can solve for $\pi_{j0},\pi_{j2},\ldots,\pi_{jk}$ (detailed derivations are found in Appendix A). Muraki [19] shows that the G term in the following

equation makes the solutions much easier to track.

$$\text{Let } G = \sum_{k=0}^{m_j} \exp\left[\sum_{h=0}^{k}(\theta_i - \beta_{jh})\right], \text{ we get, in the } K = 4 \text{ example} \tag{5}$$

$$\pi_{j0} = \frac{1}{G},$$

$$\pi_{j1} = \frac{\exp(0) \times \exp(\theta_i - \beta_{j1})}{G},$$

$$\pi_{j2} = \frac{\exp(0) \times \exp(\theta_i - \beta_{j1}) \times \exp(\theta_i - \beta_{j2})}{G}, \text{ and}$$

$$\pi_{j3} = \frac{\exp(0) \times \exp(\theta_i - \beta_{j1}) \times \exp(\theta_i - \beta_{j2}) \times \exp(\theta_i - \beta_{j3})}{G}.$$

These expressions can be interpreted intuitively as though the person 'passes through' each of the preceding response categories before finally stopping at a response [1, p. 165] that, presumably, most accurately reflects that person's standing on the latent variable continuum. The adjacent β_{jk} parameters represent the incremental item 'difficulties' that the person has to step through to reach the next response category. In educational testing, the incremental item difficulties assign credits to partially correct answers, and thus the name *partial credit model*. In HRQOL, symptom severity may be either assessed as 'present' or 'absent' or on a gradation such as 'persistent/intermittent/none'. Each numerator in the PCM equation represents one unique response step, and the denominator G is the sum of all possible steps. In the psychometrics literature, models that conform to these characteristics are named 'divide-by-total' models [26].

The PCM can accommodate items with different response scales in one HRQOL instrument, such as a combination of items assessed as 'present/absent', 'persistent/intermittent/none', or on a four-category rating scale as described previously. To obtain a general notation, we define that item j is scored $y = 0, 1, 2, \ldots, m_j$ with $K_j = m_j + 1$ response categories and the denominator G. A general model expression for $K_j = m_j + 1$ that incorporates all of the aforementioned steps is given in the PCM:

$$\Pr(y_{ij} = y|\theta_i, \beta_{jh}) = \frac{\exp\sum_{h=0}^{k}(\theta_i - \beta_{jh})}{\sum_{k=0}^{m_j}\exp\sum_{h=0}^{k}(\theta_i - \beta_{jh})}, \tag{6}$$

where the numerator is the individual response outcomes and the denominator is the sum of all the possible outcomes; $i = 1, 2, \ldots, N$ refers to individual respondents, N refers to total number of respondents in the sample, $j = 1, 2, \ldots, J$ refers to items, and $h = 1, 2, \ldots, k$ refers to the number of response categories.

Muraki [19] further extended the PCM into the GPCM by introducing a discrimination parameter α_j for each item.

$$\Pr(y_{ij} = y|\theta_i, \alpha_j, \beta_{jh}) = \frac{\exp\sum_{h=0}^{k}\alpha_j(\theta_i - \beta_{jh})}{\sum_{k=0}^{m_j}\exp\sum_{h=0}^{k}\alpha_j(\theta_i - \beta_{jh})} \tag{7}$$

The PCM is a special case of the GPCM where all $\alpha_j = 1$. The parameters θ_i, β_{jh}, and α_j are respectively interpreted as a person's underlying health status or quality of life, the inherent health status or quality of life intensity measured by the response category thresholds, and an item's ability in discriminating between persons with different underlying health status.

3. Generalized partial credit model from a Bayesian perspective

To further reduce notation, we denote the observed item response vector as $\boldsymbol{y} = (y_{i1}, y_{i2}, \ldots, y_{iJ})$, the vector of ability parameters as $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)$, and the vector of parameters for the jth item as

$\boldsymbol{\alpha} = \alpha_j$ and $\boldsymbol{\beta} = \beta_{jh}$. The likelihood of observing response vector $\boldsymbol{y}$ is

$$\ell(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_i \prod_j \prod_h \Pr(y_{ij}|\theta_i, \alpha_j, \beta_{jh}), \tag{8}$$

where $\boldsymbol{\theta}$ is a random sample from $N(\mu, \sigma^2)$ [27, Chapter 16], with hyperprior distributions $\mu \sim N(\mu_0, \sigma_0^2)$ and $\sigma^2 \sim$ Inv-Gamma(a_0, b_0). Model (7) is not identified. For example, a constant can be added to the value of α_j or β_{jh} without affecting the model's prediction. There is a long tradition to set $N(\mu, \sigma^2)$ to $N(0, 1)$ to simplify model fitting [5, 28, 29] and to address the issue of model identifiability. Another common approach to solve model indeterminacy is by centering the β_{jh} parameters by their average so that they sum to zero [28, p. 450]. These constraints are not always explicitly stated in the documentations, causing potential confusions in different estimated parameter values by different software packages.

Bafumi *et al.* [30] proposed an unconventional method to identify model parameters. Estimated parameters are normalized as follows after the estimation is complete. This method has the advantage that sample-specific distribution of $\boldsymbol{\theta}$ can be derived for future references (e.g., difference between cohorts of respondents).

1. θ_i's are normalized to have mean 0 and standard deviation 1. The normalized θ_i's are named θ_i^{adj},
2. β_{jk}'s are normalized by the mean and standard deviation of θ_i to retain a common scale for all parameters. The normalized β's are named β_{jk}^{adj}, and
3. The multiplicative α_j's are multiplied by the standard deviation of θ_i to retain a common scale, The normalized α_j's are named α_j^{adj},

hence retaining $\Pr(y|\theta^{\text{adj}}, \beta^{\text{adj}}) = \log^{-1}\left[\sigma_\theta\, \alpha\left(\frac{\theta-\mu_\theta}{\sigma_\theta} - \frac{\beta-\mu_\theta}{\sigma_\theta}\right)\right] = \log^{-1}[\alpha(\theta - \beta)] = \Pr(y|\theta, \beta)$ (subscripts omitted to simplify notations).

3.1. *Priors, hyperpriors, joint posterior density, and conditional posterior densities*

The prior distributions are as follows:

$$\begin{aligned}
&\boldsymbol{\alpha} \sim \log N(\mu_\alpha = 0.0, \sigma_\alpha^2 = 1.4),\\
&\boldsymbol{\beta} \sim N(\mu_\beta = 0, \sigma_\beta^2 = 6.25),\\
&\boldsymbol{\theta} \sim N(\mu, \sigma^2), \quad \text{with hyperpriors}\\
&\qquad \mu \sim N(\mu_0 = 0.0, \sigma_0^2 = 100); \quad \sigma^2 \sim \text{Inv-Gamma}(\text{shape} = 0.5, \text{rate} = 0.5),
\end{aligned}$$

where N and $\log N$ represent the normal and the log-normal distributions, respectively. Our prior knowledge about the unknown $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ parameters is represented in relatively noninformative prior and hyperprior distributions. Baldwin *et al.* [9] previously used the log-normal distribution for $\boldsymbol{\alpha}$ to represent common findings that $\boldsymbol{\alpha}$ is non-negative and typically near 1.0. The distribution of the latent characteristics $\boldsymbol{\theta}$ is not set to a standard normal. Instead, we set hyperpriors for μ and σ^2 so that they center on the standard normal. The hyperprior for μ is noninformative. The hyperprior for σ^2 follows an inverse gamma distribution with a mean of 1.0 and a variance of 2 (when shape $= 0.5$ and rate $= 0.5$), which places the bulk of parameter values within the typically observed $[-3, +3]$ bounds. It also minimizes the risk of floating point overflow during model estimation by WinBUGS.

The joint posterior distribution is as follows:

$$\begin{aligned}
p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{y}) &\propto p(\boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})\\
&= p(\boldsymbol{\alpha}) p(\boldsymbol{\beta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{\beta}) \ell(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}),
\end{aligned}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are assumed independent. The next steps are to derive the conditional posterior distributions for the $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ parameters from the joint posterior distribution. By way of an example, we begin the derivations with $p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{y}) / p(\boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ by the axioms of conditional probability [31, p. 3]. However, this remains hard to track. Note that $p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

Thus, $p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\beta}) = p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta})p(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}) / p(\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\beta})$. If we assume *a priori* independence of $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$, we may write this as

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\beta}) &\propto p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta})p(\boldsymbol{\theta})p(\boldsymbol{\alpha})p(\boldsymbol{\beta}) \\ &\propto \ell(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta})p(\boldsymbol{\theta}), \end{aligned}$$

because the marginal $p(\boldsymbol{\alpha})$ and $p(\boldsymbol{\beta})$ distributions for fixed $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are constants. Note that the conditional posterior is proportional to prior times likelihood of the data—the familiar form of Bayesian inference [31, p. 32]. The conditional posterior distributions for the $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ parameters, respectively, are as follows:

$$p(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{\alpha},\boldsymbol{\beta}) \propto p(\boldsymbol{\theta})\ell(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}), \tag{9}$$

$$p(\boldsymbol{\alpha}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\beta}) \propto p(\boldsymbol{\alpha})\ell(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}), \tag{10}$$

$$p(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\alpha}) \propto p(\boldsymbol{\beta})\ell(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}). \tag{11}$$

Sampling from the conditional distributions in Equations (9)–(11) may be carried out using a rejection sampling with Gibbs approach, which is the method implemented in the WinBUGS software program [12] or a Metropolis–Hastings within Gibbs algorithm (e.g., [32, 33]). Readers interested in further technical details may follow-up with other published work (e.g., [20, 29, 32–34] and [35, p. 444]). The next section provides a step-by-step guide for using R and WinBUGS to estimate the GPCM parameters.

4. Steps to fit a generalized partial credit model by R and WinBUGS

4.1. Neuroticism dataset

The dataset is the `bfi` dataset in the R package `psych` [36], which contains the responses of 2800 subjects to 25 personality self-report items. The items map onto five putative psychological dimensions, the 'Big Five' personality traits [37]: agreeableness, conscientiousness, extraversion, neuroticism, and openness. This section focuses on the five items assessing neuroticism, a tendency to easily experience anger ('1. Get angry easily'), unpleasant affect/emotion ('2. Get irritated easily' and '3. Have frequent mood swings'), depression ('4. Often feel blue'), and anxiety ('5. Panic easily').

The response categories were '1. Very inaccurate', '2. Moderately inaccurate', '3. Slightly inaccurate', '4. Slightly accurate', '5. Moderately accurate', and '6. Very accurate'. The `bfi` dataset was chosen because it is suitable to examine key aspects of IRT modeling. IRT makes inferences on an unobservable underlying psychological construct, in this case a neuroticism personality trait, on the basis of each person's responses to a few straightforward items. The response categories describe what the person is like generally and do not depend on environmental, social, and interpersonal contexts. The observed responses may be viewed as a self-reported symptom severity—a higher summation score representing more severe neuroticism. IRT goes beyond simple summary scores and takes into consideration, for each item, its symptom severity threshold (β_{jh}) and sensitivity to changes in neuroticism (α_j).

4.2. R syntax

We need two syntax files, one for R and one for WinBUGS. The R syntax file prepares the data for fitting the model in the WinBUGS syntax file. The R syntax is in Listing 1. Lines 1–3 load the required packages into R. Additional required packages are loaded automatically (e.g., packages `coda` is required by `R2WinBUGS`). Lines 6 and 7 extract only the neuroticism items. Line 9 shows that we use data from the first 500 subjects and convert them into a matrix. Lines 10–12 specify the number of subjects, the total number of items, and the response category for each item. Lines 13–18 set up the data and names of parameters to be fitted by WinBUGS. Then, `bugs()` is called to pass the data and the Gibbs sampler parameters to WinBUGS for analysis. Lines 24 and 31 calculate the elapsed time. The `bugs()` function needs to know the name of the WinBUGS syntax file and settings for the iterative chains. WinBUGS generates random initial values if the `init` option is set to `NULL`. Here, we set `codaPkg = TRUE` to save the iterative chains for further analyses by the `coda` package. Line 28 sets `bugs.seed = 7` for reproducibility. The results are returned back to R as `neuro.bugs`.

Listing 1. R syntax.

```
 1 library(ltm)
 2 library(psych)
 3 library(R2WinBUGS)
 4
 5 ### Extracting neuroticism items from bfi data (in library(psych))
 6 data(bfi)
 7 neuroticism <- as.data.frame(bfi[,16:20 ])
 8 ###
 9 Y <- matrix(unlist(neuroticism[1:500,]), nrow = 500)
10 n <- nrow(Y)
11 p <- ncol(Y)                          # number of items
12 K <- apply(Y, 2, max, na.rm = TRUE) # response categories for each item
13 m.alpha <- 0.0
14 s.alpha <- 1.2
15 m.beta <- 0
16 s.beta <- 2.5
17 data <- list("Y", "n", "p", "K", "m.alpha", "s.alpha", "m.beta", "s.beta")
18 param <- c("alpha", "beta", "mu0", "var0")
19 ###
20 n.burn <- 2000
21 n.thin <- 15
22 n.sim <- 10000 * n.thin + n.burn
23 # timing the bugs() iterative simulation
24 pr.time <- proc.time()[1:3]
25 neuro.bugs <- bugs(data = data, inits = NULL, parameters = param,
26     model.file="neuro.txt",
27     codaPkg = TRUE,
28     bugs.seed = 7,
29     n.burnin = n.burn, n.thin = n.thin, n.iter = n.sim, n.chains = 3,
30     debug = FALSE)
31 pr.time <- proc.time()[1:3] - pr.time
32 print(pr.time)
33 show(neuro.bugs)
34 [1] "/tmp/Rtmp1jVqO0/coda1.txt" "/tmp/Rtmp1jVqO0/coda2.txt"
35 [3] "/tmp/Rtmp1jVqO0/coda3.txt"
```

A total of 152,000 iterative simulations were carried out per chain, the first 2000 iterations discarded, and 10,000 iterations saved after thinning. This long chain was guided by convergence diagnostics (described in Section 5). One advantage of using the `R2WinBUGS` package is that it automatically prints out two useful statistics for each parameter estimate: (1) the effective sample size (sample size adjusted for autocorrelation across simulations); and (2) the Gelman and Rubin convergence diagnostic [38], shown as `Rhat`. Values of `Rhat` close to 1 indicate convergence [39–41].

4.3. WinBUGS syntax

Listing 2 shows the WinBUGS syntax. Lines 1–15 are adapted from Curtis [21]. Line 4 shows that the item responses can be one of `K[j]` possible values, with the probability of each response separately specified. Line 6 samples each person's latent characteristic from $N(\mu, \sigma^2)$ with hyperpriors $\mu \sim N(\mu_0 = 0.0, \sigma_0^2 = 100)$ and $\sigma^2 \sim$ Inv-Gamma(shape $= 0.5$, rate $= 0.5$), specified by lines 25 and 26, respectively.

Lines 11–13 calculate the numerators in the GPCM model in Equation (7). The denominator is harder to track. Recall that in Equation (5), the denominator $G = 1+\delta_1+\delta_2\delta_1+\delta_3\delta_2\delta_1+\cdots+\delta_{m_j}\delta_{m_j-1}\cdots\delta_1$. Thus, line 13 yields δ_1 when `k = 1`, $\delta_2\delta_1$ when `k = 2`, and so on for $\delta_{m_j}\delta_{m_j-1}\cdots\delta_1$ when `k = K[j]`. Line 14 is the WinBUGS representation of Equation (7). Line 17 specifies prior $\boldsymbol{\alpha}$ with a log-normal distribution with mean `m.alpha` and precision `pr.alpha`. Line 20 specifies prior $\boldsymbol{\beta}$ with a normal distribution with mean `m.beta` and precision `pr.beta`. Lines 23 and 24 calculate the precision parameters from the inverse of variance parameters.

Listing 2. WinBUGS syntax for the GPCM.

```
 1 model {
 2   for (i in 1:n) {
 3     for (j in 1:p) {
 4        Y[i, j] ~ dcat(prob[i, j, 1:K[j]])
 5        }
 6        theta[i] ~ dnorm(mu0, tau0)
 7   }

 8   for (i in 1:n) {
 9     for (j in 1:p) {
10       for (k in 1:K[j]) {
11         eta[i, j, k] <- alpha[j] * (theta[i] - beta[j, k])
12         psum[i, j, k] <- sum(eta[i, j, 1:k])
13         exp.psum[i, j, k] <- exp(psum[i, j, k])
14         prob[i, j, k] <- exp.psum[i, j, k] / sum(exp.psum[i, j, 1:K[j]])
15   } } }

16   for (j in 1:p) {
17     alpha[j] ~ dlnorm(m.alpha, pr.alpha)
18     beta[j, 1] <- 0.0
19     for (k in 2:K[j]) {
20         beta[j, k] ~ dnorm(m.beta, pr.beta)
21     }
22   }

23   pr.alpha <- pow(s.alpha, -2)
24   pr.beta  <- pow(s.beta, -2)

25 mu0 ~ dnorm(0, 0.01)
26 tau0 ~ dgamma(0.5, 0.5)
27 var0 <- 1/tau0
28 }
```

Table I shows that the mean parameter estimates by WinBUGS agree well with those calculated by maximum likelihood estimation.

Figure 1 is an empirical quantile–quantile plot of the β_{jh} estimates derived from the maximum likelihood method in `gpcm()` against the Bayesian estimates. The two distributions agree well, except that, for values outside $[-1.5, +1.5]$, the Bayesian estimates are distributed slightly closer to the mean than the maximum likelihood estimates. The small differences between the two sets of estimates disappear if we set $\boldsymbol{\theta} \sim N(0, 1)$ (not shown), suggesting that the normalization scales the raw parameter values slightly toward μ.

Table I. Parameter estimates by WinBUGS and the `gpcm()` function in `library(ltm)`.

	α_j		b_{j1}		b_{j2}		b_{j3}		b_{j4}		b_{j5}	
Item	W	g	W	g	W	g	W	g	W	g	W	g
1. Angry	1.576	1.589	−0.668	−0.665	0.109	0.125	−0.052	−0.008	1.027	1.055	1.630	1.694
2. Irritated	1.970	1.969	−1.468	−1.476	−0.516	−0.502	−0.357	−0.319	0.514	0.544	1.355	1.400
3. Swings	0.958	0.931	−1.207	−1.223	0.468	0.513	−0.548	−0.538	0.996	1.038	1.533	1.616
4. Blue	0.438	0.417	−1.706	−1.906	0.985	1.219	−0.741	−0.905	1.368	1.467	1.965	2.236
5. Panic	0.458	0.448	−0.656	−0.723	0.960	1.099	−0.429	−0.534	1.379	1.459	1.811	2028

W, WinBUGS parameter estimates; g, `gpcm()` function in `library(ltm)`.

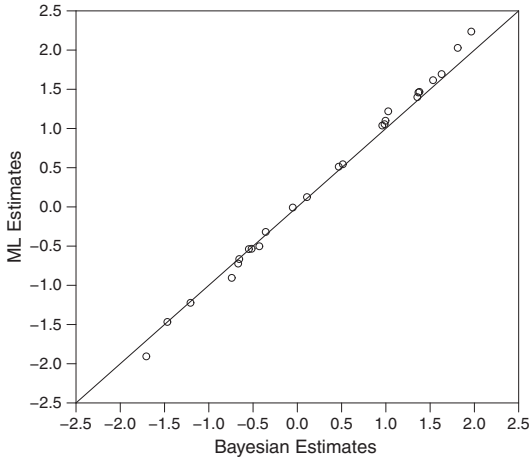

Figure 1. Quantile–quantile plot of parameter estimates by `gpcm()` using maximum likelihood estimation against the Bayesian estimates. The two distributions show good overall agreement, as seen in most of the dots falling on the diagonal line. The Bayesian method appears to underestimate the magnitude of more extreme parameters values outside [-1.5, +1.5].

5. MCMC diagnostics

Several diagnostics for the MCMC chains can be calculated by the `coda` package (or `boa`). The analyst is advised to consult several diagnostics because each has strengths and weaknesses, as discussed in review papers by Brooks and Robers [42] and Cowles and Carlin [43].

The menu-driven tool `codamenu()` calculates some of them, including the diagnostics by Gelman–Rubin [38], Geweke [44], Heidelberger–Welch [45], and the Raftery–Lewis [46] run-length estimate.

The Gelman–Rubin diagnostic requires parallel chains from dispersed initial values. The idea is to compare the between-chain and within-chain variabilities. If all chains converge to a similar posterior distribution, then the between-chain variability would be small relative to the within-chain variability. A ratio is calculated from a combined variability estimate (a weighted average of between-chain and within-chain variabilities) and the within variability alone. The square root of this ratio is the *potential scale reduction factor*; a longer simulation is indicated with a scale reduction value greater than 1.0. The Geweke diagnostic tests whether the early and latter iterative sequences (defaults are first 10% and the latter 50%) are comparable in their averages. A z-statistic is calculated, a large value (e.g., out of the ± 2 bounds) indicate the need for a longer chain.

The Heidelberger–Welch diagnostics include a *stationary* and a *half-width* test. The stationary test uses the Cramér–von Mises statistic to successively test the null hypothesis that the sampled values come from a covariance stationary process. The whole chain is first used to calculate the Cramér–von Mises statistic. If it passes the test (null hypothesis not rejected), then the whole chain is considered stationary. If it fails the test, then the initial 10% of the chain is dropped and tested again. Then, 20% is dropped and so on, until either the chain passes the test (e.g., at 30% reduction) or the remaining data are no longer sufficient (default to 50%).

The half-width test is then applied to the part of the chain that is deemed stationary. It tests whether or not the chain provides enough data to determine the confidence interval of the mean estimate to within a specific accuracy. The accuracy measure is the ratio of half of the width of the 95% confidence interval to the mean estimate. The default accuracy is 0.1 of the accuracy of the 95% confidence interval.

The Raftery–Lewis estimate focuses on achieving a prespecified precision of specific quantiles of a chain. The default is that the 2.5th percentile of a parameter estimate must be within a precision of 0.005 quantile units with 0.95 probability. The output reports the minimum length of the burn-in period (M), the estimated number of post burn-in iterations required to meet the criteria (N), the minimum number of post burn-in iterations ($N_{\min}$, assuming zero autocorrelation), and the *dependence factor* I, which is $N/N_{\min}$, the relative increase in total number of iterations due to autocorrelation. Jackman [40, Section 6.2] observes a few problems with the Raftery–Lewis estimate. It can be conservative. It can run into a confusing and frustrating cycle—another, much longer run length is prescribed after the analyst has carried out a run length exactly as prescribed previously.

Our WinBUGS chains have generally exceeded the Raftery–Lewis burn-in length and post burn-in iterations, with the obvious exceptions of `mu0` and `var0`. The estimated total run lengths are markedly reduced to 10,350 and 12,714, respectively, by lowering the default 0.025 quantile to 0.02 (Details of diagnostics available upon request). Moreover, `mu0` and `var0` pass the Gelman–Rubin, the Geweke, and the Heidelberger–Welch diagnostics. We suspect that the present run length is satisfactory after all, given Jackman's observation [40, Section 6.2] on the problems with the Raftery–Lewis, the passing of the three other diagnostics, and the markedly reduced run length estimates when the default quantile is slightly lowered.

6. Basic item analysis and selection

An item analysis helps determine the final length and response format of a new PRO instrument. The process can be rather detailed [47]. The FDA guidance provides no specific recommendations on how to carry it out. An important goal, however, is to reduce the draft questionnaire to a shorter version with carefully crafted items that are selectively sensitive to a wide range of latent θ. This goal can often be adequately addressed by visual inspections of IRT model properties. Two such properties are covered in this section: (1) fitted response probability curves; and (2) Fisher information curves with respect to the latent $\boldsymbol{\theta}$. These basic visual explanations are often the first steps after estimating the IRT parameters and are part of popular introductory texts [1, 2, 48]. More rigorous methods are usually found in journal articles [49, 50].

6.1. Category response curves

Figure 2 shows an example of marginal posterior density curves for all items, across a range of $\boldsymbol{\theta}$ values. The R commands used to plot the curves are available upon request.

These 'category response curves' [2] are often used to evaluate the need for item and scale reduction. For example, there is a visible overlap between response categories 2, 3, and 4 for all items ('2.

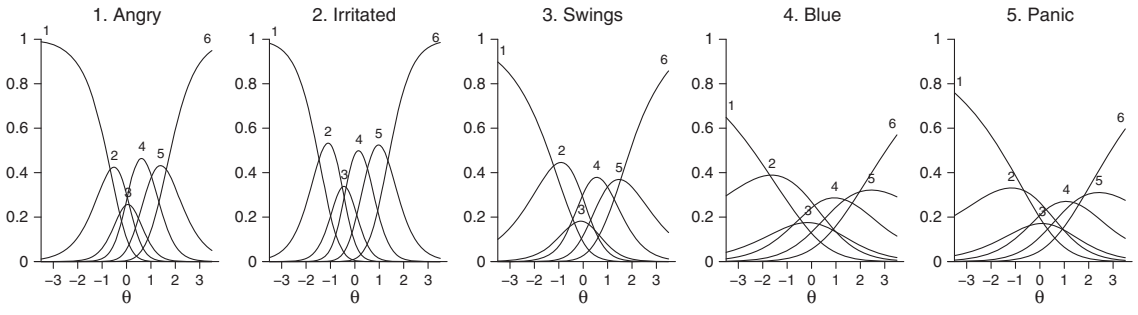

Figure 2. The marginal posterior density curves of all five items. They are more commonly referred to as 'category response curves'. They are derived from plugging in the WinBUGS parameter estimates into Equation (2) over θ values from -3.5 to $+3.5$. The response category 3 has the lowest density peak among all other curves, indicating that it is rarely endorsed. The overlapping curves between response categories indicate that some may be merged into one in the next iteration of instrument development.

Moderately inaccurate', '3. Slightly inaccurate', and '4. Slightly accurate', respectively). The parameter estimates are also very similar in value (Table I). These response categories may be merged into one in the next version of the instrument. Another noteworthy pattern is that the probability of endorsing item 4 'feeling blue', an indication of depressive symptoms, is lower than in other items. Low endorsement probability is not necessarily a problem, because, in this case, we know that depression symptoms are not prevalent.

Generally, the category response curves would preferably cover a range of values on the latent characteristic. Low peaks indicate low probability of endorsement and relatively poorer α_j discrimination parameters. They need to be further explored. For example, there is an approximately 1 in 5 probability of responding a 3 'Slightly inaccurate' in all items for persons at average levels of neuroticism (i.e., at $\theta \approx 0$). Categories 2 and 4 work better than category 3 because they cover the same range with much higher probabilities. Therefore, by visually inspecting all such plots for all items, we may consider merging response categories, removing items with similar properties or items with visibly undesirable properties.

6.2. *Item information curves*

Another frequently asked question concerns how certain we are about a person's estimated level on the latent θ continuum. This question can be addressed by visually inspecting the Fisher information of an item over a range of θ values. A higher Fisher information means lower uncertainty for the θ estimate [1, 3] and vice versa.

Figure 3 plots the item information curves across a range of θ values for all items. Item 2 ('irritated') is visibly the most informative item. The next most informative items capture 'anger' and 'mood swings'; and 'anger' tends to provide the most information at very intense levels of neuroticism (i.e., $\theta > 1.5$). These three items do not appear to provide much information for θ values outside the $[-1.5, 2.5]$ range. New items may be needed if measuring very low levels of neuroticism is of interest. Items 4 and 5, on the other hand, share a similar low information profile. Scientifically, these items may need to be kept. From the perspective of item information, they do not seem to contribute much to the information about θ. Neither item wins over the other in pinpointing a person's underlying neuroticism. They may be revised or replaced if alternate items are available.

This five-item example shows that redundant items can be useful in a draft survey. Some seemingly redundant items may provide information over a range of the latent characteristic not covered by other theoretically more promising items. This example lacks items that are sensitive to latent characteristics below 1.5 standard deviations of the norm. Additional items would be necessary if detecting these extreme latent values were of interest. (It is hard to conceive of a need to diagnose extremely low neuroticism levels.)

Proprietary software programs generate the Fisher information curves. However, except for the simplest cases [5, pp. 70–71], the computation formulae are not easily accessible to nonexperts. This can

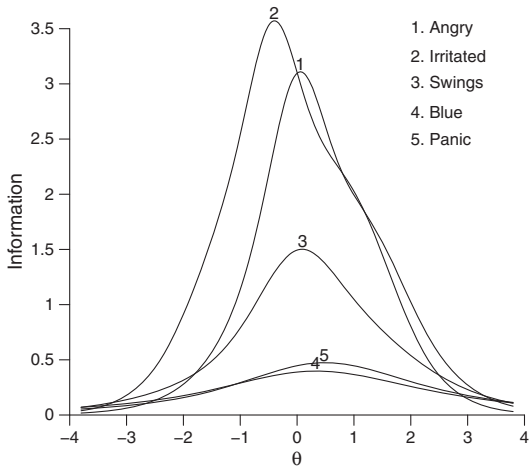

Figure 3. Item information curves of all neuroticism items.

be seen in the somewhat technical reparameterization in Muraki's [51] derivations for the GPCM. These complications may be attributed in part to the long time since Birnbaum's [25, Chapters 17 and 20] and Samejima's [52] pioneering works more than 40 years ago. There ought to be an accessible approach to the information function, one which is based on Fisher's original definition [53]. Anyone who is comfortable with Fisher's original definition should be able to follow the computation easily. That is exactly what we have attempted to do in Appendix B. We aim to provide enough details to save the readers from plowing through other sources and yet not necessarily finding the needed answers [1–3,5,25,52,54,55].

7. Discussion

We provide an illustrative example in this article to show how to fit a GPCM model using R and WinBUGS. The primary pedagogical goal is to include sufficient practical analytic techniques in one guide so that readers unfamiliar with IRT can immediately apply these skills in their own work. Also covered are a few basics in Gibbs sampler and in diagnosing the convergence of iterative simulation by MCMC methods. The parameter estimates agree well with those calculated by the maximum likelihood method implemented in the `ltm` package. This leads us to encourage the further use of R and WinBUGS to reduce the reliance on proprietary computer software programs in IRT analysis.

We focus primarily on item analysis, which is among the most technically challenging steps towards claiming a PRO instrument reliable, valid, and responsive to changes in well-being (as per Section III.E of the FDA guidance). This tutorial does not cover practical aspects of carrying out analyses of reliability, validity, and responsiveness. However, these key concepts involve statistics that are comparatively more straightforward. For example, the concept of responsiveness can be tested by comparing the PRO scores of individuals whose symptoms have changed substantially [56]. Reliability can be established by Cronbach's alpha statistic [47] and by correlating scores between repeated assessments. Validity can be established via 'concurrent validity' by correlating the scores of the target instrument and the scores derived from other instruments of the same or similar construct that have previously been validated. Another important concept is the 'construct validity', which is commonly established by a high degree of correlation between the target assessment and other assessments of similar construct (convergent validity) and the low correlation between the target assessment with other instruments of dissimilar construct (discriminant validity) [57].

Readers interested in other IRT models can find them in [21], who recently published a collection of WinBUGS code for many more IRT models, including the GPCM. However, Curtis did not provide a step-by-step link, as we have, between the mathematics and the WinBUGS code. The learner is encouraged to tackle the mathematics before plugging in the WinBUGS code—in our opinion, the recommended learning approach for biostatisticians who are already familiar with Bayesian statistics. A Bayesian approach to IRT has its own advantages. Baldwin *et al.* [9] recently reported in this journal that a Bayesian approach to IRT requires much smaller sample sizes than is required by conventional likelihood methods. The GPCM model in Equation (7) may appear complex, but a clear understanding of the mathematics reduces the model to three nested loops in the WinBUGS code. We believe that this resemblance between mathematics and computer syntax is an important advantage in preferring WinBUGS over special-purpose computer packages. It facilitates a deeper understanding of the IRT theory as well as the statistical computation. WinBUGS forces the learner to acquire a clearer understanding of the statistical model, which hopefully discourages indiscriminately applying existing data analysis 'recipes'. Similarly, experienced biostatisticians who are already familiar with the R language no longer need to learn the sometimes eccentric computer programming languages of proprietary IRT software. It opens up new possibilities as well, as seen in analyses on gene expression data [10], explanatory IRT models [58], and recent psychometric work in cognitive models of IRT [59,60].

Iterative simulation is time consuming. It is the main disadvantage against the Bayesian IRT approach. The time and effort diminish the enthusiasm for WinBUGS and its open-source version OpenBUGS in IRT analysis (also JAGS [61]). Analysts save time by using existing macros in their preferred statistical computer package, for example, the macros in SAS (SAS Institute, Cary, NC, USA) [62–64] and STATA (StataCorp, College Station, TX, USA) [65–67], and solutions based on SAS NLMIXED procedure [68], if all they want are the parameter estimates and other statistics supported by these macros. Another limitation is the somewhat steep learning curve for beginners. It is compounded by practical complications. For example, problems in model convergence are hard to diagnose and rectify. Mistakes in WinBUGS syntax are hard to debug because error messages are often nonspecific.

These are some of the reasons why beginners are intimated by WinBUGS. These limitations notwithstanding, we believe that circumstances will improve over time, with the publication of WinBUGS code collections and tutorials similar to this one. Psychometric models and methods in R are rapidly emerging. In 2007, the Journal of Statistical Software dedicated volume 20 to psychometric methods in R. Another collaborative effort is found in the online 'Task View' maintained by Mair and Hatzinger (http://cran.r-project.org/web/views/Psychometrics.html, last accessed July, 2011). New development efforts in the OpenBUGS program include improvements in the documentations as well as cross-platform compatibility. We are optimistic that, in time, as new and accessible knowledge bases accrue, R and WinBUGS will become the preferred tools for fitting IRT models.

APPENDIX A. Detailed derivations of the partial credit model

The G term in Equation (5) is crucial in understanding how the PCM equation is derived. However, a few algebraic steps were omitted in Masters' original derivations [23, p. 158] and Muraki's [19] subsequent definition of the G term. They are restored here. These details are also helpful in tracking the model syntax in WinBUGS. We begin by restating the conditional probability of endorsing each successive response category k over $k-1$ for item j:

$$\Phi_{jk} = \frac{\pi_{jk}}{\pi_{j[k-1]} + \pi_{jk}} = \frac{\exp(\theta_i - \beta_{jk})}{1 + \exp(\theta_i - \beta_{jk})}.$$

In the aforementioned $K = 4$ example,

$$\Phi_{j1} = \frac{\pi_{j1}}{\pi_{j0} + \pi_{j1}} = \frac{\exp(\theta_i - \beta_{j1})}{1 + \exp(\theta_i - \beta_{j1})}, \quad \text{(Disagree over Strongly disagree)}$$

$$\Phi_{j2} = \frac{\pi_{j2}}{\pi_{j1} + \pi_{j2}} = \frac{\exp(\theta_i - \beta_{j2})}{1 + \exp(\theta_i - \beta_{j2})}, \quad \text{(Agree over Disagree)}$$

$$\Phi_{j3} = \frac{\pi_{j3}}{\pi_{j2} + \pi_{j3}} = \frac{\exp(\theta_i - \beta_{j3})}{1 + \exp(\theta_i - \beta_{j3})}. \quad \text{(Strongly agree over Agree)}$$

Thus, $\pi_{j1} = [(\pi_{j0} + \pi_{j1})\exp(\theta_i - \beta_{j1})]/[1 + \exp(\theta_i - \beta_{j1})]$, which can be simplified as follows:

$$\pi_{j1}[1 + \exp(\theta_i - \beta_{j1})] = (\pi_{j0} + \pi_{j1})\exp(\theta_i - \beta_{j1}),$$

$$\pi_{j1} + \exp(\theta_i - \beta_{j1}) \cdot \pi_{j1} = \exp(\theta_i - \beta_{j1}) \cdot \pi_{j0} + \exp(\theta_i - \beta_{j1}) \cdot \pi_{j1},$$

$$\pi_{j1} + \cancel{\exp(\theta_i - \beta_{j1}) \cdot \pi_{j1}} = \exp(\theta_i - \beta_{j1}) \cdot \pi_{j0} + \cancel{\exp(\theta_i - \beta_{j1}) \cdot \pi_{j1}},$$

so $\pi_{j1} = \exp(\theta_i - \beta_{j1}) \cdot \pi_{j0}$ because the $\exp(\theta_i - \beta_{j1}) \cdot \pi_{j1}$ terms cancel out. To make the notations easier to track, we temporarily replace the somewhat cumbersome term of $\exp(\theta_i - \beta_{jk})$ with δ_k. We have $\pi_{j1} = \delta_1 \pi_{j0}$. Similarly, $\pi_{j2} = \delta_2 \pi_{j1}$ and $\pi_{j3} = \delta_3 \pi_{j2}$. Because each person must choose one of the four possible responses for item j, $\pi_{j0} + \pi_{j1} + \pi_{j2} + \pi_{j3} = 1$. We can solve for π_{j0}:

$$\pi_{j0} = \frac{1}{1 + \delta_1 + \delta_2\delta_1 + \delta_3\delta_2\delta_1}, \text{and because } \exp(0) = 1,$$

$$\delta_1 = 1 \times \delta_1 = \exp(0) \times \exp(\theta_i - \beta_{j1}),$$

$$\delta_2\delta_1 = \exp(0) \times \exp(\theta_i - \beta_{j1}) \times \exp(\theta_i - \beta_{j2}),$$

$$\delta_3\delta_2\delta_1 = \exp(0) \times \exp(\theta_i - \beta_{j1}) \times \exp(\theta_i - \beta_{j2}) \times \exp(\theta_i - \beta_{j3}).$$

Let $G = 1 + \delta_1 + \delta_2\delta_1 + \delta_3\delta_2\delta_1$, we have

$$\pi_{j0} = \frac{1}{G},$$

$$\pi_{j1} = \frac{1 \times \delta_1}{G} = \frac{\exp(0) \times \exp(\theta_i - \beta_{j1})}{G},$$

$$\pi_{j2} = \frac{1 \times \delta_2\delta_1}{G} = \frac{\exp(0) \times \exp(\theta_i - \beta_{j1}) \times \exp(\theta_i - \beta_{j2})}{G},$$

$$\pi_{j3} = \frac{1 \times \delta_3\delta_2\delta_1}{G} = \frac{\exp(0) \times \exp(\theta_i - \beta_{j1}) \times \exp(\theta_i - \beta_{j2}) \times \exp(\theta_i - \beta_{j3})}{G}.$$

To obtain a general notation, we define that item j is scored $y = 0, 1, 2, \ldots, m_j$ with $K_j = m_j + 1$ response categories and the denominator $G = 1 + \delta_1 + \delta_2\delta_1 + \delta_3\delta_2\delta_1 + \ldots + \delta_{m_j}\delta_{m_j-1}\cdots\delta_1$. This

sum follows a highly regular pattern, which involves $\exp(0)$ in the first term, $\exp(0) \times \exp(\theta_i - \beta_{j1})$ in the second term, $\exp(0) \times \exp(\theta_i - \beta_{j1}) \times \exp(\theta_i - \beta_{j2})$ in the third term, and so on until the last term $\exp(0) \times \exp(\theta_i - \beta_{j1}) \times \exp(\theta_i - \beta_{j2}) \cdots \times \exp(\theta_i - \beta_{jm_j})$. Thus,

$$G = \sum_{k=0}^{m_j} \exp\left[\sum_{h=0}^{k} (\theta_i - \beta_{jh})\right], \tag{A1}$$

where $\sum_{h=0}^{0} (\theta_i - \beta_{jh}) \equiv 0$ for convenience. Finally, a general model expression for $K_j = m_j + 1$ that incorporates all of the aforementioned steps is given in the PCM in Equation (6).

APPENDIX B. More on item information

This appendix covers the derivations of the item information function that assesses the amount of information that can be expected from an item. The derivations are pedagogically useful, especially for biostatisticians who are familiar with Fisher's original definition [69] but are not familiar with psychometrics. Samejima [52, Chapter 6] was first to introduce Fisher's information into IRT in her 1969 monograph, according to Baker and Kim [5, p. 222]. The detailed derivations in Samejima's work are no longer available in the literature (e.g., [5, p. 222]; [70, p. 374]; [1, p. 208]; [54, p. 308]). Here, we summarize our independent derivations starting from Fisher's definition.

On the basis of Fisher's original definition, the information function for the latent characteristic θ given data y is as follows:

$$I(\theta|y) = \int \left[\frac{\partial}{\partial\theta} \log \Pr(y|\theta)\right]^2 \Pr(y|\theta)\, \mathrm{d}y, \tag{B1}$$

where, for the GPCM, the log likelihood of the probability density function is defined in Equations (7) and (8). There is no need to include $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ into the log-likelihood function because $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are fixed, which also makes the notations easier to track. Incidentally, the first derivative of the log likelihood, $\partial \log \Pr(y|\theta) \big/ \partial\theta$, is also known as the *score* of the probability density function. Hogg and Craig [53, pp.373–374]) show that Equation (B1) is the same as the following:

$$I(\theta|y) = -\int \left[\frac{\partial^2}{\partial\theta^2} \log \Pr(y|\theta)\right] \Pr(y|\theta)\, \mathrm{d}y, \tag{B2}$$

and thus the following definition given in texts such as Lee [31, p. 82], Berger [71, Section 3.3.3], and Gelman *et al.* [72, Section 2.9]:

$$I(\theta|y) = -\mathsf{E}\left[\partial^2 \left(\log \Pr(y|\theta)\right) / \partial\theta^2\right], \tag{B3}$$

the expectation being taken over all possible values of y for fixed item parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. To save notations further, we write p for the probability density function $\Pr(\boldsymbol{y}|\theta)$.

Muraki's [19] derivations for the item information function begin with Equation (B2). The integration is replaced with summation across m_j discrete response categories for item j.

$$I_j(\theta|\boldsymbol{y}) = \sum_{k=0}^{m_j} p_{jk} \left[-\frac{\partial^2}{\partial\theta^2} \log p_{jk}\right], \tag{B4}$$

where the information of all possible response categories are weighted by their corresponding response probabilities. Up to this point, the definition of item information should be familiar to biostatisticians. However, an alternate but unfamiliar definition is used in the psychometric literature (e.g., [54, Equation (2)]):

$$I_j(\theta|y) = \sum_{k=0}^{m_j} \frac{\left[\frac{\partial}{\partial\theta} p_{jk}\right]^2}{p_{jk}}. \tag{B5}$$

Lee [31, pp. 82–83] provides a general proof that Equations (B4) and (B5) are indeed equivalent. However, the seemingly different equations may be a source of confusion and frustration in learning

Fisher information in IRT. The intermediate derivations from Equations (B4) and (B5) are not found in recently published texts (e.g., [5, p. 222]; [70, p. 374]; [1, p. 208]; [54, p. 308]). The missing derivations are summarized next.

In Equation (B4), we need to find the second derivative of $\log p$ with respect to θ. We take the first derivative and obtain

$$\frac{\partial}{\partial\theta}\log p = \frac{1}{p}\cdot\frac{\partial}{\partial\theta}p$$

by the general power rule in calculus. We then take the negative of the second derivative, $-\frac{\partial}{\partial\theta}\left[\frac{1}{p}\cdot\frac{\partial}{\partial\theta}p\right]$, and follow the product rule to get the following:

$$\begin{aligned}
-\frac{\partial}{\partial\theta}\left[\frac{1}{p}\cdot\frac{\partial}{\partial\theta}p\right] &= -\left[\frac{1}{p}\cdot\frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta}p\right] + \frac{\partial}{\partial\theta}p\cdot\frac{\partial}{\partial\theta}\frac{1}{p}\right]\\
&= -\left[\frac{1}{p}\cdot\frac{\partial^2}{\partial\theta^2}p + \frac{\partial}{\partial\theta}p\cdot\frac{\partial}{\partial\theta}p^{-1}\right]\\
&= -\left[\frac{1}{p}\cdot\frac{\partial^2}{\partial\theta^2}p + \frac{\partial}{\partial\theta}p\cdot\left[-1\cdot p^{-2}\cdot\frac{\partial}{\partial\theta}p\right]\right]\\
&= -\left[\frac{1}{p}\cdot\frac{\partial^2}{\partial\theta^2}p - \left[\frac{\frac{\partial}{\partial\theta}p}{p}\right]^2\right]\\
&= \left[\left[\frac{\partial\partial\theta\, p}{p}\right]^2 - \frac{1}{p}\cdot\frac{\partial^2}{\partial\theta^2}p\right].
\end{aligned}$$

By plugging in the previous result back into Equation (B4), we get

$$\begin{aligned}
I_j(\theta) &= \sum_{k=0}^{m_j} p_{jk}\left[\left[\frac{\frac{\partial}{\partial\theta}p_{jk}}{p_{jk}}\right]^2 - \frac{\frac{\partial^2}{\partial\theta^2}p_{jk}}{p_{jk}}\right]\\
&= \sum_{k=0}^{m_j}\frac{\left[\frac{\partial}{\partial\theta}p_{jk}\right]^2}{p_{jk}} - \sum_{k=0}^{m_j}\frac{\partial^2}{\partial\theta^2}p_{jk}.
\end{aligned} \tag{B6}$$

The second term of Equation (B6), $\sum_{k=0}^{m_j}\frac{\partial^2}{\partial\theta^2}p_{jk}$, must sum to zero. It is easier to see why this must be the case using the four-category example mentioned previously. When θ is fixed, $p_0+p_1+p_2+p_3=1$, because each person must choose one of the m_j possible responses for item j. By taking the second partial derivatives with respect to θ on both sides of the equation, we get $\frac{\partial^2}{\partial^2\theta}p_0+\frac{\partial^2}{\partial^2\theta}p_1+\frac{\partial^2}{\partial^2\theta}p_2+\frac{\partial^2}{\partial^2\theta}p_3 = \frac{\partial^2}{\partial^2\theta}1=0$. The sum of the second partial derivatives of the response probabilities must be 0. Thus,

$$I_j(\theta) = \sum_{k=0}^{m_j}\frac{\left[\frac{\partial}{\partial\theta}p_{jk}\right]^2}{p_{jk}}.$$

References

1. de Ayala RJ. *The Theory and Practice of Item Response Theory*. The Guilford Press: New York, NY, 2009.
2. Embretson SE, Reise SP. *Item Response Theory for Psychologists*. LEA: Mahwah, NJ, 2000.
3. Lord FM. *Application of Item Response Theory to Practical Testing Problems*. Erlbaum: Hillsdale, NJ, 1980.
4. van der Linden W, Hambleton RK. *Handbook of Modern Item Response Theory*. Springer-Verlag: New York, 1997.
5. Baker FB, Kim S-H. *Item Response Theory: Parameter Estimation Techniques*. Marcel Dekker, Inc.: New York, 2004.
6. Bjorner JB, Chang CH, Thissen D, Reeve BB. Developing tailored instruments: item banking and computerized adaptive assessment. *Quality of Life Research* 2007; **16 Suppl 1**:95–108.
7. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Thissen D, Revicki DA, Weiss DJ, Hambleton RK, Liu H, Gershon R, Reise SP, Lai JS, Cella D. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the patient-reported outcomes measurement information system (PROMIS). *Medical Care* 2007; **45**:S22–31.

8. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, Amtmann D, Bode R, Buysse D, Choi S, Cook K, Devellis R, DeWalt D, Fries JF, Gershon R, Hahn EA, Lai JS, Pilkonis P, Revicki D, Rose M, Weinfurt K, Hays R. The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology* 2010; **63**(11):1179–94.
9. Baldwin P, Bernstein J, Wainer H. Hip psychometrics. *Statistics in Medicine* 2009; **28**:2277–2292.
10. Li H, Hong F. Cluster-Rasch models for microarray gene expression data. *Genome Biology* 2001; **2**(8):RESEARCH0031.
11. R Development Core Team. R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2010. http://www.R-project.org/, ISBN 3-900051-07-0.
12. Lunn DJ, Thomas A, Best N, Spiegelhalter D. Winbugs – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**:325–337.
13. Food and Drug Administration. Guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims, 2009. Last accessed: December, 2009. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf.
14. Food and Drug Administration. Guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims, 2006. Last accessed: December, 2009. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM155480.pdf.
15. European Medicines Agency Committee for medicinal products for human use (CHMP). Reflection paper on the regulatory guidance for the use of health-related quality of life (HRQL) measures in the evaluation of medicinal products, 2005. Last accessed: December, 2009. human/ewp/13939104en.pdf.
16. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. The University of Chicago Press: Chicago, 1980.
17. Weesie J. The Rasch model in Stata, 1999. http://www.stata.com/support/faqs/stat/rasch.html, Last accessed: December, 2009.
18. Agresti A. *Categorical Data Analysis*. Wiley: Hoboken, NJ, 2002.
19. Muraki E. A generalized partial credit model: application of an em algorithm. *Applied Psychological Measurement* 1992; **16**:159–176.
20. Fox JP. *Bayesian Item Response Modeling: Theory and Applications*. Springer: New York, 2010.
21. Curtis SM. Bugs code for item response theory. *Journal of Statistical Software* 2010; **36**:1–34.
22. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of complexity and fit. *Journal of the Royal Statistical Society, SeriesB* 2002; **64**:583–639.
23. Masters GN. A Rasch model for partial credit scoring. *Psychometrika* 1982; **47**(2):149–174.
24. Thissen D. Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika* 1982; **47**(2): 175–186.
25. Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores*, Lord FM, Novick MR (eds). Addison-Wesley: Reading, MA, 1968.
26. Thissen D, Steinberg L. A taxonomy of item response models. *Psychometrika* 1986; **51**(4):567–577.
27. Lord FM, Novick MR (eds). *Statistical Theories of Mental Test Scores*. Addison-Wesley: Reading, MA, 1968.
28. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 1981; **46**(4):443–459.
29. Albert JH. Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics* 1992; **17**:251–269.
30. Bafumi J, Gelman A, Park DK, Kaplan N. Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis* 2005; **13**(2):171–187.
31. Lee PM. *Bayesian Statistics: An Introduction*, 3rd ed. Arnold Publishers: London, 2004.
32. Patz RJ, Junker BW. Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics* 1999; **24**(4):342–366.
33. de la Torre J, Stark S, Chernyshenko OS. Markov chain Monte Carlo estimation of item parameters for the generalized graded unfolding model. *Applied Psychological Measurement* 2006; **30**(3):216–232.
34. Martin AD, Quinn KM, Park JH. *MCMCpack: Markov chain Monte Carlo (MCMC) package*, 2010. http://CRAN.R-project.org/package=MCMCpack,Rpackageversion1.0-8,.
35. Congdon P. *Bayesian Statistical Modeling*. John Wiley & Sons Ltd.: Chichester, England, 2006.
36. Revelle W. psych: Procedures for psychological, psychometric, and personality research, Northwestern University, Evanston, Illinois, 2010. http://personality-project.org/r/psych.manual.pdf,Rpackageversion1.0-90.
37. Goldberg LR. The development of markers for the big-five factor structure. *Psychological Assessment* 1992; **4**:26–42.
38. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**(4): 457 –511.
39. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press: New York, 2007.
40. Jackman S. *Bayesian Analysis for the Social Sciences*. John Wiley & Sons, Ltd.: Chichester, United Kingdom, 2009.
41. Ntzoufras I. *Bayesian Modeling Using WinBUGS*. John Wiley & Sons, Inc.: Hoboken, NJ, 2009.
42. Brooks SP, Roberts GO. Assessing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing* 1998; **8**(319–335).
43. Cowles MK, Carlin BP. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association* 1996; **91**:883–904.
44. Geweke J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics*, Vol. 4, Bernardo J, Berger J, Dawid A, Smith A (eds). Claredon Press: Oxford, UK, 1992.
45. Heidelberger P, Welch PD. A spectral method for confidence interval generation and run length control in simulations. *Communication of the ACM* 1981; **24**:233–245.

46. Raftery A, Lewis S. How many iterations in the Gibbs sampler? In *Bayesian Statistics*, Vol. 4, Bernardo J, Berger J, Dawid A, Smith A (eds). Claredon Press: Oxford, UK. 763–774, 1992.
47. Nunnally JC, Bernstein IH. *Psychometric Theory*, 3rd ed. McGraw-Hill: New York, 1994.
48. Bond TG, Fox CM. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Lawrence Erlbaum Associates: Mahwah, NJ, 2001.
49. Karabatsos G. Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education* 2003; **16**:277–298.
50. Meijer RR, Sijtsma K. Methodology review: evaluating person fit. *Applied Psychological Measurement* 2001; **25**: 107–135.
51. Muraki E. Information functions of the generalized partial credit model. *Applied Psychological Measurement* 1993; **17**:351–363.
52. Samejima F. Estimation of latent ability using a response pattern of graded scores, 1969. Psychometrika Monograph, No. 17. Richmond, VA: Psychometric Society. Retrieved in August, 2011 from http://www.psychometrika.org/journal/online/MN17.pdf.
53. Hogg RV, Craig AT. *Introduction to Mathematical Statistics*. Prentice Hall: Englewood Cliffs, NJ, 1995.
54. Samejima F. Some critical observations of the test information function as a measure of local accuracy in ability. *Psychometrika* 1994; **59**(3):307–329.
55. Samejima F. A general model for free-response data, 1972. Psychometrika Monograph, No. 18, 37 (1, Pt. 2).
56. Juniper EF, Guyatt GH, Feeny DH, Griffith LE, Ferrie PJ. Minimum skills required by children to complete health-related quality of life instruments for asthma: comparison of measurement properties. *European Respiratory Journal* 1997; **10**(10):2285–2294. English.
57. Downing SM. Validity: on meaningful interpretation of assessment data. *Medical Education* 2003; **37**(9):830–837.
58. De Boeck P, Wilson M. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Spinger-Verlag: New York, 2004.
59. de la Torre J, Douglas JA. Model evaluation and multiple strategies in cognitive diagnosis: an analysis of fraction subtraction data. *Psychometrika* 2008; **73**:595–624.
60. Junker BW, Sijtsma K. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement* 2001; **25**:258–272.
61. Plummer M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), March 20 – 22*, Vienna, Austria, 2003.
62. Lee S-H, Terry R. IRT-FIT: SAS macros for fitting item response theory (IRT) models. *In Proceedings of the SAS Users Group International (SUGI 30) Conference, April 9-13, 2005*, Philadelphia, PA, 2005.
63. Lee S-H, Terry R. MDIRT-FIT: SAS macros for fitting multidimensional item response. *In Proceedings of the SAS Users Group International (SUGI 31) Conference, March 26-29, 2006*, San Francisco, California, 2011.
64. Pan T, Chen Y. Using proc logistic to estimate the Rasch model. *Proceedings of the SAS Global Forum 2011 Conference, April 4 - 7, 2011.*, Las Vegas, Neveda, 2011.
65. Hardouin JB. Rasch analysis: estimation and tests with raschtest. *Stata Journal* 2007; **7**(1):22–44.
66. Weesie J. The Rasch model in Stata. Stata Corp. 1999. Accessed August 31, 2011, from http://www.stata.com/support/faqs/stat/rasch.html.
67. Zajonc T. OpenIRT - Bayesian and maximum likelihood estimation of item response theory (IRT) models in Stata, 2011. Accessed August 31, 2011 at http://wwwpeoplefasharvardedu/tzajonc/openirthtml.
68. Sheu CF, Chen CT, Su YH, Wang WC. Using SAS PROC NLMIXED to fit item response theory models. *Behavioral Research Methods* 2005; **37**:202–218.
69. Fisher RA. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society* 1925; **22**:700–725.
70. Dodd BG, Koch WR. Effects of variations in item step values on item and test information in the partial credit model. *Applied Psychological Measurement* 1987; **11**(4):371–384.
71. Berger JO. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag: New York, 1985.
72. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*, 2nd ed. Chapman & Hall: New York, 2003.