Mathematical Optimization Models for Nonparametric Item Response Theory

Yuichi Takano
School of Network and Information, Senshu University
2-1-1 Higashimita, Tama-ku, Kawasaki-shi, Kanagawa 214-8580, JAPAN

Shintaro Tsunoda **Masaaki Muraki**
Department of Industrial Engineering and Management
Graduate School of Decision Science and Technology, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, JAPAN

Abstract

This paper explores a mathematical optimization approach to nonparametric item response theory (NIRT). Specifically, we develop mathematical optimization models for estimating nonparametric item characteristic curves and latent abilities of examinees simultaneously. These models maximize the log likelihood function under the monotone homogeneity and double monotonicity constraints and are formulated as mixed integer nonlinear programming problems. Since these problems are very hard to solve exactly, we devise heuristic optimization algorithms to efficiently find a good-quality solution. Through the computational experiments, the effectiveness of our mathematical optimization models and heuristic optimization algorithms are demonstrated by comparison to the common two-parameter logistic IRT model.

Keywords: Nonparametric IRT, Mathematical optimization model, Heuristic optimization algorithm, Item characteristic curve estimation, Latent ability estimation

1 Introduction

Item response theory (IRT) is a modern test theory for the design, analysis, and scoring of tests. A key component of IRT is the item characteristic curve (ICC), which shows the relationship between the examinee's latent ability and the probability of correctly answering a question item. ICCs of question items and the latent abilities of examinees are estimated from the item response data of examinees. The aim of IRT is to investigate not the test score, but the latent (i.e., not directly observable) ability of each examinee. Moreover, this methodology allows one to closely examine item characteristics, such as the difficulty and discrimination. According to approaches taken to estimating the ICCs, IRT models can be divided into two categories, i.e., parametric item response theory (PIRT) and nonparametric item response theory (NIRT). PIRT models typically force ICCs to be parametric functions (e.g., logistic curves or normal ogives). On the other hand, this paper focuses on NIRT models, which do not assume any particular parametric form for the ICCs.

NIRT has its origin in Meredith's work [10] and Mokken scale analysis [11], and it has seen steady development in both its theory and applications (see, e.g., [13, 19, 20, 21, 22]). The greatest benefit of NIRT models is their ability to estimate various forms of ICCs given only mild assumptions. Indeed, it has been demonstrated, e.g., in [4, 5, 15], that PIRT models do not always fit the data well. In this case, NIRT models, which provide a more flexible framework, are particularly beneficial. They are also useful for determining whether PIRT model assumptions are valid or not (see, e.g., [6]). However, greater flexibility of nonparametric ICCs sometimes makes a model overfit the data. As pointed out by Molenaar [13], an estimation based on NIRT models may consequently be unstable especially when there is not much item response data.

There are several methods of estimating nonparametric ICCs, e.g., kernel smoothing [3, 15], isotonic regression [7], and B-spline models [5, 16]. The most commonly used approach is kernel smoothing; however, they sometimes estimate ICCs that decrease with respect to the latent ability. In other words, kernel smoothing dose not always preserve monotone homogeneity [10, 11], which is the most fundamental property required by ICCs. In contrast, isotonic regression [7] and B-spline model [5] ensure that ICCs are nondecreasing. A number of studies have assessed the goodness of fit of PIRT models by means of these estimation procedures for nonparametric ICCs (see, e.g., [4, 8, 9, 23, 25]).

The ordered latent class models [1, 2, 24] estimate the nonparametric ICCs and the latent classes of examinees simultaneously. Croon [1, 2] and van Onna [24] used the expectation-maximization (EM) algorithm and Markov chain Monte Carlo (MCMC) method, respectively in these models. On the other hand, the purpose of the present paper is to build a new computational framework for estimating the nonparametric ICCs and the latent abilities of examinees simultaneously. To accomplish this, we provide a mathematical optimization approach. Mathematical optimization models make it possible to place various restrictions on excessively flexible ICCs. Accordingly, our model can incorporate two basic constraints on nonparametric ICCs, i.e., the monotone homogeneity constraint [10, 11] and the double monotonicity constraint [11, 12], as in the ordered latent class models [1, 2, 24]. Moreover, we conducted computational experiments to assess the effectiveness of our NIRT models in comparison with the common two-parameter logistic IRT model.

Our contributions are summarized as follows:

- We formulate mathematical optimization models for NIRT as mixed integer nonlinear programming (MINLP) problems. These formulations determine the nonparametric ICCs and the latent abilities of examinees simultaneously under the required constraints.
- We devise heuristic optimization algorithms to efficiently find good-quality solutions to NIRT models that are very hard to solve exactly. The computational results demonstrated that the solutions provided by our algorithms were good enough to achieve positive results for our models.

The rest of the paper is organized as follows: In Section 2, we explain nonparametric ICC estimation and its basic assumptions. In Section 3, we present mathematical optimization models for NIRT. Section 4 is devoted to our heuristic optimization algorithm for solving the NIRT model with the monotone homogeneity constraint. Computational results are reported in Section 5. Finally, conclusions are presented in Section 6.

2 Nonparametric Item Characteristic Curve Estimation

Let us suppose that examinees $i = 1, 2, \ldots, I$ have taken a test consisting of dichotomously scored question items $j = 1, 2, \ldots, J$. More specifically, we are given the binary item response data,

$$\boldsymbol{U} = (u_{i,j};\ i = 1, 2, \ldots, I,\ j = 1, 2, \ldots, J) \in \{0, 1\}^{I \times J},$$

where $u_{i,j} = 1$ if examinee i gave a correct answer to question item j, and $u_{i,j} = 0$ otherwise. The item characteristic curves (ICCs) and the latent abilities of examinees are estimated from this item response data.

This paper addresses nonparametric item response theory (NIRT) that is characterized by a nonparametric ICC estimation. In the conventional way, the following two assumptions are made throughout the paper:

Unidimensionality: the latent abilities of all examinees can be evaluated unidimensionally.

Local Independence: item responses are conditionally independent of each other given an individual latent ability.

In addition, we shall evaluate the latent abilities of examinees on a discrete scale of $t = 1, 2, \ldots, T$, which we call the ability class. To describe the nonparametric ICCs, we introduce the decision variable,

$$\boldsymbol{X} = (x_{j,t};\ j = 1, 2, \ldots, J,\ t = 1, 2, \ldots, T) \in \mathbb{R}^{J \times T},$$

where $x_{j,t}$ is the probability of question item j being answered correctly by examinees of ability class t. Figure 1 illustrates a nonparametric ICC represented as a piecewise linear function.

The most fundamental property required for ICCs is monotone homogeneity (MH) [10, 11]. This implies that all ICCs are nondecreasing with a latent ability. In other words, the probability of a correct answer does not decrease with the ability class of the examinee. This property can be expressed as the following constraints:

$$\text{Monotone Homogeneity}: \quad 0 \le x_{j,1} \le x_{j,2} \le \cdots \le x_{j,T} \le 1 \quad (\forall j = 1, 2, \ldots, J). \tag{1}$$

An additional assumption of nonparametric ICCs is double monotonicity (DM) [11, 12]. This assumption implies that the ICC of one item does not intersect with the other. In other

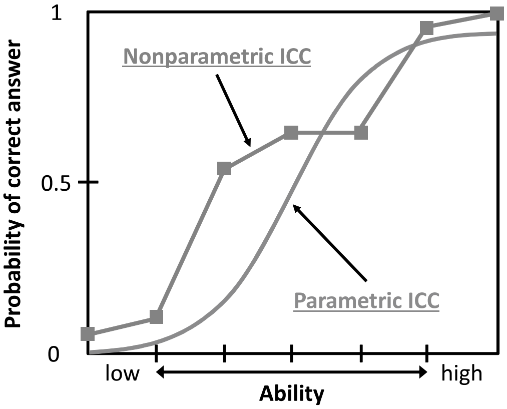

Figure 1: Parametric and Nonparametric Item Characteristic Curves

words, for all classes of examinees, the difficulties of two question items are never reversed. To formulate a clear definition, we suppose that there is a permutation,

$$\sigma : \ \{1, 2, \dots, J\} \to \{1, 2, \dots, J\},$$

such that $\sigma(k) = j$ means that the k-th most difficult item is question item j. We refer to σ as the difficulty ranking function. Accordingly, the DM constraints are expressed as follows:

$$\text{Double Monotonicity :} \quad x_{\sigma(1),t} \le x_{\sigma(2),t} \le \cdots \le x_{\sigma(J),t} \quad (\forall t = 1, 2, \dots, T). \tag{2}$$

That is, for all classes of examinees, the probability of correctly answering a high-ranking item is lower than that of correctly answering a low-ranking one.

3 Mathematical Optimization Models

This section presents mathematical optimization models for NIRT. We first formulate a log likelihood function to be maximized. We then develop a monotone homogeneity model and a double monotonicity model.

3.1 Log likelihood function

Let us introduce the decision variable to estimate the ability class of examinees,

$$\boldsymbol{Y} = (y_{i,t};\ i = 1, 2, \dots, I,\ t = 1, 2, \dots, T) \in \mathbb{R}^{I \times T},$$

where $y_{i,t} = 1$ if the ability class of examinee i is t, and $y_{i,t} = 0$ otherwise. Since only one ability class should be assigned to each examinee, the following constraints must be satisfied,

$$\sum_{t=1}^{T} y_{i,t} = 1 \quad (\forall i = 1, 2, \dots, I), \tag{3}$$

$$y_{i,t} \in \{0, 1\} \quad (\forall i = 1, 2, \dots, I,\ \forall t = 1, 2, \dots, T). \tag{4}$$

Now, we can define a log likelihood function to be maximized. Given $\boldsymbol{x}_j := (x_{j,1}, x_{j,2}, \dots, x_{j,T})$ and $\boldsymbol{y}_i := (y_{i,1}, y_{i,2}, \dots, y_{i,T})$, we can see from (3) and (4) that the probability of having the response $u_{i,j} \in \{0, 1\}$ becomes

$$\Pr(u_{i,j} \mid \boldsymbol{x}_j, \boldsymbol{y}_i) = \sum_{t=1}^{T} y_{i,t} (x_{j,t})^{u_{i,j}} (1 - x_{j,t})^{1 - u_{i,j}}.$$

Accordingly, under the local independence assumption, the probability of examinee i giving the response $\boldsymbol{u}_i := (u_{i,1}, u_{i,2}, \dots, u_{i,J})$ is

$$\Pr(\boldsymbol{u}_i \mid \boldsymbol{X}, \boldsymbol{y}_i) = \prod_{j=1}^{J} \Pr(u_{i,j} \mid \boldsymbol{x}_j, \boldsymbol{y}_i).$$

Since the responses of different examinees are independent, the overall item response $\boldsymbol{U}$ occurs with the probability,

$$\Pr(\boldsymbol{U} \mid \boldsymbol{X}, \boldsymbol{Y}) = \prod_{i=1}^{I} \Pr(\boldsymbol{u}_i \mid \boldsymbol{X}, \boldsymbol{y}_i) = \prod_{i=1}^{I} \prod_{j=1}^{J} \left(\sum_{t=1}^{T} y_{i,t} (x_{j,t})^{u_{i,j}} (1 - x_{j,t})^{1 - u_{i,j}} \right).$$

By treating $\boldsymbol{X}$ and $\boldsymbol{Y}$ as decision variables, the log likelihood function can be defined as follows:

$$\ell(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{U}) = \log \Pr(\boldsymbol{U} \mid \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{I} \sum_{j=1}^{J} \log \left(\sum_{t=1}^{T} y_{i,t} (x_{j,t})^{u_{i,j}} (1 - x_{j,t})^{1 - u_{i,j}} \right).$$

In view of constraints (3) and (4), the log likelihood function can be rewritten as follows:

$$\begin{aligned} \ell(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{U}) &\overset{(3),(4)}{=} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} y_{i,t} \log \left((x_{j,t})^{u_{i,j}} (1 - x_{j,t})^{1 - u_{i,j}} \right) \\ &= \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} y_{i,t} \left(u_{i,j} \log(x_{j,t}) + (1 - u_{i,j}) \log(1 - x_{j,t}) \right). \end{aligned} \tag{5}$$

3.2 Monotone homogeneity model

The monotone homogeneity (MH) model estimates $\boldsymbol{X}$ and $\boldsymbol{Y}$ so that the log likelihood function, $\ell(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{U})$, is maximized under conditions (1), (3), and (4). Consequently, the MH model can be framed as a mixed integer nonlinear programming (MINLP) problem,

$$\text{(MHM)}\quad\left|\begin{array}{ll}\underset{\boldsymbol{X},\boldsymbol{Y}}{\text{maximize}} & \displaystyle\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T} y_{i,t}\left(u_{i,j}\log(x_{j,t})+(1-u_{i,j})\log(1-x_{j,t})\right)\\ \text{subject to} & 0\le x_{j,1}\le x_{j,2}\le\cdots\le x_{j,T}\le 1\quad(\forall j=1,2,\ldots,J),\\ & \displaystyle\sum_{t=1}^{T} y_{i,t}=1\quad(\forall i=1,2,\ldots,I),\\ & y_{i,t}\in\{0,1\}\quad(\forall i=1,2,\ldots,I,\ \forall t=1,2,\ldots,T).\end{array}\right.$$

3.3 Double monotonicity model

Next, we deal with a mathematical optimization problem with double monotonicity (DM) constraints (2).

Let us recall that $\sigma(k)=j$ means that the k-th most difficult item is question item j. In what follows, we shall represent this difficulty ranking function with the permutation matrix,

$$\boldsymbol{Z}=(z_{j,k};\ j=1,2,\ldots,J,\ k=1,2,\ldots,J)\in\mathbb{R}^{J\times J}, \tag{6}$$

$$z_{j,k}=1 \iff \sigma(k)=j. \tag{7}$$

It follows from the definition that the permutation matrix satisfies

$$\sum_{k=1}^{J} z_{j,k}=1\quad(\forall j=1,2,\ldots,J), \tag{8}$$

$$\sum_{j=1}^{J} z_{j,k}=1\quad(\forall k=1,2,\ldots,J), \tag{9}$$

$$z_{j,k}\in\{0,1\}\quad(\forall j=1,2,\ldots,J,\ \forall k=1,2,\ldots,J). \tag{10}$$

The optimization model presented below finds an appropriate difficulty ranking by treating $\boldsymbol{Z}$ as a decision variable.

To estimate ICCs under the DM constraints, we further use a new decision variable,

$$\boldsymbol{W}=(w_{k,t};\ k=1,2,\ldots,J,\ t=1,2,\ldots,T)\in\mathbb{R}^{J\times T},$$

which represents the probability of the k-th most difficult item being answered correctly by examinees of ability class t. The MH and DM constraints on $\boldsymbol{W}$ can be expressed as follows:

$$0\le w_{k,1}\le w_{k,2}\le\cdots\le w_{k,T}\le 1\quad(\forall k=1,2,\ldots,J), \tag{11}$$

$$w_{1,t}\le w_{2,t}\le\cdots\le w_{J,t}\quad(\forall t=1,2,\ldots,T). \tag{12}$$

The associated log likelihood function becomes

$$\begin{aligned}
\ell(\boldsymbol{W},\boldsymbol{Y},\boldsymbol{Z}\mid\boldsymbol{U}) &\overset{(5)}{=} \sum_{i=1}^{I}\sum_{k=1}^{J}\sum_{t=1}^{T} y_{i,t}\left(u_{i,\sigma(k)}\log(w_{k,t})+(1-u_{i,\sigma(k)})\log(1-w_{k,t})\right)\\
&\overset{(7)}{=} \sum_{i=1}^{I}\sum_{k=1}^{J}\sum_{t=1}^{T} y_{i,t}\left(\sum_{j=1}^{J} z_{j,k}\left(u_{i,j}\log(w_{k,t})+(1-u_{i,j})\log(1-w_{k,t})\right)\right)\\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{J}\sum_{t=1}^{T} y_{i,t}z_{j,k}\left(u_{i,j}\log(w_{k,t})+(1-u_{i,j})\log(1-w_{k,t})\right).
\end{aligned}$$

We are now in a position to formulate the DM model, i.e., the problem of maximizing the log likelihood function, $\ell(\boldsymbol{W},\boldsymbol{Y},\boldsymbol{Z}\mid\boldsymbol{U})$, subject to constraints (3), (4) and (8)–(12). Accordingly, the DM model can be cast as an MINLP problem,

$$\text{(DMM)}\quad\left|\begin{array}{ll}
\underset{\boldsymbol{W},\boldsymbol{Y},\boldsymbol{Z}}{\text{maximize}} & \displaystyle\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{J}\sum_{t=1}^{T} y_{i,t}z_{j,k}\left(u_{i,j}\log(w_{k,t})+(1-u_{i,j})\log(1-w_{k,t})\right)\\
\text{subject to} & 0\le w_{k,1}\le w_{k,2}\le\cdots\le w_{k,T}\le 1 \quad (\forall k=1,2,\ldots,J),\\
& w_{1,t}\le w_{2,t}\le\cdots\le w_{J,t} \quad (\forall t=1,2,\ldots,T),\\
& \displaystyle\sum_{k=1}^{J} z_{j,k}=1 \quad (\forall j=1,2,\ldots,J),\\
& \displaystyle\sum_{j=1}^{J} z_{j,k}=1 \quad (\forall k=1,2,\ldots,J),\\
& z_{j,k}\in\{0,1\} \quad (\forall j=1,2,\ldots,J,\ \forall k=1,2,\ldots,J),\\
& \displaystyle\sum_{t=1}^{T} y_{i,t}=1 \quad (\forall i=1,2,\ldots,I),\\
& y_{i,t}\in\{0,1\} \quad (\forall i=1,2,\ldots,I,\ \forall t=1,2,\ldots,T).
\end{array}\right.$$

4 Heuristic Optimization Algorithm

The optimization models presented in Section 3 are mixed integer nonlinear programming (MINLP) problems, which are very hard to solve exactly. Because of that, we decided to develop heuristic optimization algorithms for efficiently computing good-quality solutions. An algorithm for solving problem (MHM) is described in this section, and that for solving problem (DMM) is described in Appendix.

We begin by giving an ability class to each examinee as an initial solution. To do this, one may use the number of question items that each examinee answered correctly. We denote the initial ability classes by

$$\bar{\boldsymbol{Y}}=(\bar{y}_{i,t};\ i=1,2,\ldots,I,\ t=1,2,\ldots,T).$$

Next, we solve problem (MHM) in which the decision variable $\boldsymbol{Y}$ is fixed to $\bar{\boldsymbol{Y}}$. Since this problem can be decomposed into ones of each ICC, we solve

$$(\mathrm{MHM}(j \mid \bar{\boldsymbol{Y}})) \quad \left| \begin{array}{ll} \underset{\boldsymbol{x}_j}{\text{maximize}} & \displaystyle\sum_{i=1}^{I}\sum_{t=1}^{T} \bar{y}_{i,t} \left(u_{i,j} \log(x_{j,t}) + (1-u_{i,j}) \log(1-x_{j,t})\right) \\ \text{subject to} & 0 \leq x_{j,1} \leq x_{j,2} \leq \cdots \leq x_{j,T} \leq 1, \end{array} \right.$$

for $j = 1, 2, \ldots, J$. Since problem $(\mathrm{MHM}(j \mid \bar{\boldsymbol{Y}}))$ is a maximization of a concave function with linear constraints, we can solve it exactly and efficiently with a standard nonlinear optimization solver.

Let

$$\bar{\boldsymbol{X}} = (\bar{x}_{j,t};\ j = 1, 2, \ldots, J,\ t = 1, 2, \ldots, T)$$

be composed of optimal solutions to problems $(\mathrm{MHM}(j \mid \bar{\boldsymbol{Y}}))$ for $j = 1, 2, \ldots, J$. Now, we solve problem (MHM) in which the decision variable $\boldsymbol{X}$ is fixed to $\bar{\boldsymbol{X}}$. Similarly to the above problems, this problem can be decomposed into ones of each examinee. Consequently, we solve

$$(\mathrm{MHM}(i \mid \bar{\boldsymbol{X}})) \quad \left| \begin{array}{ll} \underset{\boldsymbol{y}_i}{\text{maximize}} & \displaystyle\sum_{j=1}^{J}\sum_{t=1}^{T} y_{i,t} \left(u_{i,j} \log(\bar{x}_{j,t}) + (1-u_{i,j}) \log(1-\bar{x}_{j,t})\right) \\ \text{subject to} & \displaystyle\sum_{t=1}^{T} y_{i,t} = 1, \\ & y_{i,t} \in \{0, 1\} \quad (\forall t = 1, 2, \ldots, T). \end{array} \right.$$

for $i = 1, 2, \ldots, I$. To solve problem $(\mathrm{MHM}(i \mid \bar{\boldsymbol{X}}))$, it is only necessary to select one ability class t^* such that the objective function is maximized, and set $y_{i,t^*} = 1$. In this manner, we update $\bar{\boldsymbol{Y}}$ and return to the first step to find a better $\bar{\boldsymbol{X}}$.

By repeating this procedure, the log likelihood function, $\ell(\bar{\boldsymbol{X}}, \bar{\boldsymbol{Y}} \mid \boldsymbol{U})$, monotonically increases. We terminate this algorithm when the solution $\bar{\boldsymbol{Y}}$ stops changing. Our heuristic optimization algorithm is summarized as follows:

Algorithm 1: Heuristic Optimization Algorithm for Solving Problem (MHM)

Step 0. (Initialization) Set the initial ability classes, $\bar{\boldsymbol{Y}}$.

Step 1. (ICC Estimation) Solve problems $(\mathrm{MHM}(j \mid \bar{\boldsymbol{Y}}))$ for all $j = 1, 2, ..., J$. Let $\bar{\boldsymbol{X}}$ be an optimal solution.

Step 2. (Ability Estimation) Solve problems $(\mathrm{MHM}(i \mid \bar{\boldsymbol{X}}))$ for all $i = 1, 2, ..., I$. Let $\bar{\boldsymbol{Y}}$ be an optimal solution.

Step 3. (Termination Condition) If $\bar{\boldsymbol{Y}}$ remains the same as the previous one, terminate the algorithm with the solution $(\bar{\boldsymbol{X}}, \bar{\boldsymbol{Y}})$. Otherwise, return to Step 1.

5 Computational Experiments

The computational results reported in this section compare the effectiveness of our NIRT models with that of the common PIRT model.

5.1 Experimental design

The number of examinees, I, was set to 1000 and 3000, and the number of question items, J, was set to 30 and 60, similarly to Nozawa [14]. Since the ordinal scale of neural test theory grades examinees into about ten classes (see, e.g., [17, 18]), the number of ability classes, T, was set to ten.

We evaluated the IRT models through the process illustrated in Figure 2.

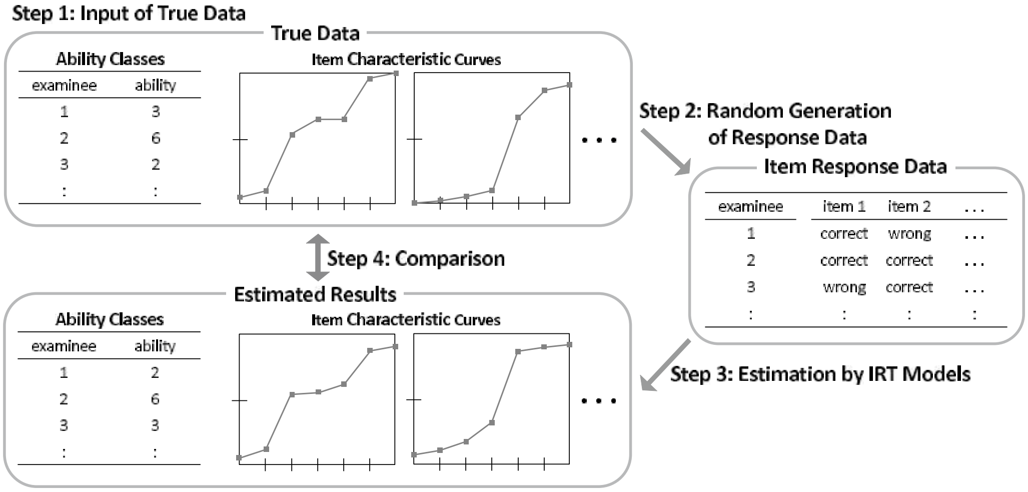

Figure 2: Process of Model Evaluation

In Step 1, we randomly generated θ_i from a standard normal distribution for $i = 1, 2, \ldots, I$. Next, we converted θ_i into an ability class t in view of the second column "range of θ" of Table 1. For instance, if $0 \leq \theta_i < 0.23$, we gave a true ability class $t_i^{\mathrm{true}} = 6$ to examinee i. The ranges of θ were determined so that each ability class is assigned to approximately the same number of examinees.

To define the ICCs of question items $j = 1, 2, \ldots, J$, we used two types of function. One was the two-parameter logistic (2PL) model,

$$p_j^{\mathrm{2PL}}(\theta) = \frac{1}{1 + \exp(-1.7a_j(\theta - b_j))}, \tag{13}$$

Table 1: Relationship between the Ability Class t and the Continuous Value θ

t	range of θ	median of θ
1	$[-\infty, -1.29)$	-1.73
2	$[-1.29, -0.81)$	-1.02
3	$[-0.81, -0.49)$	-0.64
4	$[-0.49, -0.23)$	-0.36
5	$[-0.23, 0)$	-0.12
6	$[0, 0.23)$	0.12
7	$[0.23, 0.49)$	0.36
8	$[0.49, 0.81)$	0.64
9	$[0.81, 1.29)$	1.02
10	$[1.29, \infty)$	1.73

where a_j and b_j are parameters of discrimination and difficulty that are uniformly drawn from the respective intervals $[0.5, 2.0]$ and $[-1.5, 1.5]$. Similarly to Nozawa [14], the other was the extended three-parameter normal ogive (3PN) model of order two,

$$p_j^{\mathrm{3PN}}(\theta) = \Phi(a_{j,2}(\theta - b_j)^3 + \sqrt{3a_{j,1}a_{j,2}}(\theta - b_j)^2 + a_{j,1}(\theta - b_j)), \tag{14}$$

where Φ is the normal ogive; $a_{j,1}$ and $a_{j,2}$ are shape parameters; and b_j is a parameter of difficulty. These parameters, $a_{j,1}$, $a_{j,2}$ and b_j, are uniformly drawn from the intervals $[0.4, 0.8]$, $[0.1, 0.5]$, and $[-0.5, 0.5]$. This model defines ICCs based on the multimodal distribution of the examinees' abilities. Although two-parameter logistic IRT models can accurately estimate ICCs defined by the 2PL model, they have difficulty in fitting ICCs defined by the 3PN model.

The third column "median of θ" of Table 1 shows the median of the corresponding range of θ. When the true ICC of question item j was based on the 2PL model (13), it was defined as $x_{j,1}^{\mathrm{true}} = p_j^{\mathrm{2PL}}(-1.73), x_{j,2}^{\mathrm{true}} = p_j^{\mathrm{2PL}}(-1.02), \ldots, x_{j,10}^{\mathrm{true}} = p_j^{\mathrm{2PL}}(1.73)$ in correspondence with the median values of Table 1. The true ICCs based on the 3PN model (14) were defined in the same way. We denote by ρ the percentage of ICCs defined by the 3PN model, and we set ρ to 0%, 20% and 50% in the manner of Nozawa [14]. For instance, when $J = 60$ and $\rho = 20\%$, true ICCs of 12 question items were created by the 3PN model.

In Step 2, item response data, $\boldsymbol{U}$, was randomly generated with a binomial distribution based on the data from Step 1. Precisely, examinees of ability class t answered item j correctly with probability $x_{j,t}^{\mathrm{true}}$.

In Step 3, ability classes and ICCs were estimated using the following IRT models from the item response data $\boldsymbol{U}$,

2PLM: two-parameter logistic IRT model,

MHM: monotone homogeneity model (MHM),

DMM: double monotonicity model (DMM).

We used EasyEstimation Ver. 1.4.3 (`http://irtanalysis.main.jp/english`), a program for analyzing IRT models, to perform computations of 2PLMs. For comparison, a continuous ability θ_i estimated by 2PLM was converted into an ability class t in view of the second column of Table 1. We used Algorithm 1 to solve optimization model (MHM) and a similar heuristic optimization algorithm (see Appendix) to solve optimization model (DMM). MATLAB R2011b (`http://www.mathworks.com/products/matlab`) and a MATLAB optimization toolbox, fmincon, were used to implement these heuristic optimization algorithms. In these algorithms, examinees were equally divided into ten groups based on the number of correct answers, and the initial $\bar{\boldsymbol{Y}}$ was set by assigning one ability class to each group. The heuristic optimization algorithms employed the following MH constraints:

$$\begin{aligned} &0.01 \le x_{j,1} \le x_{j,2} \le \cdots \le x_{j,T} \le 0.99 \quad (\forall j = 1, 2, \ldots, J), \\ &0.01 \le w_{k,1} \le w_{k,2} \le \cdots \le w_{k,T} \le 0.99 \quad (\forall k = 1, 2, \ldots, J) \end{aligned}$$

to avoid numerical instabilities caused by $\log(\,\cdot\,)$ going to $-\infty$.

In Step 4, we evaluated the estimation accuracy of each IRT model by comparing the data generated in Step 1 with the estimates obtained in Step 3. We took the root mean square error (RMSE) to be the measure for the evaluation. The RMSE of the ability classes was calculated as follows:

$$\text{RMSE of ability classes} = \sqrt{\frac{\sum_{i=1}^{I} (t_i^{\text{true}} - \hat{t}_i)^2}{I}},$$

where $\hat{t}_i$ is the estimated ability class. The RMSE of ICCs was calculated as follows:

$$\text{RMSE of ICCs} = \sqrt{\frac{\sum_{j=1}^{J} \sum_{t=1}^{T} (x_{j,t}^{\text{true}} - \hat{x}_{j,t})^2}{JT}},$$

where $\hat{x}_{j,t}$ is the estimated probability of a correct answer. We repeated Steps 1 to 4 ten times and show the average RMSE in what follows.

5.2 Computational results

Tables 2 and 3 show the RMSEs of the ability classes and ICCs for the 12 experimental conditions. Note that the minimum RMSE for each experimental condition is bold-faced in the tables.

Table 2: Root Mean Square Error of Ability Classes

I	J	ρ	2PLM	MHM	DMM
1000	30	0%	0.796	0.883	**0.795**
		20%	**0.826**	0.905	0.835
		50%	1.009	1.035	**0.951**
	60	0%	0.619	0.610	**0.580**
		20%	0.680	0.652	**0.630**
		50%	0.826	**0.680**	0.681
3000	30	0%	0.787	0.901	**0.784**
		20%	0.837	0.950	**0.825**
		50%	0.942	0.979	**0.898**
	60	0%	0.630	0.627	**0.585**
		20%	0.668	0.629	**0.609**
		50%	0.834	0.705	**0.676**

Table 3: Root Mean Square Error of Item Characteristic Curves

I	J	ρ	2PLM	MHM	DMM
1000	30	0%	**0.025**	0.068	0.047
		20%	**0.049**	0.070	0.059
		50%	0.079	0.073	**0.054**
	60	0%	**0.022**	0.047	0.048
		20%	0.051	**0.048**	0.063
		50%	0.080	**0.050**	0.063
3000	30	0%	**0.015**	0.066	0.042
		20%	**0.046**	0.068	0.055
		50%	0.074	0.067	**0.060**
	60	0%	**0.016**	0.038	0.046
		20%	0.047	**0.038**	0.059
		50%	0.078	**0.041**	0.055

We can see from Table 2 that when the number of question items was 30, the RMSE of the ability classes obtained by MHM was larger than that of 2PLM. When the number of question items was 60, on the other hand, MHM had a smaller RMSE than 2PLM did, and the difference got larger as the percentage of 3PN ICCs increased. As for the RMSE of the ICCs in Table 3, when the percentage of 3PN ICCs was 0%, MHM was always worse than 2PLM. Conversely, when the percentage of 3PN ICCs was 50%, MHM was always better than 2PLM. MHM has the potential of fitting ICCs based on the 3PN model (14) well, but its estimation results may be unstable when there is not much item response data. Thus, when the number of question items and percentage of 3PN ICCs were sufficiently large, nonparametric MHM outperformed parametric 2PLM.

The estimation accuracy of MHM was not always high, mostly because of overfitting. In contrast, DMM attained the minimum RMSE of the ability classes for 10 experimental conditions in Table 2. However, as shown in Table 3, it failed to estimate the ICCs accurately. Indeed, when the number of question items was 60, DMM had a larger RMSE for the ICCs than MHM did. This is because the true ICCs did not satisfy the DM constraints, and consequently, DMM had difficulty estimating them.

Figures 3 and 4 show illustrative examples of estimated ICCs together with the true ICCs for $(I, J, \rho) = (3000, 60, 50\%)$. The true ICC was defined by the 3PN model in Figure 3 and by the 2PL model in Figure 4. It is clear from Figure 3 that the ICC estimated by 2PLM did not fit the true 3PN-based ICC well because 2PLM can only create a logistic curve. On the other hand, the other nonparametric IRT models estimated relatively accurate shapes of the true ICC. Figure 4 reveals that the ICC estimated by DMM was very different from the true 2PL-based one because the DM constraints are too tight. Additionally, we should notice that the ICC estimated by MHM moved away from the true ICC for the ability classes $t = 2$, 3, 5 and 6. Meanwhile, it is reasonable that 2PLM estimated the true 2PL-based ICC very accurately.

6 Conclusions

This paper described a mathematical optimization approach to nonparametric item response theory (NIRT). Specifically, to estimate nonparametric item characteristic curves (ICCs) and latent abilities of examinees simultaneously, we developed mathematical optimization models and heuristic optimization algorithms. The computational results demonstrated that NIRT models outperformed the common two-parameter logistic IRT model especially when many ICCs were based on a multimodal ability distribution.

The contributions of this research are twofold. First, we formulated mathematical optimization models to determine the nonparametric ICCs and the latent abilities of examinees simultaneously under the monotone homogeneity and double monotonicity constraints. Second, we developed heuristic optimization algorithms to efficiently find good-quality solutions to the

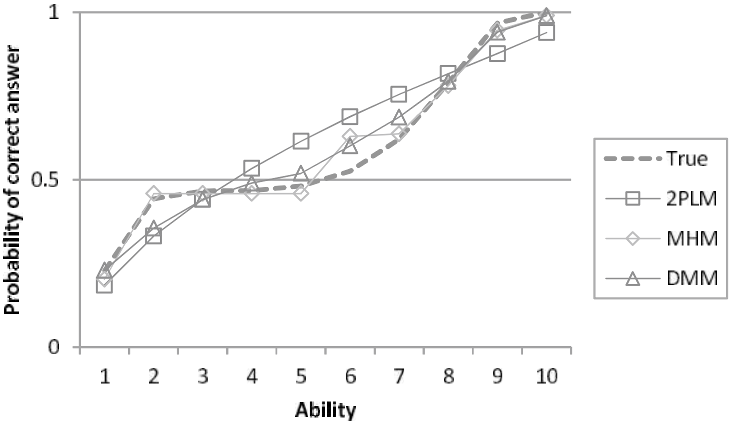

Figure 3: Estimated Item Characteristic Curves Together with the True 3PN (Extended Three-Parameter Normal Ogive) One

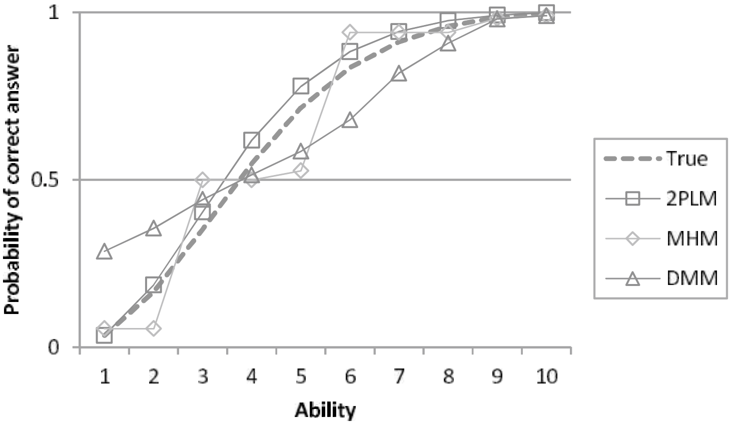

Figure 4: Estimated Item Characteristic Curves Together with the True 2PL (Two-Parameter Logistic) One

NIRT models. By means of these algorithms, we verified the effectiveness of our mathematical optimization models for NIRT.

This study illustrates the fact that the mathematical optimization approach can be a powerful tool for nonparametric ICC estimation. Mathematical optimization models make it possible to estimate ICCs under the various effective constraints. Indeed, the double monotonicity constraint is useful for improving the estimation accuracy of the latent abilities.

A future direction of study will be to extend our formulation to polytomous NIRT models (see, e.g., [20]). In addition, there is room for further research into algorithms especially for solving the double monotonicity model.

Acknowledgments

The authors wish to thank Yoshitsugu Yamamoto, University of Tsukuba, and Tomomi Matsui, Tokyo Institute of Technology for helpful suggestions and comments. This work was supported by a Grant-in-Aid for Young Scientists ((B) 26750114) by the Japan Society for the Promotion of Science.

Appendix

This appendix describes a heuristic optimization algorithm for solving the optimization model (DMM). Step 0 and Step 1 are the same as those of Algorithm 1. In Step 2, we determine a difficulty ranking of question items on the basis of the estimated ICCs. Specifically, for all question items $j = 1, 2, \ldots, J$, we calculate the sum of probabilities of the correct answer, $\bar{x}_j^{\mathrm{sum}} = \sum_{t=1}^{T} \bar{x}_{j,t}$. If $\bar{x}_j^{\mathrm{sum}}$ is small, the question item j is difficult to answer correctly; accordingly, we set a difficulty ranking such that if $\bar{x}_j^{\mathrm{sum}}$ is the k-th smallest of all question items, then $\bar{z}_{j,k} = 1$. Next, we estimate the ICCs again by solving the following optimization problem under the DM constraints given the difficulty ranking,

$$
(\mathrm{DMM}(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{Z}})) \quad \left|
\begin{array}{ll}
\underset{\boldsymbol{W}}{\text{maximize}} & \displaystyle\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{J}\sum_{t=1}^{T} \bar{y}_{i,t}\bar{z}_{j,k}\left(u_{i,j}\log(w_{k,t}) + (1-u_{i,j})\log(1-w_{k,t})\right) \\
\text{subject to} & 0 \le w_{k,1} \le w_{k,2} \le \cdots \le w_{k,T} \le 1 \quad (\forall k = 1, 2, \ldots, J), \\
 & w_{1,t} \le w_{2,t} \le \cdots \le w_{J,t} \quad (\forall t = 1, 2, \ldots, T).
\end{array}
\right.
$$

The next step is similar to Step 2 of Algorithm 1. We solve the following optimization problems to determine the ability classes of the examinees,

$$
(\mathrm{DMM}(i \mid \bar{\boldsymbol{W}}, \bar{\boldsymbol{Z}})) \quad \left|
\begin{array}{ll}
\underset{\boldsymbol{y}_i}{\text{maximize}} & \displaystyle\sum_{j=1}^{J}\sum_{k=1}^{J}\sum_{t=1}^{T} y_{i,t}\bar{z}_{j,k}\left(u_{i,j}\log(\bar{w}_{k,t}) + (1-u_{i,j})\log(1-\bar{w}_{k,t})\right) \\
\text{subject to} & \displaystyle\sum_{t=1}^{T} y_{i,t} = 1, \\
 & y_{i,t} \in \{0, 1\} \quad (\forall t = 1, 2, \ldots, T),
\end{array}
\right.
$$

for $i = 1, 2, ..., I$. These problems are easily solved similarly to $(\mathrm{MHM}(i \mid \bar{\boldsymbol{X}}))$.

Finally, we obtain the solution $(\bar{\boldsymbol{W}}, \bar{\boldsymbol{Y}}, \bar{\boldsymbol{Z}})$. We do not return to Step 1 because our preliminary experiment showed that such a repetition did not improve a solution significantly. This heuristic optimization algorithm is summarized as follows:

Algorithm 2: Heuristic Optimization Algorithm for Solving Problem (DMM)

Step 0. (Initialization) Set the initial ability classes, $\bar{\boldsymbol{Y}}$.

Step 1. (Tentative ICC Estimation) Solve problems (MHM($j \mid \bar{\boldsymbol{Y}}$)) for all $j = 1, 2, ..., J$. Let $\bar{\boldsymbol{X}}$ be an optimal solution.

Step 2. (Difficulty Ranking Estimation) Set a difficulty ranking, $\bar{\boldsymbol{Z}}$, such that if $\bar{x}_j^{\text{sum}} = \sum_{t=1}^{T} \bar{x}_{j,t}$ is the k-th smallest of all question items, then $\bar{z}_{j,k} = 1$.

Step 3. (ICC Estimation with DM Constraints) Solve problem (DMM($\bar{\boldsymbol{Y}}, \bar{\boldsymbol{Z}}$)). Let $\bar{\boldsymbol{W}}$ be an optimal solution.

Step 4. (Ability Estimation) Solve problems (DMM($i \mid \bar{\boldsymbol{W}}, \bar{\boldsymbol{Z}}$)) for all $i = 1, 2, ..., I$. Let $\bar{\boldsymbol{Y}}$ be an optimal solution.

Step 5. (Termination) Terminate the algorithm with the solution ($\bar{\boldsymbol{W}}, \bar{\boldsymbol{Y}}, \bar{\boldsymbol{Z}}$).

References

[1] Croon, M. A. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology, 43*, 171–192.

[2] Croon, M. A. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology, 44*, 315–331.

[3] Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika, 62*, 7–28.

[4] Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement, 25*, 234–243.

[5] Johnson, M. S. (2007). Modeling dichotomous item responses with free-knot splines. *Computational Statistics & Data Analysis, 51*, 4178–4192.

[6] Junker, B., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement, 25*, 211–220.

[7] Lee, Y.-S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement, 31*, 121–134.

[8] Lee, Y.-S., Wollack, J. A., & Douglas, J. (2009). On the use of nonparametric item characteristic curve estimation techniques for checking parametric model fit. *Educational and Psychological Measurement, 69*, 181–197.

[9] Liang, T. & Wells, C. S. (2009). A model fit statistic for generalized partial credit model. *Educational and Psychological Measurement, 69*, 913–928.

[10] Meredith, W. (1965). Some results based on a general stochastic model for mental tests. *Psychometrika, 30*, 419–440.

[11] Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research.* New York, Berlin: Walter de Gruyter, Mouton.

[12] Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*, 417–430.

[13] Molenaar, I. W. (2001). Thirty years of nonparametric item response theory. *Applied Psychological Measurement, 25*, 295–299.

[14] Nozawa, Y. (2008). *Comparison of parametric and nonparametric IRT equating methods under the common-item nonequivalent groups design (Doctoral dissertation).* Available from ProQuest Dissertations and Theses database. (UMI No. 3347237)

[15] Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611–630.

[16] Rossi, N., Wang, X., & Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of Educational and Behavioral Statistics, 27*, 291–317.

[17] Shojima, K. (2007). *Neural test theory* (DNC Research Note No. 07-02). The National Center for University Entrance Examinations.

[18] Shojima, K. (2008). *Neural test theory: A latent rank theory for analyzing test data* (DNC Research Note No. 08-01). The National Center for University Entrance Examinations.

[19] Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 22*, 3–31.

[20] Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage.

[21] Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589–617.

[22] Stout, W. F. (2001). Nonparametric item response theory: A maturing and applicable measurement modeling approach. *Applied Psychological Measurement, 25*, 300–306.

[23] Sueiro, M. J., & Abad, F. J. (2011). Assessing goodness of fit in item response theory with nonparametric models: A comparison of posterior probabilities and kernel-smoothing approaches. *Educational and Psychological Measurement, 71*, 834–848.

[24] van Onna, M. J. H. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika, 67*, 519-538.

[25] Wells, C. S., & Bolt, D. M. (2008). Investigation of a nonparametric procedure for assessing goodness-of-fit in item response theory. *Applied Measurement in Education, 21*, 22–40.