Are All Item Response Functions Monotonically Increasing?

BY

Wenhao Wang

Submitted to the graduate degree program in the Department of Psychology and Research in Education and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairperson Neal Kingston

Bruce Frey

John Poggio

William Skorupski

Carol Woods

Date Defended: 4/6/12

The Dissertation Committee for Wenhao Wang

certifies that this is the approved version of the following dissertation:

Are All Item Response Functions Monotonically Increasing?

Chairperson Neal Kingston

Date approved: 4/18/12

Abstract

Item response functions of the parametric logistic IRT models follow the logistic form which is monotonically increasing. However, item response functions of some real items are nonmonotonic which might lead to examinees with lower proficiency levels receiving higher scores. This study compared three nonparametric IRF estimation methods—the nonparametric smooth regression method, the item-ability regression method, and the B-spline nonparametric IRF method—to determine whether they could detect the nonmonotonic IRF accurately using simulated data. In addition, these methods were used to identify items with nonmonotonic IRFs on real assessments. Results present that three nonparametric methods can detect the nonmonotonic IRF equally and each real assessment has some items with nonmonotonic IRFs. Investigations on the reasons for and the consequences of the nonmonotonicity were conducted for several items and indicate that the nonmonotonicity can affect the fairness and comparability of the test score. Thus, the nonmonotonicity should be checked before applying the parametric logistic models.

Acknowledgments

I am grateful to all the people who support, encourage, and care about me during the dissertation work.

I would like to express my sincere thanks and appreciation to my advisor Dr. Neal Kingston. He helped me to start my dissertation work when I just got into my graduate program. I learned knowledge and gained wise advice related to my dissertation from him. Dr. Kingston also edited my dissertation with great patience. Most importantly, he taught me the greatest motivation for studying and working in this field is interest, helped me find the purpose to work in this field, which is to use assessment to help students' learning, and made me realize critical thinking is so important. I also would like to thank the other members of my dissertation committee: Dr. Bruce Frey, Dr. John Poggio, Dr. William Skorupski, and Dr. Carol Woods. Especial thanks to Dr. Skorupski for his support at the beginning of my study and to Dr. Woods for her help on the functional analysis. I am also grateful to Dr. Rob McKinley, Dr. Angela Broaddus, Dr. Patrick Irwin, reading item development group, and editors from CETE. Especially, I would like to thank Dr. McKinley for his help on item-ability regression method and Dr. Broaddus for her help on detailed analysis on math items from content aspect.

I also would like to give my thanks to God for "my times are in Your hands". My parents have always been very supportive and considering during my graduate study. Thanks to all my friends in Lawrence who makes my studies in the United States more fun and less lonely.

TABLE OF CONTENTS

List of Tables

Table Page

List of Figures

Figures Pages

List of Equations

Chapter 1: Introduction

Background

Item response theory (IRT) uses mathematical models to describe the relationship between the probability of an examinee's item responses and latent trait level, the latter referred to as proficiency. The regression of the probability of a correct response on the latent trait is called the item response function (IRF) (De Ayala, 2008, p. 16; Lord, 1980, p. 12). The most commonly used IRT models are the logistic models with one, two, or three parameters. These models have two advantages: they can be expressed concisely and parameters estimation is relatively easy. These models have proven to be powerful tools through their application in various areas: constructing test and item banks, scoring, reporting scores, scaling, equating test scores, applying adaptive tests, and identifying differential item functions (De Ayala, 2008, p. 316; Hambleton & Swaminathan, 1985, p. 7).

Before applying the parametric logistic IRT models, two assumptions must be met (Lord, 1980, p. 15): local independence and the logistic form of IRFs. Local independence requires the number of dimensions in the model matches the number of dimensions in the data. In this study only one-dimensional models were considered. The logistic form of IRFs is defined as that the IRF of the parametric logistic IRT model follows a logistic form with one to three parameters. This logistic form is monotonically increasing, which means that the IRF is not decreasing as the latent trait increases (Lord, 1980, p. 12). The fit assessments of the parametric logistic IRT models should always include the investigation of whether the model assumptions are met (Hambleton & Swaminathan, 1985, p. 151; Hambleton, Swaminathan, & Rogers, 1991, p. 53). The model fit is the degree to which the model adequately explains the data. If the models do not fit the data because the assumptions were not met, this issue could lead to reduced validity of interpretations one could make from test scores.

Statement of the Problem

Parametric logistic IRT model fit. Some statistical tests have been used to provide quantitative evidence of the model fit. The most commonly used statistical test is the Pearson chi-square test, which compares the observed item score and model predicted item score at several proficiency levels (Yen, 1981). McKinley and Mills (1985) and Orlando and Thissen (2000) introduced the likelihood-based statistics to access the model fit. Recently, some researchers have run the same model fit statistical tests from a Bayesian perspective through the posterior predictive model checking method (Glas & Meijer, 2003; Levy, Mislevy, & Sinharay, 2009). Researchers have applied these statistical tests on the data to provide evidence of the model fit (Yen, 1981; Orlando & Thissen, 2000; Glas & Meijer, 2003).

However, these statistical tests can only answer the global yes/no question about whether certain data items fit the model. If the model does not fit with the data, these statistics cannot explain why the model is not a good fit. Because of this, an exploratory diagnostic assessment of the model fit is needed.

Applying graphical representation methods may meet the need by providing diagnostic information on why the model does not fit the data; in fact, several researchers have presented different methods for using graphical representations to address the need (Douglas & Cohen, 2001; Kingston & Dorans, 1985; Lord, 1970). Graphical representation methods estimate the IRF without assuming any of its mathematical form, and then the estimated IRF is compared with the parametric logistic IRF. Since estimating IRF without assuming any mathematical form is more flexible, practically significant differences between two IRFs are visible and indicate that the parametric logistic model does not fit the data (Junker & Sijtsma, 2001; Molenaar, 2001; Sijtsma & Junker, 2006; Stout, 2001). Lord (1970) used the true-score regression method to plot the IRF on the scale of the true score and then used a test characteristic curve to transform the scale to the theta scale. Later, Kingston and Dorans (1985) used the item-ability regression method to estimate IRF. They first estimated theta from the parametric logistic model, and then grouped the theta estimates based on empirical experiences. Once this was done, the estimated IRF was formed using the ratio of examinees correctly answering the item to examinees answering the item. In addition, Douglas and Cohen (2001) used the nonparametric smooth regression method to estimate the IRF on the scale of theta based on the examinees' response patterns and sum scores of the test. Research from these graphical representation methods provided very useful diagnostic information on the model fit of the parametric logistic IRT models of real data (Lord, 1970; Kingston & Dorans, 1985; Douglas & Cohen, 2001).

Previous research did not focus on assessing the monotonicity of the IRF because the simulated data, while not necessarily fitting these parametric logistic IRT models, were monotonically increasing. However, researchers did present some real items with nonmonotonic IRFs (Kingston & Doran, 1985).

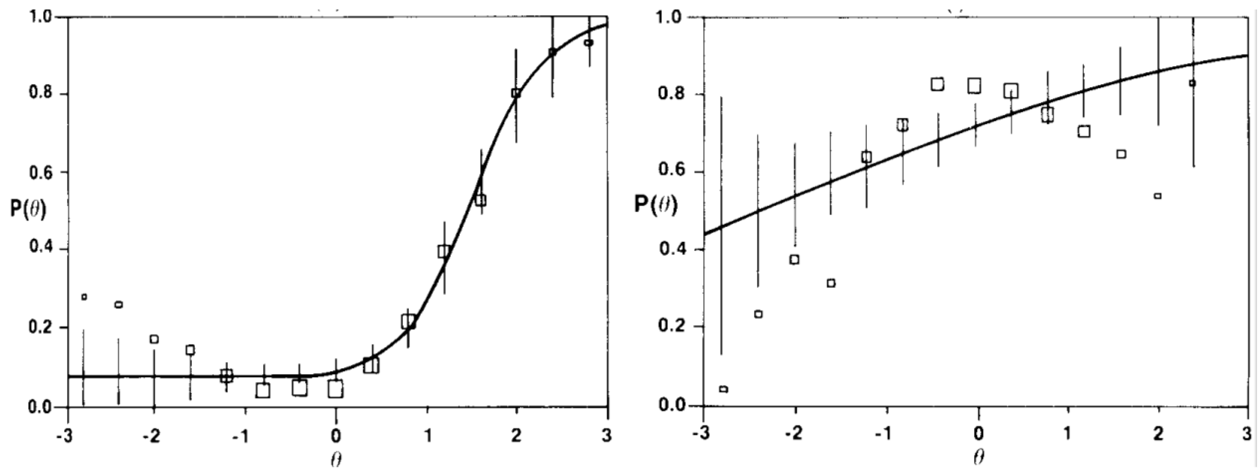

*Figure 1.*Two real items with the nonmonotonic IRF.Reprinted from “The analysis of item-ability regressions: An exploratory IRT model fit tool,” by N. M. Kingston and N. J. Dorans, 1985, *Applied Psychological Measurement, 4,* p. 283. The squares formed the estimated IRF.

A lack of monotonicity might suggest the item is not functioning properly and should be removed. Alternatively, if the nonmonotonicity is not properly modeled, examinees with lower proficiency levels could receive higher scores. Thus, the author’s study aimed to check whether there are IRFs on existing achievement tests that are not monotonically increasing.

Monotonic IRF checking methods. The model fit checking method focusing on the monotonicity of the IRF should be able to estimate the true IRF from the observed data without assuming monotonicity. If one item has an obvious nonmonotonic IRF estimated from these methods, the nonmonotonic IRF presents that this item does not meet the assumption that the IRF is monotonically increasing. Nonparametric smooth regression, item-ability regression, and B-spline nonparametric IRT are three methods which can estimate the nonmonotonic IRF. One goal of this study was to test whether these three methods can detect the nonmonotonic IRF accurately through simulated data.

Posterior predictive model checking. If the simulated data are studied, then the true IRFs are known and can be used to generate the null distribution for evaluating whether the estimated IRFs are monotonic. Also, the true IRFs are criteria used to compare three monotonic IRF checking methods and to judge whether these methods detect the nonmonotonic IRF accurately. However, under a real data situation, the true IRF can never be known because the latent variable cannot be directly observed (Douglas, 1997). Even if one IRF estimated by three monotonic IRF checking methods is nonmonotonic, which is different from the IRF of the parametric logistic model, it is not true to conclude that the item has nonmonotonic IRF and the parametric logistic model does not fit data. One additional step during the real data situation should be conducted to judge whether the difference is large enough to conclude that the parametric logistic model is misfit. The Posterior predictive model checking (PPMC) method could serve this purpose.

The PPMC method (Gelman, Meng, & Stern, 1996; Guttman, 1967; Rubin, 1984) is a statistical inference method from a Bayesian perspective. It compares the statistic of the observed data to the distribution of the same statistic calculated from replicated data. The replicated data are generated or predicted by the model being checked and posterior distributions of parameters. Glas and Meijer (2003) used this method to analyze the person fit in IRT models; similarly, Levy, Mislevy, and Sinharay (2009) used it to check the dimensionality assumption of the multidimensional IRT models; and furthermore, Sinharay (2005) used this method to conduct several model fit tests of the IRT parametric logistic models. Since the PPMC method is a popular and flexible Bayesian model checking tool (Sinharay, 2005), this study used the PPMC method to judge whether the IRF is monotonically increasing in the real data study.

Purpose

The purposes of this study were to test whether three monotonic IRF checking methods can accurately detect the nonmonotonic IRF using simulated data and to apply these methods on real data. The three monotonic IRF checking methods are the nonparametric smooth regression method, the item-ability regression method, and the B-spline nonparametric IRT method. The

PPMC method was used to determine the extent of nonmonotonicity of IRFs estimated by three monotonic IRF checking methods in the real data study, which was discussed in more detail in the next chapter.

Variables. The independent variables were three monotonic IRF checking methods. The dependent variables were the differences between the estimated IRFs and the true IRFs for the simulated data, and the extent of the nonmonotonicity of IRFs estimated by three monotonic IRF checking methods for the real data.

Research questions. This study addressed two research questions:

1. Can nonparametric smooth regression, item-ability regression, and B-spline nonparametric IRT detect the nonmonotonic IRF accurately in a simulated data study?
2. Do real items with nonmonotonic IRFs exist, and if so, how common are they?

Hypotheses. The hypotheses, which correspond to the three research questions, were as follows:

1. Nonparametric smooth regression, item-ability regression, and B-spline nonparametric IRT can detect the nonmonotonic IRF in a simulated data study equally well.
2. Real items with the nonmonotonic IRF exist.

Summary

This study aimed to identify and compare three monotonic IRF checking methods using simulated data. This study also used the PPMC method to determine the extent of the nonmonotonicity.

Chapter 2: Literature Review

This chapter provides background on the estimation of item response function (IRF), a discussion of the common assumption among methods used in this study, and a description of data features used in this study. Other topics include background on parametric logistic IRT models, three methods used to estimate non-monotonic IRFs, and the use of PPMC method for judging the nonmonotonicity.

Background on IRF Estimation

Researchers have been interested in methods that can estimate IRFs without using the parametric logistic IRT models. These models are separated into two categories: parametric methods and nonparametric methods.

Parametric IRF estimation method. One parametric model that allows nonmonotonic form is Thissen and Steinberg's (1984) multiple choice model. This model is an extension of Bock's (1972) and Samejima's (1979) nominal response model. For an item with *m* possible responses, the probability that a randomly selected examinee with the proficiency θ give response *h* (from 1 to *m*) is as follows:

$$P(x = h|\theta) = \frac{\exp(a_h\theta + c_h)}{\sum_{k=0}^{m}\exp(a_k\theta + c_k)} + \frac{\mathrm{d_h}\exp(a_0\theta + c_0)}{\sum_{k=0}^{m}\exp(a_k\theta + c_k)}\,. \tag{1}$$

This model counts the guessing and includes two parts. The first part of the formula is the probability of giving response *h* without guessing. The second part of the formula is the guessing probability of giving response *h:* $a_0\theta+c_0$, and d_h are for the "don't know" response category, which means the examinee is "undecided" of the response to choose (Thissen & Steinberg, 1984).

Nonparametric IRF estimation method. Even if Thissen and Steinberg's (1984) model is able to estimate nonmonotonic IRFs, it is still constrained by the parameters and is not very flexible (Junker & Sijtsma, 2001). When the purpose is to check the model fit, more flexible methods might be needed. This leads us to nonparametric IRF estimation methods. They are known for their flexibility and can be used to check the model fit (Junker & Sijtsma, 2001; Molenaar, 2001; Stout, 2001).

Nonparametric IRF estimation methods are exploratory methods used to assess the fit of the parametric logistic IRT models because they can adapt to irregularities in the data (Junker & Sijtsma, 2001). Some methods, like the monotone homogeneity method (Junker & Sijtsma, 2001; Mokken, 1971), assume the monotonic IRF and cannot be used to detect the items with nonmonotonic IRFs. Conversely, five methods do not assume the monotonic IRF and can be used to detect items with nonmonotonic IRFs.

One nonparametric IRF estimation method that does not assume the monotonic IRF came from Lord and Novick’s (1968, p. 363) item-test regression method, which estimated the proportion of the number of examinees who give the correct response to the number of examinees at every observed score per item. Since this method requires a large size sample and the item score is regressed on the manifest variable instead of the latent variable, this research will not compare the item-test regression method.

Subsequently, Lord (1970) modified the item-test regression method when he plotted the IRF on a scale of the true score through estimating the true-score distribution. The scale was then transformed to the theta scale from the true-score scale by the test characteristic curve, a method called the true-score regression. However, the true-score distribution estimation might constrain the range of the true score and lead to the estimated IRF regressing on a small range of theta. These procedures constrained the flexibility of the true-score regression method and thus, this research will not compare the true-score regression method.

In addition, Kingston and Dorans (1985) used the item-ability regression method to estimate IRF. They first estimated theta from the traditional parametric logistic IRT model and then grouped the estimated theta into several groups based on empirical experiences. The estimated IRF was formed using the ratio of examinees correctly answering the item to examinees answering the item. Because the IRF estimation could be affected by the arbitrary theta grouping, this item-ability regression method is modified and compared in this research.

A fourth nonparametric IRF estimation method that does not assume the monotonic IRF came from Ramsay (1991) and Douglas and Cohen (2001), who used the nonparametric smooth regression method to estimate the IRF on the scale of theta based on the examinees' response patterns and sum scores of the test. The nonparametric smooth regression method is a "relatively easy-to-implement nonparametric regression method" and an "exploratory tool for assessing IRF monotonicity" (Junker & Sijtsma, 2001, p. 213). Thus, this nonparametric smooth regression method is compared in this research.

Lastly, Rossi, Wang, and Ramsay (2002) presented a nonparametric IRF estimation method that used the B-spline basis to form a linear IRF and transformed it into a logistic form. This method is flexible to any IRF shape (Rossi et al., 2002; Sijtsma & Junker, 2006). Thus, Rossi et al.'s (2002) B-spline nonparametric IRT method is also compared in this research.

In conclusion, this study compared three nonparametric IRF estimation methods, which do not assume the monotonic IRF. These three methods—nonparametric smooth regression method, item-ability regression method, and B-spline nonparametric IRT method—were discussed in more detail.

Common Assumptions and Data Description

The common assumption of the three nonparametric methods and the parametric logistic IRT models is local independence. Because these models are based on a single underling construct, local independence is equivalent to unidimensionality (De Ayala, 2008, p.20; Douglas, 1997; Hambleton & Swaminathan, 1985, p.25; Hambleton et al., 1991, p.10; Lord, 1980, p.16; Rossi et. al., 2002; Ramsay, 1996; Sijtsma & Junker, 2006). Unidimensionality states that the proficiency to influence the performance of an item is one dimensional. Local independence states that the probability of giving one response on one item is independent with the probability of giving one response on another item when the proficiency level is held constant (Lord, 1980, p.17). All the methods discussed in this study were applied to dichotomous data. The majority of the IRT models handle dichotomous data (Hambleton & Swaminathan, 1985, p.34): 1 for correct response and 0 for incorrect response (Glöckner-Rist & Hoijtink, 2003).

Parametric Logistic IRT Models

The three most commonly used parametric logistic IRT models are the one-, two-, and three -parameter logistic model—the 1-PL model, 2-PL model and 3-PL model, respectively (De Ayala, 2008, p.152; Hambleton & Swaminathan, 1985, p.34; Hambleton et al., 1991, p.12). These three parametric logistic IRT models have similar mathematical expressions, but the number of parameters of each model is different.

The 1-PL model is also called the Rasch model (Rasch, 1960). Subsequently, Birnbaum (1968) proposed the 2-PL model and the 3-PL model, which have more parameters and can characterize the item more accurately than the 1-PL model. But the 2-PL and the 3-PL model also require more data to estimate the parameters. Sometimes the simpler 1-PL model is used because of the estimation difficulty of the 2- and 3-PL models. The 1-, 2-, and 3-PL model were used to replace the one-, two-, and three-parameter normal ogive model proposed by Lord (1952) because of their simplicity (De Ayala, 2008, p.14). The difference in the parameter values of the corresponding normal ogive model and parametric logistic model is less than 0.01 (Lord, 1980, p.14).

The IRF is assumed to follow the mathematical expression of the parametric logistic model being used. For example, if the 3-PL model is used, the IRF of the item is assumed to follow the 3-PL model. The following is the 3-PL model expression. The IRF of 3-PL model follows this expression and has a logistic form (Lord, 1980, p.12):

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-D * a_i * (\theta - b_i)]\,.} \tag{2}$$

The proficiency parameter is θ. The $P_i(\theta)$ is the probability that a randomly selected examinee with proficiency θ gives a correct answer to item i. It can also be interpreted as the expectation of the item scores of the examinees with proficiency θ. The a_i, b_i, and c_i are the item parameters which characterize item i; The discriminating parameter is a_i. The greater the value of a_i, the more the item is able to discriminate the examinees according to their proficiency levels. The value of a_i is proportional to the slope of the IRF at the point of inflection. The item difficulty parameter is b_i. The greater the value of b_i, the more difficult the item is. The value of b_i is the proficiency level whose $P_i(\theta)$ is the midpoint between lower asymptote and higher asymptote of IRF. The guessing parameter is c_i, which exists to account for examinee guessing on multiple-choice items. For example, if an examinee with very low proficiency responds to a four-choice item randomly, he or she might have a 25% chance of guessing the right choice. The value of c is the lower asymptote of IRF. In the 3-PL model expression, D is a constant and approximately equal to 1.702, which is included in the model to keep the value of a calculated from the logistic model on the same scale of the normal ogive model. When the item parameters are known, θ is the only parameter influencing the probability of examinees' responses.

In the 1-PL model, *a*s are equal for all the items. Often *a* is set to 0.588 so that *D*a* will be 1, but since the scale is arbitrary, this choice is of no significance. The *c* for every item in the 1-PL model is equal to 0. Only *b* characterizes the item. In the 2-PL model, the *c* for every item is equal to 0. The discriminating parameter and difficulty parameter work together to characterize the item. Thus, we can get the 1- and 2-PL model from the 3-PL model by fixing some parameters at certain values and call the 1- or 2-PL model as a special version of the 3-PL model.

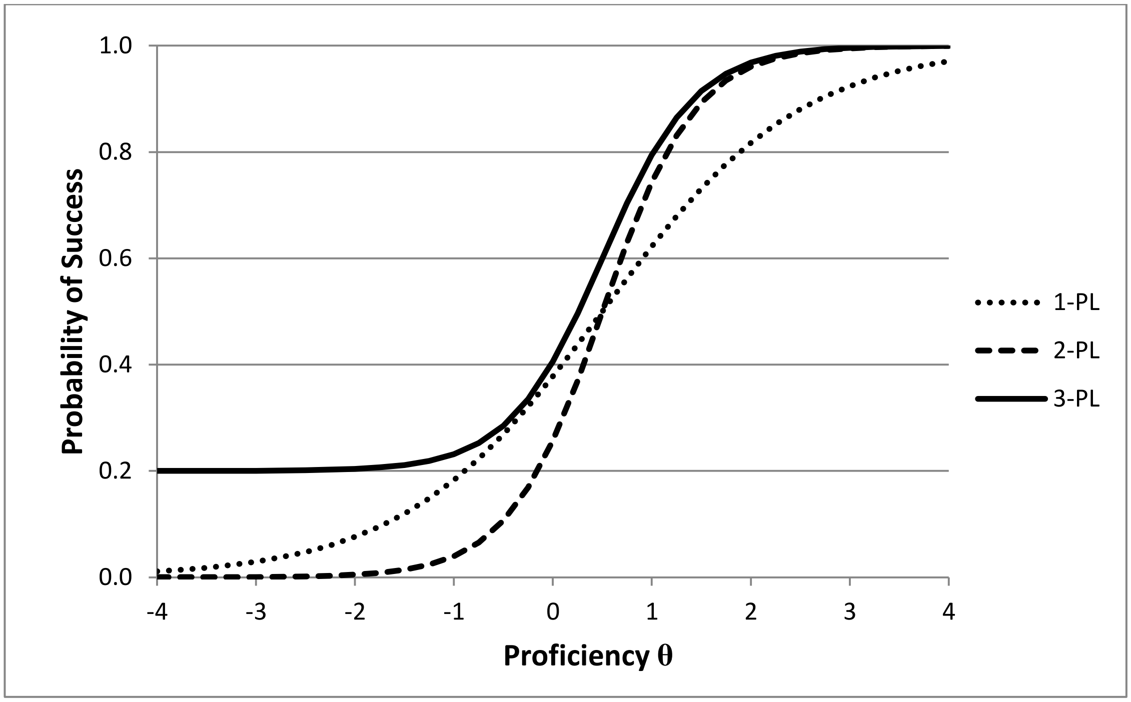

*Figure 2.*IRFs of parametric logistic models. The doc line is the IRFs of 1-PL model; the dash like is the IRF of 2-PL, and 3-PL models are presented.

Figure 2 displayed the IRFs of the 1-, 2-, and 3-PL models. The difficulty parameters for three IRFs are equal to 0.5. IRFs of the 2- and 3-PL models' discrimination parameters are equal to 1.25. Lower asymptotes (i.e., guessing parameters) of the 1- and 2-PL models are equal to 0; the low asymptote of the 3-PL model is equal to 0.2. If the 1- or 2-PL model is used, then the examinees with the proficiency level θ, equal to the item difficulty level 0.5, will be assumed to have a probability of 0.5 to correctly answer this item. If the 3-PL model is used, then the examinees with θ equal to 0.5 will be assumed to have a probability of (1-c)/2+c, to correctly answer this item. The 2- and 3-PL models discriminate examinees in the mid-proficiency range (θ from -2 to 2) better than at extreme proficiency levels.

All three IRFs of parametric logistic models in Figure 2 are monotonically increasing. IRFs, which follow the parametric logistic models, do not decrease as the proficiency level increases. For example, if the proficiency level of examinee A is higher than examinee B, the probability that examinee A will correctly respond to the item will be higher or equal to the probability that examinee B will do so. The mathematical representation of the IRF monotonicity is $for\ \theta_a < \theta_b, P(\theta_a) < P(\theta_b)$.

However, sometimes the IRFs might decrease as the proficiency level increases at a certain proficiency range, like the IRF in Figure 1. In this case, it is not appropriate to use the parametric logistic models because the examinees with lower proficiency could receive the higher scores if parametric logistic models are used. This research was interested in investigating methods that are able to identify the nonmonotonic IRFs.

Nonparametric Smooth Regression

The nonparametric smooth regression method has been used in many areas since it was proposed by Nadaraya (1964) and Watson (1964). Later, Copas (1983) first applied this method to binary data, and Ramsay (1991) was the first to apply this method in the field of psychometrics. Additionally, Azzalini, Bowman, and Härdle (1989) and Douglas and Cohen (2001) used this method to check the fit of the parametric models but did not focus on the monotonicity. Although it has been used in many areas, the basic model of this method is same.

In the nonparametric smooth regression method, the (*X*, *Y*) will be two random variables with a joint bivariate distribution. The *J* observations are drawn from the joint bivariate distribution independently and (x_j, y_j) is the *j*-th observation. The regression function of variable *Y* on *X* is given by the following:

$$E[Y|X=x] = \frac{\sum_{j=1}^{N} K(\frac{x-x_j}{h}) y_j}{\sum_{j=1}^{N} K(\frac{x-x_j}{h})}. \tag{3}$$

If *X* is a discrete variable and the *J* observations contain all possible population values of *X*, *K*((*x*-x_j)/*h*) is equal to 1 when *x* equals x_j and 0 otherwise, but in a real situation *J* observations only contain some of all possible values of *X* and *X* might be continuous variable. Thus, we have to use some methods to average y_js in the small range around every evaluation point *x*. The simplest way is to divide the range of *X* into intervals and then average the value of y_js in each interval. Since this method is very sensitive to the observations and arbitrarily chosen interval boundaries, a better method is to use weights for averaging. When we use weights to average y_js, we set *K*(*u*) to be a nonnegative symmetric kernel function with mode at 0,which is the monotonic decreasing of the absolute value of *u* (Copas, 1983; Douglas, 1997). In the regression function, *h* is a parameter called the bandwidth, which controls the smoothing amount. The choice of *h* is a trade-off between the fluctuation of the regression function and the bias of the regression function estimation. The large bandwidth will increase the effect of x_js far from *x* on the estimation of *E*[*Y*|*X*=*x*], leading to a very smooth *E*[*Y*|*X*=*x*] curve and a high bias but low random error estimation; and the small bandwidth will decrease the effect of x_js far from *x* on the estimation of *E*[*Y*|*X*=*x*], leading to a not very smooth *E*[*Y*|*X*=*x*] curve and a low bias but high random error estimation (Copas, 1983; Douglas, 1997). The procedures such as generalized cross-validation can be used to estimate the value of *h* from sample data (Craven & Wahba, 1978). Moreover, Ramsay (1991) presented the effect of the choice of h on the smoothing and estimation in the graph and suggested an optimal value of h for psychometric binary data.

Applying to educational dichotomous assessment data. Beyond the general understanding of the nonparametric smooth regression method, it is important to understand how to apply it to the educational dichotomous assessment data and why it could be used to check the model fit of parametric logistic models. To apply the nonparametric smooth regression method in this way, Y_i should be a binary random variable denoting the score of one item. When Y_i is equal to 1, it represents that the item has been answered correctly; otherwise Y_i is equal to 0. Also, X will be the latent trait variable Θ, which influences the item score Y under the assumption of unidimensionality. The regression function $E[Y|X=x]$could be written as $P(\Theta=\theta)$, which is the expression of the IRF. However, the latent trait variable Θ could not be observed directly. An estimator $\hat{\Theta}$ will replace Θ to get the estimated$\hat{P}_i(\hat{\Theta} = \theta)$. The IRF estimated by the nonparametric smooth regression method replaces X with $\hat{\Theta}$ in equation 3:

$$\hat{P}_i(\hat{\Theta} = \theta) = \frac{\sum_{j=1}^{N} K(\frac{\theta - \hat{\theta}_j}{h}) y_{ij}}{\sum_{j=1}^{N} K(\frac{\theta - \hat{\theta}_j}{h})}. \tag{4}$$

In this research, we use the Gaussian function as the kernel function in equation 4 because it is a typical choice of the kernel function (Douglas, 1997):

$$K(u) = \exp\left(-\frac{u^2}{2}\right). \tag{5}$$

In this research, we choose h in equation 4 depending on the sample size (N): h=1.1*$N^{0.2,}$ because Ramsay (1991) presented that this value worked well under the Gaussian kernel function. Ramsay's TESTGRAF program (2000) also used this value as the default.

To explore the convergence of this nonparametric smooth regression method when applied to educational dichotomous assessment data, Douglas (1997) has proven that $\hat{P}(\hat{\Theta} = \theta)$ converges to the true IRF with probability 1 as the test length and sample size increase under a set of weak assumptions. The longer the test and the larger the sample size, $\hat{P}_i(\hat{\Theta} = \theta)$ approaches the true IRF. Furthermore, Douglas (1997) also stated that even if the true IRFs do not follow any particular parametric logistic models, the convergence of nonparametric smooth regression method is unaffected. Additionally, Douglas (1999) stated that two corresponding IRFs estimated from different models with the same response distribution will be nearly identical for the long tests. In conclusion, Douglas (1999) assumed that even if two distinct IRFs are estimated, only one is correct. Thus, for a given response distribution and a long test, the IRF estimated by the less restricted nonparametric smooth regression method is the correct one. Moreover, any substantial differences of IRF between the logistic parametric models and the nonparametric smooth regression models will demonstrate that the logistic parametric models do not fit the data (Douglas & Cohen, 2001).

Ordinal ability estimation. In order to plot the IRF estimated from the equation 4, an estimate of each examinee's proficiency is needed. The $\hat{\theta}_j$ (the estimate of the *j*-th examinee's proficiency) can be calculated by the maximum likelihood method using the parametric logistic models. If these models are used to estimate proficiency, we need to assume the parametric logistic models fit the data and estimate the item parameters of the parametric logistic models, which will narrow the application of the nonparametric smooth regression method. Therefore, we use another way to estimate the latent variable, which Douglas (1997) called the ordinal ability estimation.

In the ordinal ability estimation method, the order of the examinees' proficiency is the key information needed for estimating proficiency (Ramsay, 1991). The sum of item scores is often used to order examinees (Douglas and Cohen, 2001). Since the sum of item scores does not consider the variation of item psychometric characteristics, Ramsay (1991) stated that the sum of item scores is not an ideal indicator for ranking examinees. Therefore, Ramsay (1991) presented alternative statistics to rank the examinees. However, Douglas (1997) proved the maximum error in proficiency estimation based on the ranking of the sum of item scores converging to 0 with a probability of 1 as the test length increases under a set of weak assumptions. Consequently, this research uses the order of the sum of item scores to estimate the proficiency. Douglas and Cohen (2001) and Ramsay (1991) also used the ordinal ability estimation method to estimate the proficiency. The steps to apply the ordinal ability estimation are as follows:

1. Determine the empirical percentile of the examinee in the latent trait distribution using the sum of item scores without including the item score being examined $X_{n\text{-}i}$ (n is the number of items and i is the index for the item being examined) because this score is locally independent with the IRF of this item.
2. Use the latent trait distribution G's inverse function G^{-1} to calculate the proficiency based on the empirical percentile.

It may seem that the choice of the latent trait distribution, G, might affect the proficiency estimation and then affect the estimation of IRFs. Actually, it only affects the scale of the proficiency variable and does not change the estimation of IRFs (Ramsay, 1991). For example, τ=$g(\theta)$ is a strictly monotonic transformation of the proficiency variable, θ. Then it becomes:

$$P(\theta) = P\left[g^{-1}\big(g(\theta)\big)\right] = P[g^{-1}(\tau)] = P^{*}(\tau). \tag{6}$$

In the formula, for any θ there is a τ corresponded to it with the same probability of success. Since the strictly monotonic transformation is used, the order of the proficiency is the only thing that affects the probability of success. The change of the proficiency variable scale is like plotting the IRF on a different axis (Ramsay, 1991). In this research, we choose the standard normal distribution as the latent trait distribution.

Item-Ability Regression

The item-ability regression method was first presented by Kingston and Dorans (1985) who used it as a graphical technique to examine the fit of the parametric logistic model. In Kingston and Dorans' (1985) study, the proficiency scale is split into 15 equal intervals ranging from -3.0 to 3.0. In their study, n_i^+ is the number of examinees in the interval *i* who correctly answered the item; n_i^0 is the number of examinees in the interval *i* who omitted the item; and n_i is the number of examinees in the *i*-th interval who answered the item. Once these numbers are calculated, the estimated probability of success in interval *i* was as follows:

$$P(\theta\ in\ ith\ interval) = \frac{n_i^+ + \frac{n_i^0}{A}}{n_i}. \qquad (7).$$

Kingston and Dorans (1985) plotted these probabilities "as squares whose areas are proportional to n_i"(p. 282) and compared these probabilities with the 95% confidence interval of the probability of success of the parametric logistic model at the proficiency θ_i (θ_i is the midpoint of the *i*-th interval). Since the item-ability regression does not impose any restrictions on the shape of IRF, any large difference between the IRF estimated by the item-ability regression method and the IRF estimated by the parametric logistic model concludes that the parametric logistic model does not fit the data. Kingston and Dorans (1985) used the number of times that the estimated probability of success according to the item-ability regression method was not included in the 95% confidence interval as an indicator that the parametric logistic model did not fit the data. However, the 95% confidence interval is estimated by replacing the binomial distribution with the symmetric standard normal distribution and the estimation of 15 probabilities. This estimation is influenced by arbitrary grouping. Most importantly, a large sample size is needed (Kingston &Doran, 1985). These two features constrain the flexibility of the item-ability regression method.

To increase the flexibility of this method, R. McKinley (personal communication, spring, 2010) proposed a modified item-ability regression method. McKinley's modified item-ability regression method calculates prior probability from the empirical proficiency distribution, posterior probability of every examinee on each proficiency evaluation node from the prior probability, and probabilities of success on these evaluation nodes instead of the proficiency intervals from the posterior probabilities and item scores. These probabilities of success form the estimated IRF. The following is the introduction of this modified item-ability regression method in the order of prior probability calculation, posterior probability calculation using prior probabilities, and IRF calculation using posterior probabilities and item scores.

Prior probability and empirical proficiency distribution. The IRF mathematical form and the latent proficiency distribution mathematical form are both unknowns. If one is assumed, the other one can be estimated with or without assuming any mathematical form. The IRF mathematical form is usually assumed as the parametric logistic model and the latent proficiency distribution mathematical form is usually assumed as the standard normal distribution (Woods &Thissen, 2006). For example, the nonparametric smooth regression method does not assume any mathematical form for the IRF, but assumes the standard normal distribution for the latent proficiency distribution. Also, if the IRF is assumed to follow the parametric logistic model, then the latent proficiency distribution can be estimated without assuming any mathematical form. This distribution is called the empirical proficiency distribution.

In order to calculate the empirical proficiency distribution, IRFs are first assumed to follow the parametric logistic model and the expectation-maximization (EM) algorithm is applied to estimate the item parameters (Bock & Aitkin, 1981). One output of the EM algorithm is the empirical proficiency distribution. The probabilities of this distribution are used as priors for calculating the posterior probability of every examinee on each proficiency evaluation node in item-ability regression. If there are *K* proficiency evaluation nodes, the prior probability for each node is $p_{0k}(\theta_k)$.

Posterior probability. In posterior probability, $\boldsymbol{y}_j$ is the score vector of the examine *j* for n items and u_{ij} is the response of the examine *j* on the item *i*. Also, $P_i(\theta_k)$ is the probability of success of item *i* at the proficiency level, θ_k, which is another output of the EM estimation algorithm. The conditional probability of the examinee *j* with the response vector $\boldsymbol{y}_j$ under the condition that he or she is at the proficiency evaluation node *k*, and the assumption of the local independence is as follows:

$$p_{jk}\left(y_j|\theta_k\right) = \prod_{i=1}^{n} P_i(\theta_k)^{u_{ij}}(1 - P_i(\theta_k))^{1-u_{ij}}. \tag{8}$$

Based on Bayes' theorem (Jacod&Protter, 2003, p.17), the posterior probability of the examinee *j* at the proficiency level *k* is the following:

$$p_{1jk}\left(\theta_k|y_j\right) = \frac{p_{jk}\left(y_j|\theta_k\right) * p_{0k}(\theta_k)}{p\left(y_j\right)} . \tag{9}$$

$$p\left(y_j\right) = \sum_{k=1}^{K} p_{jk}\left(y_j|\theta_k\right) * p_{0k}(\theta_k) . \tag{10}$$

The posterior probability p_{1jk} could be interpreted as a weight of examinee *j* at the proficiency evaluation node *k*. The sum of p_{1jk} across all the examinees at each proficiency evaluation node forms the posterior empirical proficiency frequency distribution and could be interpreted as the number of examinees at the proficiency evaluation node *k*.

IRF. During estimating IRF, *N* is the number of examinees, and the estimated probability of success at proficiency evaluation node *k* for item *i* by the item-ability regression method is as follows:

$$P_i(\theta_k) = \frac{\sum_{j=1}^{N} p_{1jk} * u_{ij}}{\sum_{j=1}^{N} p_{1jk}} \,. \tag{11}$$

When an examinee answers item *i* correctly, his or her u_{ij} is equal to 1and the weight counted for the numerator. When an examinee answers the item *i* incorrectly, his or her u_{ij} is equal to 0 and the weight is not counted for the numerator. Thus, the numerator is the estimated number of examinees who answered the *i*-th item correctly at the proficiency evaluation node *k*. The denominator is the number of examinees at the proficiency level *k*. Since the estimated probabilities of success are calculated using probabilities of the empirical proficiency distribution, it is not affected by any arbitrary grouping. Plotting these probabilities will present the shape of the estimated IRF in the graph. If the nonmonotonic part of the IRF is obvious, it will be noticeable from the shape of plot and indicate this item has a nonmonotonic IRF.

B-Spline Nonparametric IRT

The B-spline nonparametric IRT method uses the functional data analysis method to estimate the IRF nonparametrically. Rossi et al. (2002) first presented this method as a more flexible alternative method compared to the 3-PL model. The following section includes the nonparametric model and the estimation procedures of this method.

Nonparametric Model. The B-spline nonparametric IRT method uses the functional data analysis method to model the probability of success in a linear regression function. It then transforms this linear function into a logistic form to make the probability in the range from 0 to 1. In this section, the model of this method and the functional data analysis method along with its basis function will be introduced.

Model. The model of the B-spline nonparametric IRT method, which estimates the IRF, is as follows:

$$\lambda = ln\frac{P}{1-P} \;\; or\; P = \frac{\exp\lambda}{1+\exp\lambda}\,. \tag{12}$$

In this equation, λ is a continuous linear function of theta and needs to be estimated. Also, P is the logistic transformation of λ and P is continuous on the scale from 0 and 1(Rossi et al., 2002). Since the logistic transformation is a monotonic transformation, P will increase as λ increases. However, λ might not increase as theta increases. Thus, the nonmonotonic IRF can be estimated by this model.

The B-spline nonparametric IRT method also uses the logistic form to model the IRF, which is the same as the parametric logistic models. A difference between the parametric logistic models and this model is that the degree of the linear function is 1 for the parametric logistic model, but the degree of the linear function, λ, could be bigger than 1 for the B-spline nonparametric IRT model. Another difference is that the form of linear function is fixed with one to three parameters for the parametric logistic model, but the form of λ is not constrained and nonparametric for the B-spline nonparametric IRT model. If certain constraints are added to λ, the parametric logistic models could be considered a special version of the B-spline nonparametric IRT's model. Since the B-spline nonparametric IRT method uses a very general model and imposes fewer restrictions on the shape of the IRF (Sijtsma & Junker, 2006), it can be concluded that the parametric logistic model is misfit if it differs substantially from the IRF estimated by the B-spline nonparametric IRT method.

Functional data analysis. The linear function, λ, is estimated through the functional data analysis method, which assumes that the function being estimated is smooth (Ramsay & Silverman, 1997, p.1). When the function data analysis method is applied to λ estimation, the λ is represented as a linear combination of *K* basis functions, which are preselected known functions:

$$\lambda_i(\theta) = \sum_{k=1}^{K} \beta_{ik} B(\theta)_{k\,.} \tag{13}$$

In this equation, $\beta_{ik}$is the coefficient of the basis function $B(\theta)_k$. The basis functions are the same across different items. The coefficients are different for different items, which are the item parameters that need to be estimated.

There are two types of basis functions: B-spline basis functions for the nonperiodic functions and Fourier series functions for the periodic functions (Rossi et al., 2002). The B-spline basis functions are used for the estimation in this research.

B-spline basis functions. The B-spline basis functions are a set of piecewise polynomials with the same degree. These polynomials have three properties: nonzero on a few intervals, evaluated at any values, and positive on nonzero intervals (de Boor, 2001, p.87; Nürnberger, 1989). The mathematical expression of these basis functions are below.

The degree of basis functions is *d* and *U* is a set of *m*+1 nondecreasing knots on θ, u_0<=u_1<=…<=u_x<=… <=u_m. The interval [u_x, u_{x+1}) is the *x*-th knot span. If a knot appears several times (i.e.,u_x=u_{x+1}…) and their knot spans are equal to 0, this knot is called a multiple knot. If a knot appears only once, it is called a simple knot. The knot spans could be equal. These equally spaced knots are called uniform knots. Otherwise, the unequally spaced knots are called cardinal knots. The range from *u*$_0$ to *u*$_m$ is the evaluated range and all the B-spline basis functions are zero out of evaluated range (de Boor, 2001, p.88).The *x*-th B-spline basis function of degree *d* is as follows:

$$B_x^d(\theta) = \frac{\theta - u_x}{u_{x+d} - u_x} B_x^{d-1}(\theta) + \frac{u_{x+d+1} - \theta}{u_{x+d+1} - u_{x+1}} B_{x+1}^{d-1}(\theta) \qquad (14)$$

$$and\ B_x^0(\theta) = \begin{cases} 1 & u_x \le \theta \le u_{x+1} \\ 0 & otherwise \end{cases}.$$

This B-spline basis function formula is also called the Cox-de Boor recursion formula (de Boor, 2001, p.89). This formula shows the recurrence relation of the basis functions because the basis functions with different degrees are defined over the same knot span, and higher degree function is a function of lower degree ones.

There is a relationship among the number and degree of the B-spline basis functions, and the number of knots. When *K* is the number of the B-spline basis functions, *d* is the degree of the B-spline basis functions and *m*+1 is the number of knots, then (de Boor, 2001, p. 89) the relationship is the following:

$$K = d + 1 + (m + 1) - 2. \qquad (15)$$

In order to complete the calculation of *K* B-spline basis functions, another 2**d* multiple endpoint knots (*d* multiple knots for each end) are needed. De Boor (2001, p.89) referred to this situation as the “not-a-knot” condition.

Since the B-spline basis functions are piecewise polynomials and zero everywhere except a few intervals, it is a very flexible and a computationally convenient tool for the estimation of λ. This research uses nine uniformly spaced simple knots ranging from -2.5 to 2.5. The degree of the basis function is three and there are 11 B-spline basis functions. Also, another six multiple endpoint knots are needed for computations. Figure 3 plots the five B-spline basis functions of degree three defined by the uniform spaced knots -2, 0, and 2,as well as six multiple endpoint knots.

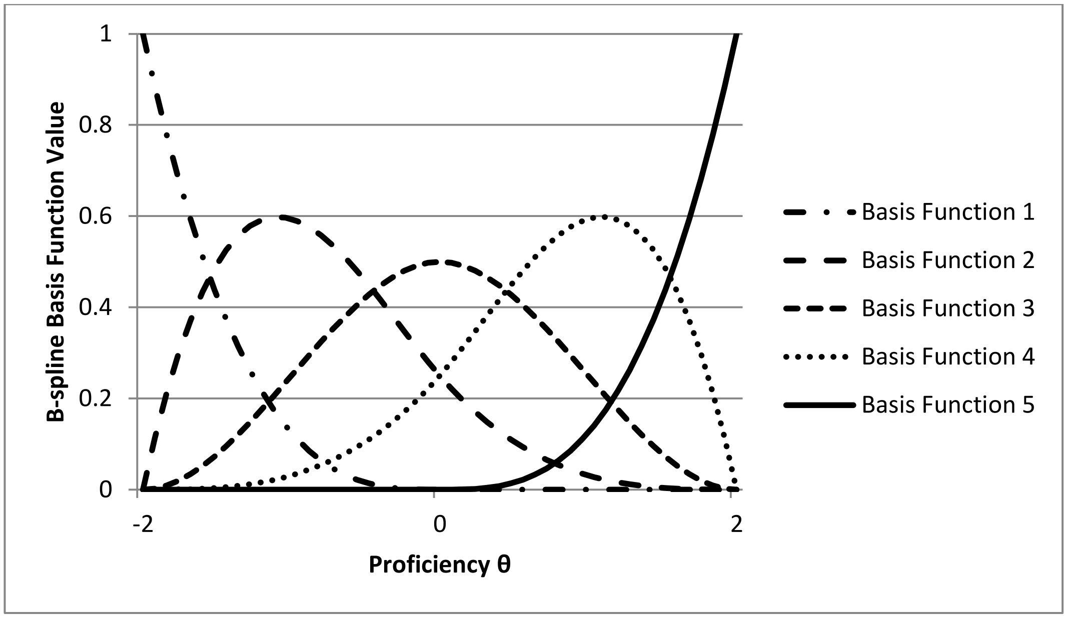

*Figure 3.*B-spline basis function. The five B-spline basis functions of degree three is defined by the uniform spaced knots -2, 0 and 2.

Estimation. In order to estimate the IRF using the B-spline nonparametric IRT method, we only need to estimate the coefficients of each B-spline basis function. The EM algorithm is used for the estimation of these coefficients. The EM algorithm (Bock & Aitkin, 1981) has three parts: initialization, expectation phase (E-phase), and maximization phase (M-phase). Initialization determines the initial values of the estimated parameters. The E-phase fixes the parameters and calculates the expected values of sample distributions frequency. The M-phase maximizes the marginal likelihood with respect to the coefficients. After initialization, the E-phase and M-phase alternate and the converging criterion is evaluated at the end of the E-phase. If the converging is reached, the estimation stops. In addition, the relationship between the smooth and flexibility of B-spline nonparametric IRT method should be considered during estimation. Since increasing the number of basis functions improves the flexibility but leads to a less smooth IRF, there should be a balance between the smooth and the flexibility. The roughness penalty method could be used to find the balance (Ramsay & Silverman, 1997). The roughness penalty method adds a roughness measure, which penalizes the curvature of the IRF, to the converging criterion for judging the convergence. The following information is an introduction on initialization, E-phase, converging criterion, and the M-phase.

Initialization. In initialization, the vales of the coefficients of every item and the prior probability at each evaluation points should be initialized for the calculation in the E-phase. Rossi et al. (2002) suggested using the nonparametric smooth regression method to estimate initial coefficients and the standard normal distribution to calculate initial prior probabilities.

To apply the nonparametric smooth regression method for calculating initial coefficients, the researcher set N_{qp} to be the number of theta evaluation points (θ_1, θ_2, . . .θ_q, . . .θ_{Nqp}) for estimating the initial values and evaluating the expected sample frequency in E-phase. First, the probabilities of success at each proficiency evaluation point of every item are estimated by the nonparametric smooth regression method. Then N_{qp} $\lambda_i^0 s$ for the *i*-th item at each proficiency evaluation point can be calculated from equation 12 using these probabilities. Since $\lambda^0 s$ are linear combinations of *K* B-spline basis functions and coefficients, it could be interpreted as the linear regression model except that the "independent variables" are functions. The values of $\lambda^0 s$ are the dependent variables. The *K* B-spline basis functions are independent variables. The *K* basis functions evaluated at each theta evaluation point forms a K*N_{pq} matrix, which are the data for estimating regression coefficients. Then for every item, initial coefficients, $\beta_{ik}^0 s$, could be estimated in the same ways as for linear regression, using the least square or maximum likelihood estimation method. These initial coefficients values, $\beta_{ik}^0 s$, are used in E-phases in the first cycle.

The prior probability of one examinee at the proficiency evaluation point θ_q, which could be interpreted as the weight of θ_q in the sample empirical latent proficiency distribution, is equal to the following:

$$w_{0q}^{0} = \frac{N^{0}(\theta_q)}{N}. \tag{16}$$

The estimated number of examinees at the evaluated node θ_q is $N^{0}(\theta_q)$. The total number of examinees in the sample is N. To estimate the initial prior probabilities, we could assume that the sample empirical latent proficiency distribution is a standard normal distribution (Rossi et al., 2002).

It should be noted that the initial estimated number of examinees at the proficiency evaluation point θ_q who give the correct answer to item i is as follows:

$$R_i^{0}(\theta_q) = P_i^{0}(\theta_q) * N^{0}(\theta_q). \tag{17}$$

The sum of $N^{0}(\theta_q)$ across ability evaluation points is equal to the sample size.

E-phase. The E-phase aims to find two expectations using the estimated coefficients (e.g., $\beta_{ik}^{0}s$ for the i-th item) and the prior probability of one examinee at a certain proficiency evaluation point (e.g., $w_{0q}^{0}$of the proficiency evaluation point θ_q) either from the initialization step for the first cycle or from the M-phase for all the other cycles. The two expectations are the expectation of the number of examinees at each proficiency evaluation points who answers each item (e.g., $N_i(\theta_q)$ for the i-th item), and the expectation of the number of examinees at each proficiency evaluation points who answers each item correctly (e.g.,$R_i(\theta_q)$ for the i-th item).These expectations are the expectations of the indicator random variables. Using item i as an example, one of the indicator random variables is G_i and g_{ij} is the observation of this indicator random variable for examinee j. If examinee j answered the item i, g_{ij} is equal to 1. Otherwise, g_{ij}

is equal to 0. Another indicator random variable is I_i and i_{ij} is the observation of this indicator random variable for examinee *j*. If examinee *j* answered the item *i* correctly, i_{ij} is equal to 1. Otherwise, i_{ij} is equal to 0.

For each item, the expectations of these two indicator random variables are calculated using the posterior probabilities of each examinee at proficiency evaluation points. These posterior probabilities for each examinee are calculated from the estimated prior probabilities and coefficients. In order to calculate the posterior probabilities, $\boldsymbol{y}_j$ is set to be the score vector of the examine *j* for *n* items and u_{ij} is set to be the response of the examine *j* on the item *i*. The probability of success of item *i* of an examinee at the proficiency evaluation point θ_q, $P_i(\theta_q)$, is calculated from equations12 and 13 using the estimated coefficients. When the assumption of the local independence is held, then the conditional probability of examinee *j* with the response vector $\boldsymbol{y}_{j,}$ under the condition that he or she is at the proficiency evaluation point $\theta_{q,}$is as follows:

$$L_{jq}\left(y_j \middle| \theta_q\right) = \prod_{i=1}^{n} P_i\left(\theta_q\right)^{u_{ij}} (1 - P_i\left(\theta_q\right))^{1-u_{ij}}. \tag{18}$$

Based on Bayes' theorem (Jacod & Protter, 2003, p.17), the posterior probability of the examinee *j* at the proficiency evaluation point θ_q is as follows:

$$w_{1jq} = \frac{L_{jq}\left(y_j \middle| \theta_q\right) * w_{0q}}{\sum_{q=1}^{N_{qp}} L_{jq}\left(y_j \middle| \theta_q\right) * w_{0q}} \, . \tag{19}$$

In this equation, w_{0q} is the estimated prior probability updated every cycle. The posterior probability w_{1jq} could be interpreted as the weight of examinee *j* at the proficiency evaluation point θ_q. Then for item *i*, two expectations are as follows:

$$N_i\left(\theta_q\right) = \sum_{j=1}^{N} w_{1jq} g_{ij} \, . \tag{20}$$

$$R_i(\theta_q) = \sum_{j=1}^{N} w_{1jq} i_{ij} \ . \tag{21}$$

It should be noted that the sum of indicator function G_i over all the examinees is equal to the sum of the number of examinees at each proficiency evaluation point:

$$\sum_{j=1}^{N} g_{ij} = \sum_{q=1}^{N_{qp}} N_i(\theta_q). \tag{22}$$

Also, using the expectation of the number of examinees at each proficiency evaluation point, the prior probability can be updated. The updated prior probability, w'_{0qi}, of an examinee at the proficiency evaluation point θ_q for item i is as follows:

$$w'_{0qi} = \frac{N_i(\theta_q)}{\sum_{j=1}^{N} g_{ij}} \ . \tag{23}$$

If every examinee answers all the items on the test, the estimated number of examinees at each proficiency evaluation point and the updated prior probability are the same across items. Then, the calculation formula of the updated prior probabilities is as follows:

$$w'_{0q} = \frac{N(\theta_q)}{N} \ . \tag{24}$$

Convergence. The log likelihood can be used for judging convergence (Bock & Aitkin, 1981). The log likelihood is calculated using the updated item coefficients and the updated prior probabilities. The log likelihood of this estimation algorithm is as follows:

$$L = \sum_{j=1}^{N} \ln[\sum_{q=1}^{N_{qp}} w_q L_{jq}(y_j|\theta_q)] \ . \tag{25}$$

The log likelihood first sums over the N_{qp} proficiency evaluation points and then sums over the N examinees. The roughness measure added to the log likelihood is given by the linear function's squared second derivative after integrating over θ:

$$J(\beta_i) = \int_{-\infty}^{\infty} (\frac{d^2\lambda_i(\theta)}{d\theta^2})^2 d\theta. \tag{26}$$

The larger the number of wiggle in the IRF is, the larger the roughness measure will be. Then the converging criterion is:

$$F = -\sum_{j=1}^{N} \ln[\sum_{q=1}^{N_{qp}} w_q L_{jq}\left(y_j \middle| \theta_q\right)] + \gamma \sum_{i=1}^{n} \int_{-\infty}^{\infty} (\frac{d^2\lambda_i(\theta)}{d\theta^2})^2 d\theta. \tag{27}$$

In this equation, γ is the smoothing parameter. As γ increases, the roughness measure is more significant compared to the log likelihood. Thus, a small roughness measure and smooth $\lambda$are needed. Finally, if γ increases to ∞, the linear function, λ, will be forced to be a straight line and the estimated IRF will follow a 2PL model. As γ decreases to 0, the penalty of the curvature vanishes and λ can be complex and rough. If the converging criterion in equation 27 is "no longer judged to be improving from one iteration to the next and-or item parameters cease to change substantially", the convergence has been reached (Rossi et al., 2002, p. 299).

M-phase. The M-phase aims to maximize the log likelihood L with respect to the coefficients β_{ik}s (Rossi et al., 2002). Coefficients that maximize the log likelihood L could be solved by letting the first derivative of L be equal to 0. The following equation is an example that is used to solve the coefficient of the k-th basis function for item i, β_{ik}, which maximizes L by letting the first derivative of L be equal to 0.

$$\frac{\partial L}{\partial \beta_{ik}} = \sum_{q=1}^{N_{qp}} [R_i(\theta_q) - P_i(\theta_q) * N_i(\theta_q)] \frac{\partial \lambda_i(\theta_q)}{\partial \beta_{ik}} = 0. \tag{28}$$

The $R_i(\theta_q)$ and $N_i(\theta_q)$ are treated as known using the calculation from E-phase, whose values are updated every cycle. $R_i(\theta_q)$ and $N_i(\theta_q)$ are dependent on the probabilities of success $P_i(\theta_q)$. Once the coefficients β_{ik}s are calculated from M-phase, probabilities of success, $P_i(\theta_q)$, will be updated, and then $R_i(\theta_q)$ and $N_i(\theta_q)$ can be updated in E-phase.

Rossi et al. (2002) gave a reexpression of the log likelihood and introduced another way to estimate $\lambda_i(\theta_q)$ in M-phase if $\lambda_i(\theta_q)$ is the parameters that researchers were seeking. This reexpression log likelihood L'could be interpretedas the sum of the log likelihoods over the N_{qp} proficiency evaluation points. For one proficiency evaluation point θ_q, the log likelihoods of getting $R_i(\theta_q)$ successful events in $N_i(\theta_q)$ trials with the probability of success $P_i(\theta_q)$ is as follows:

$$L_q = R_i(\theta_q) * lnP_i(\theta_q) - (N_i(\theta_q) - R_i(\theta_q)) * \ln\Big(1 - P_i(\theta_q)\Big). \tag{29}$$

The sum log likelihood is also as follows:

$$L' = \sum_{q=1}^{N_{qp}} L_q = \sum_{q=1}^{N_{qp}} [R_i(\theta_q) * lnP_i(\theta_q) - \Big(N_i(\theta_q) - R_i(\theta_q) * \ln\Big(1 - P_i(\theta_q)\Big)\Big]. \tag{30}$$

The first derivative of this expression with respect to β_{ik} is as same as equation25.

If $\lambda_i(\theta_q)$ is the parameter that researchers are seeking, $\lambda_i(\theta_q)$ that maximizes the log likelihood L'could be solved by letting the first derivative be equal to 0:

$$\frac{\partial L'}{\partial \lambda_i(\theta_q)} = R_i(\theta_q) - P_i(\theta_q) * N_i(\theta_q) = 0. \tag{31}$$

Then,

$$P_i(\theta_q) = \frac{R_i(\theta_q)}{N_i(\theta_q)}\ and\ \lambda_i(\theta_q) = ln\frac{P_i(\theta_q)}{1 - P_i(\theta_q)}\,. \tag{32}$$

Since $\lambda_i(\theta_q)$s are known, they could be treated as the observations of the dependent variable of a linear regression and the coefficients β_{ik}s could be calculated as the regression coefficients.

Posterior Predictive Model Checking

The posterior predictive model checking (PPMC) method (Gelman et al., 1996; Guttman, 1967; Rubin, 1984) is a statistical inference method from a Bayesian perspective. The PPMC method is mainly used for checking model fit (Gelman et al., 1996). Guttman (1967) introduced this method, and Rubin (1984) applied it to several examples and gave a formal definition. Furthermore, Gelman et al. (1996) gave a clear description of this method and applied it to several statistical problems. The PPMC method has also been applied in psychometrics to check model fit by several researchers (Glas & Meijer, 2003; Levy et al., 2009; Sinharay, 2005; Sinharay, Johnson, & Stern, 2006). In this study, PPMC was used to assess the fit of the parametric logistic model by focusing on the monotonicity of the parametric logistic model's IRF. The following sections describe this method, introduce the advantages of this method, and apply this method to assessing the fit of the parametric logistic model.

Description of the method. ***Posterior distribution of unknown parameters.*** The terms for calculating the posterior distribution include: $\boldsymbol{\omega}$, the unknown parameters of the assumed model *H*, and $\boldsymbol{y}$, the observed data. The prior distribution of the unknown parameters is $p(\boldsymbol{\omega})$ and the likelihood distribution of observed data assuming that the model *H* is true is $p(\boldsymbol{y}|\boldsymbol{\omega})$ (Sinharay, 2005; Sinharay et al., 2006). Using Bayes' theorem (Jacod & Protter, 2003, p.17), the posterior distribution of unknown parameters is $p(\boldsymbol{\omega}|\boldsymbol{y})$ and $p(\boldsymbol{\omega}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\omega})p(\boldsymbol{\omega})$.

Replicated data's posterior predictive distribution. Another type of data, $\boldsymbol{y}^{rep}$, is called the replicated data. The replicated data could be interpreted as the data that will be observed in the future or predicted and are then replicated using the same model *H* and parameters drawn from the $p(\boldsymbol{\omega}|\ \mathbf{y})$. The posterior predictive distribution of $\boldsymbol{y}^{rep}$ is as follows:

$$p(\boldsymbol{y}^{rep}|\boldsymbol{y}) = \int p(\boldsymbol{y}^{rep}|\boldsymbol{\omega})p(\boldsymbol{\omega}|\boldsymbol{y})d\boldsymbol{\omega}. \tag{33}$$

This distribution was calculated from a Bayesian perspective to eliminate the nuisance parameters by integrating them out (Bayarri & Berger, 2000). A statistic that was calculated from the observed data is compared to the distribution of the same statistic calculated from the replicated data drawn from this distribution. This was done to check the model fit (Gelman et al., 1996).

Posterior predictive p-value (PPP-value). PPP-value provides a quantitative measure of the degree to which the model is able to capture the features of the observed data, in other words, the fit of the model to the observed data. The PPP-value is defined as (Bayarri & Berger, 2000):

$$p = p(T(\boldsymbol{y}^{rep}) \geq T(\boldsymbol{y})|\boldsymbol{y}) = \int_{T(\boldsymbol{y}^{rep}) \geq T(\boldsymbol{y})} p(\boldsymbol{y}^{rep}|\boldsymbol{y})\, d\boldsymbol{y}^{rep}. \qquad (34)$$

In this equation, $T(\boldsymbol{y})$ is a discrepancy measure which is chosen to address the aspect of the model that the researcher is interested in to make inference about the fit and to measure data characteristics which cannot be addressed by the probability model (Gelman, Carlin, Stern, & Rubin, 2003, p.172). Extreme PPP-values, those close to 0, 1, or both (depending on the nature of the discrepancy measure), indicate the model does not fit the data (Sinharay et al., 2006).

Rubin (1984) suggested a practical way to form the posterior predictive distribution and calculate PPP-value, which aims to deal with the difficulty of the analytical calculations of equations 33 and 34.The steps of this calculation are as follows:

1. A total of M simulations $\boldsymbol{\omega}^1$, $\boldsymbol{\omega}^2$, …, and $\boldsymbol{\omega}^M$ is drawn from the posterior distribution, $p(\boldsymbol{\omega}|\ \mathbf{y})$, via the Markov Chain Monte Carlo (MCMC) algorithm.
2. A total of M data sets $\boldsymbol{y}^{rep,1}$, $\boldsymbol{y}^{rep,2}$, …, and $\boldsymbol{y}^{rep,M}$ is drawn from the likelihood distribution $p(\boldsymbol{y}|\boldsymbol{\omega}^m)$.
3. The discrepancy measure is calculated from each replicated data set, $T(\boldsymbol{y}^{rep,m})$.

4. PPP-value is equal to the proportion that $T(\boldsymbol{y}^{rep,m})$ is larger than $T(\boldsymbol{y})$ when $T(\boldsymbol{y})$ is compared to $MT(\boldsymbol{y}^{rep,m})$s.

Advantages of the method. The PPMC has three advantages compared to the classic model-checking method (Glas & Meijer, 2003; Rubin, 1984; Sinharay, 2005):

1. The PPMC method works when the theoretical sampling distribution of the statistic is unknown. The sampling distribution is not necessary for the PPMC method.
2. The uncertainty of the parameter estimation is taken into account by integrating over the parameters. Rubin (1984) suggested it is “scientifically valuable and appropriate to expose this sensitivity” (p.1157).
3. The procedures of the PPMC method are easily generalized (Glas& Meijer, 2003).

Application to checking monotonicity of the IRF. The model *H* in this study was 3-PL model. The parameters $\boldsymbol{\omega}$ were item parameters, *a*s, *b*s, *c*s, and proficiency parameters, θs (e.g., a_i, $b_{i,}$, and c_i for item *i* as well as θ_j for examinee *j*). The observed data $\boldsymbol{y}$ were the responses of examinees for all the items. The discrepancy measure $T(\boldsymbol{y})$ for the *i*-th item was as follows:

$$T_i = \int_{\theta_{\mathrm{lk}}}^{\theta_{\mathrm{uk}}} (P(\theta_{lk}) - P(\theta))d\theta \tag{35}$$

$$for\ any\ \theta_a\ and\ \theta_b, \qquad \theta_{lk} < \theta_a \leq \theta_b < \theta_{mk}\,, P(\theta_b) \leq P(\theta_a) \leq P(\theta_{lk}),$$

$$for\ any\ \theta_a\ and\ \theta_b, \qquad \theta_{mk} < \theta_a \leq \theta_b < \theta_{uk-1}\,, P(\theta_a) \leq P(\theta_b) \leq P(\theta_{uk-1}),$$

$$and\ \theta_{lk}\ \leq \theta_{mk}\ \leq \cdots \leq \theta_{uk-1}\ \leq \theta_{uk}, P(\theta_{mk}) \leq P(\theta_{uk-1}) \leq P(\theta_{lk}) \leq P(\theta_{uk});$$

$$\theta_{lk}, \theta_{mk}, \theta_{uk-1}\ and\ \theta_{uk}\ are\ node.$$

For item i, T_i was the area of nonmonotonic IRF of this item. It is the area between the straight line and the IRF in the oval in Figure 4.

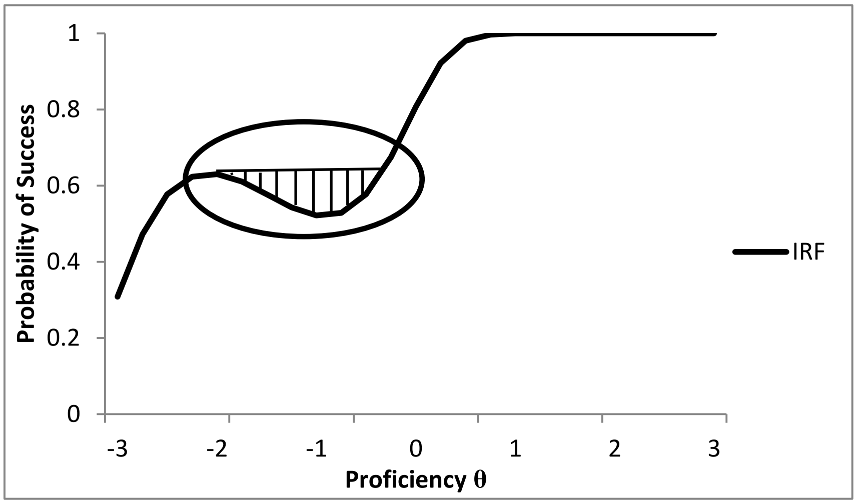

*Figure 4.*Area of nonmonotonic IRF. The area between the straight line and the IRF in the oval is the area of nonmonotonic IRF.

For the parametric logistic models, T_i of was 0 for every item. Since IRFs that were estimated from the nonparametric smooth regression method and the item-ability regression methods were not continuous, the numeric integration was used to calculate this area.

The following were procedures for applying the PPMC method to check the monotonicity of IRF (Sinharay et al., 2006):

1. The MCMC algorithm was used to simulate the posterior distributions of item parameters and proficiency parameters using the observed data and the 3-PL model (Patz & Junker, 1999).
2. A total of N θs was drawn from corresponding posterior distributions. The sample size was N. For example, θ_j of the j-th examinee was drawn from the posterior distribution $p(\theta_j|\mathbf{y})$.
3. The item parameters of n items were drawn from their posterior distributions.

4. A data set was generated from the 3-PL model using item parameters and proficiency parameters.
5. The discrepancy measure was calculated for this data set and compared with the same discrepancy measure calculated from the observed data.
6. Steps 2 to 5 were repeated for *M* times to compute the PPP-value for every item.

Summary

The IRF could be estimated using both the parametric methods and the nonparametric methods. Since the nonparametric methods have few restrictions on the shape of IRFs, some of these methods allow nonmonotonic IRFs and can be used as tools to check the monotonicity of IRFs. This research investigated three approaches (a) the nonparametric smooth regression method, (b) the item-ability regression method, and (c) the B-spline nonparametric IRF method. The nonparametric smooth regression method used the order of examinees' sum scores to estimate their proficiency levels and then estimated the IRF using examinees' responses and proficiency estimates. The item-ability regression method formed the IRF by using the ratio of examinees correctly answering the item to examinees answering the item at every proficiency evaluation node. The B-spline nonparametric IRF method estimated the coefficients of the B-spline basis functions to form a linear function of proficiency and transformed this linear function to a logistic form, which was the model of IRF.

Once an IRF has been estimated by each of these three nonparametric methods in a real data study, the PPMC method was used to judge the degree of the parametric logistic model fit focusing on the nonmonotonicity of IRFs. The PPMC method used the replicated data, which were generated by the model, and the unknown parameters drawn from the posterior distribution to calculate the PPP-value of certain discrepancy measure. Extreme PPP-values indicate that the model does not fit the data. Overall, this study was interested in and explored the area of the nonmonotonic IRF for the PPMC method.

Chapter 3: Method

This study included an investigation and comparison of three nonparametric IRF estimation methods—the nonparametric smooth regression method, the item-ability regression method, and the B-spline nonparametric IRF method—to determine whether they could detect the nonmonotonic IRF accurately using simulated data. After a comparison of the three methods, identification of items with nonmonotonic IRFs on real assessments using these three methods occurred. This chapter includes descriptions of assessments, participants, software, design, simulation study and real data study.

Assessments

For the real data study, standard end-of-year summative assessments administered to high school students across the state were used. Five assessments (three mathematics assessments from 2008-2010 and two reading assessments from 2008-2009) were examined. These assessments aim to test all the content standards of the high school math and reading curricula and only include multiple-choice items that have one correct answer out of the four answer choices.

Participants

For each assessment, 10,000 high school examinees (grades 9 to 11) who took the state, end-of-year summative assessment in mathematics or reading were randomly selected. Selected participants answered all the items on the assessment.

Software

The software used in this research includes BILOG-MG, WinBUGS, Fortran programs, and SAS. To implement the EM algorithm for the item-ability regression estimation, BILOG-MG was used. BILOG-MG applied the EM estimation algorithm to calculate the prior probability of each proficiency evaluation node for the item-ability regression estimation. WinBUGS was used to implement the MCMC algorithm on the real data set. WinBUGS drew item parameters and proficiency parameters from the posterior distributions through the MCMC algorithm for the PPMC method. Five Fortran programs were written to simulate the data, apply the nonparametric smooth regression method, apply the item-ability regression method, apply the B-splines nonparametric IRT method, and calculate the area of nonmonotonic IRF. Finally, SAS was used to read the output of the Fortran programs and calculate the PPP-values.

Design

In the simulation study, an experimental design was used. The independent variable was the nonparametric method which was used to estimate the IRFs. The dependent variable was the accuracy with which each method can detect the items with nonmonotonic IRFs. The real data study was an exploratory study. Real data study aimed to identify the prevalence of nonmonotonic IRFs.

Simulation Study

The simulation study had three phases: data generation, IRF estimation along with nonmonotonic area calculation, and results analysis. The simulated data were generated from two models: monotonic IRF and nonmonotonic IRF. Three nonparametric methods were used to estimate IRFs of 60 items. The result analyses were consisting of the type I and type II error rate calculation and the nonmonotonic area correlation among three nonparametric methods.

Data generation. In this section, the models and their parameters to generate simulated data are introduced. The simulation method and the number of simulated data sets generated are also described. These simulated data sets were used to compare three nonparametric methods.

Models. Because the 1-PL or 2-PL model is a special version of the 3-PL model, the 3-PL model (equation 2) was used to generate responses of items with monotonic IRFs. On the other hand, the model for generating responses of items with nonmonotonic IRFs is as follows:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[0.5 * a_i * (b_i - (\theta + 1.5)^3 + (\theta + 2))]}. \qquad (36)$$

These two models all have three item parameters. The difference is that the degree of the polynomial in exponential function is 1 for the monotonic model and 3 for the nonmonotonic model.

Proficiency and item parameters. Ten thousand proficiency parameters, θs, were randomly drawn from a standard normal distribution. Sixty groups of item parameters were randomly selected from the real item parameter estimates of the high school summative mathematics assessment. Six groups of item parameters were for the nonmonotonic model and 54 groups of item parameters were for the monotonic model. These six groups of item parameters (item 10, 20, 30, 40, 50, and 60) were selected purposely. Items10 and 20 either have low or high item difficulty parameters. Items 30 and 40 either have low or high item discriminate parameters. Items 50 and 60 either have low or high item guessing parameters. Table 1 indicated the true item parameters for these six items. Figure 5-10 presented the true IRF for these six items.

Table 1
True Item Parameters for Nonmonotonic IRFs

Item	a	b	c
10	0.855	1.654	0.139
20	0.695	-2.416	0.336
30	0.369	0.863	0.163
40	1.8	0.123	0.295
50	0.833	-1.509	0.089
60	0.784	-0.841	0.522

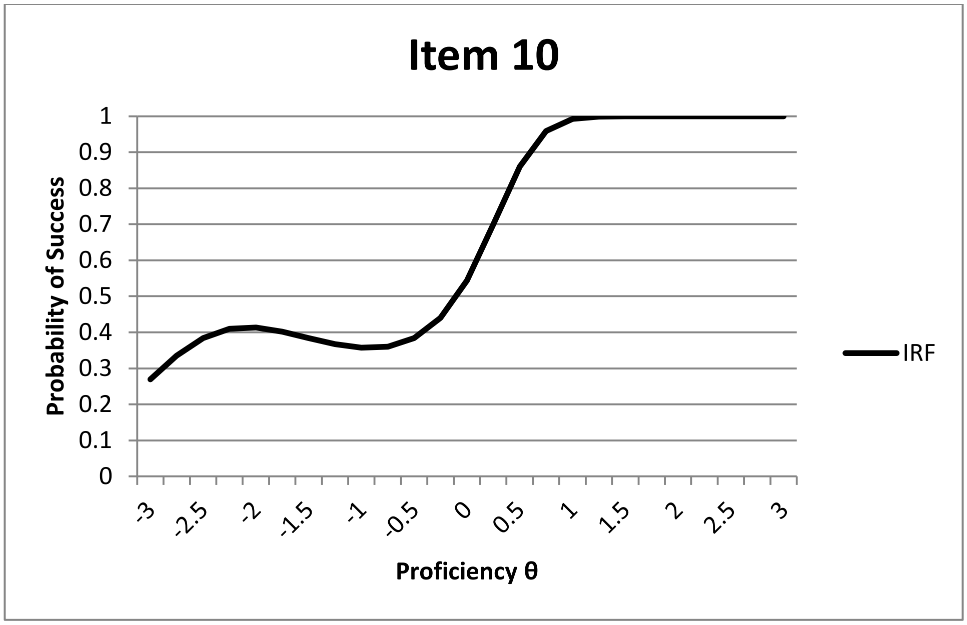

Figure 5. True IRF of simulated item 10. This item has high item difficulty parameter.

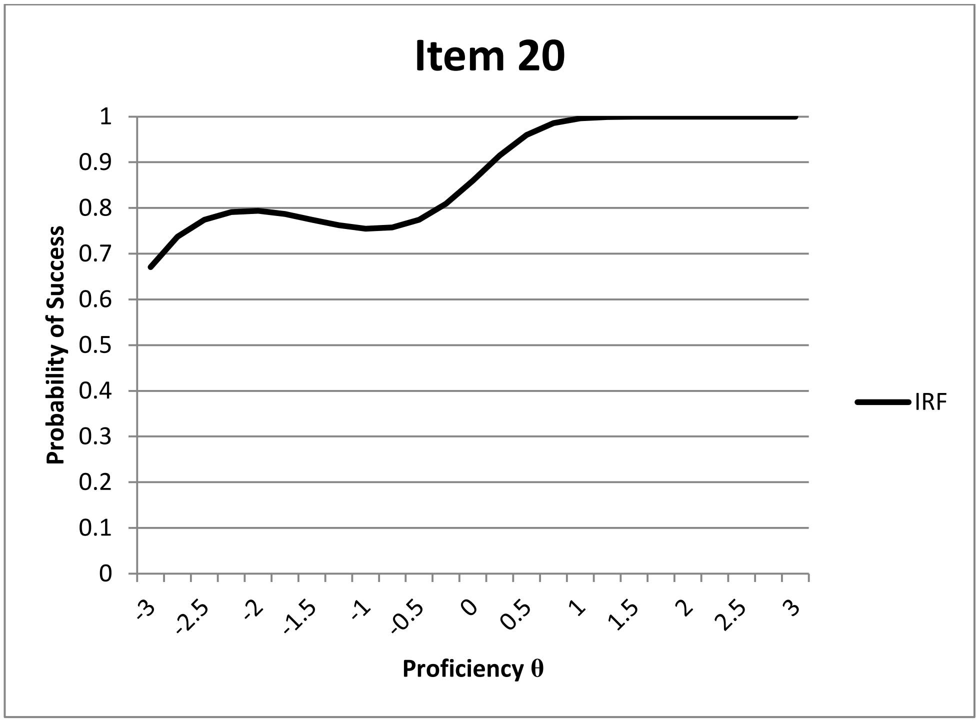

Figure 6. True IRF of simulated item 20. This item has low item difficulty parameter.

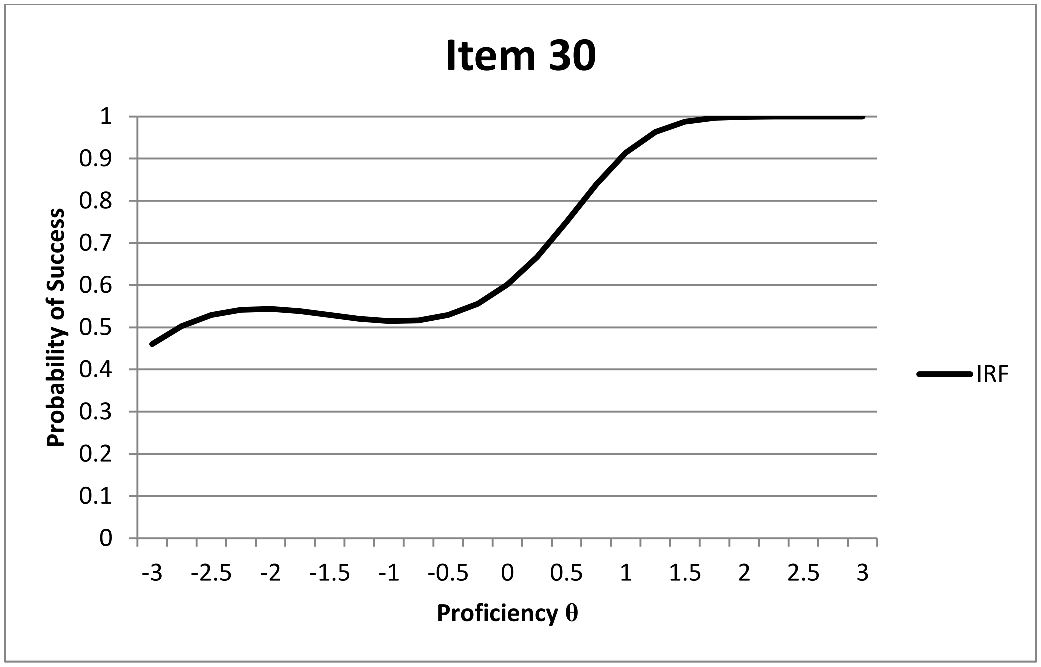

Figure 7. True IRF of simulated item 30. This item has low item discriminate parameter.

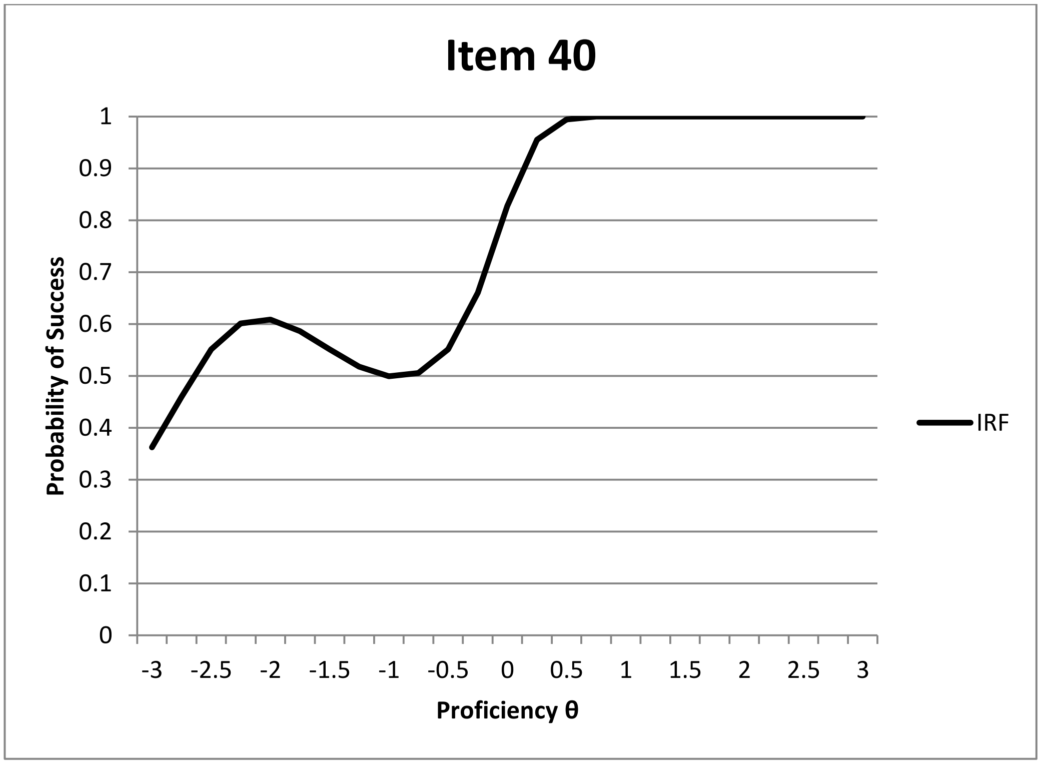

Figure 8. True IRF of simulated item 40. This item has high item discriminate parameter.

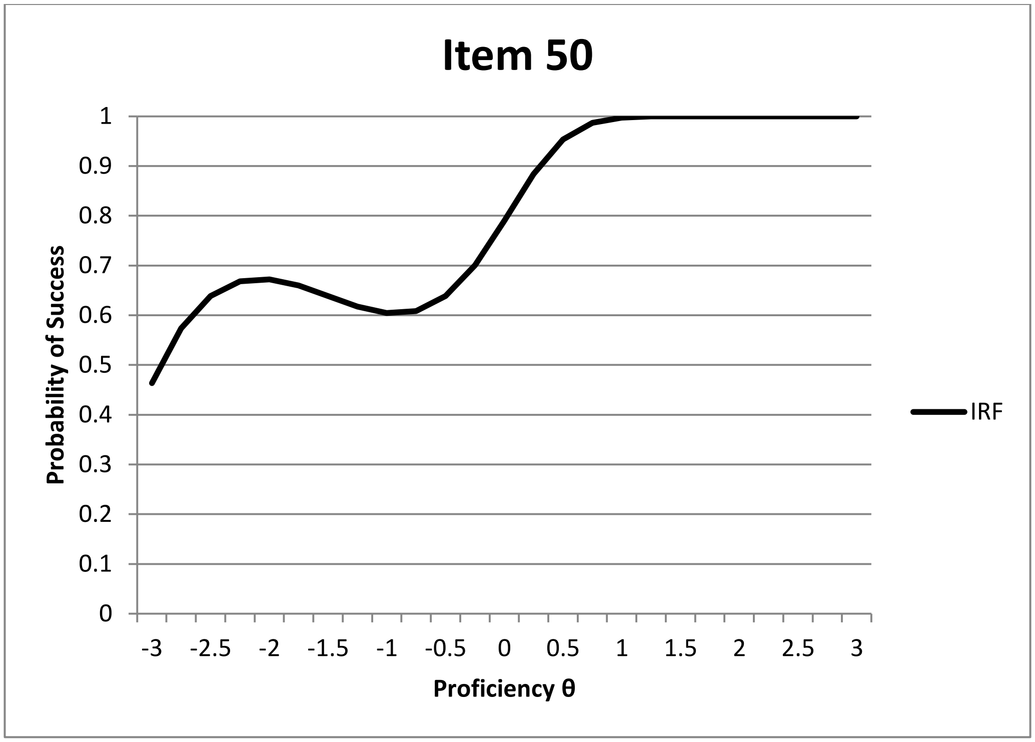

Figure 9. True IRF of simulated item 50. This item has low item guessing parameter.

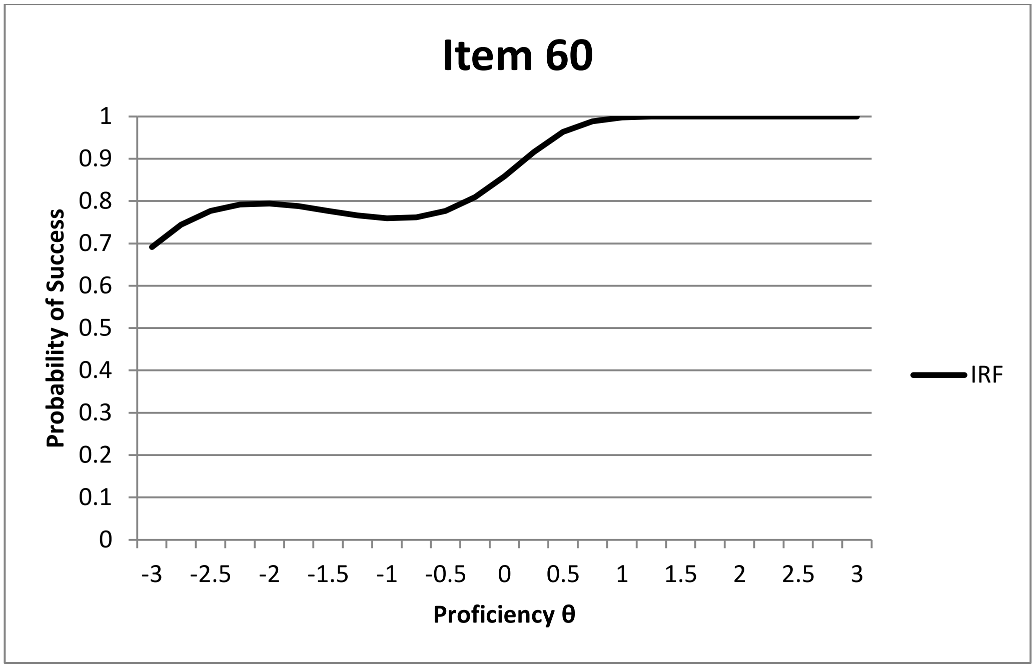

Figure 10. True IRF of simulated item 60. This item has high item guessing parameter.

Simulation method. The Monte Carlo simulation method was used to generate the data. To generate an item's response, a uniformly distributed random number in the range from 0 to 1 was generated first. Then this random number was compared with the probability of success calculated from the model. If the probability of success is bigger than the random number, the response is 1. Otherwise, the response is 0.

Replication. For one replication, the responses of 10,000 examinees on a simulated 60-item test were generated. The responses of six items were generated from the nonmonotonic model and the responses of 54 items were generated from the 3-PL model. Five hundred replications were generated.

IRF estimation and nonmonotonic area calculation. For every item, three nonparametric methods were used to estimate IRF, and numeric integration was used to calculate the area of nonmonotonic IRF. In one estimated IRF, there might be several small nonmonotonic IRF parts. Therefore, the sum of nonmonotonicity areas in one IRF was considered as the area of nonmonotonic IRF. The details about the IRF estimation and area calculation for each method are described below.

Nonparametric smooth regression. One hundred and one evaluation nodes were selected for estimating IRFs. These nodes range from -3 to 3 and the intervals between nodes are equal. These nodes were also the nodes for the numeric integration to calculate the nonmonotonic area of true and estimated IRF.

Item-ability regression. Since this estimation method needs the empirical proficiency distribution probability for each evaluation node from the output of the BILOG-MG program, and there were minor changes of the predefined evaluation nodes, which were ranged from -3 to 3, after the EM-estimation, 101 evaluation nodes are a little bit different from the evaluation nodes of the nonparametric smooth regression and B-splines nonparametric IRT method and the intervals between two nodes might not be equal. These nodes were also the nodes for the numeric integration to calculate the nonmonotonic area of true and estimated IRF.

B-splines nonparametric IRT. The model of this method to estimate IRF is the logistic transformation of the linear combination of the eleven B-spline basis functions with degree three and their coefficients. Nine uniformly spaced simple knots ranging from -3 to 3 and another six multiple endpoint knots were selected to estimate the coefficients of these B-spline basis functions. One hundred and one evaluation nodes were selected for plotting the IRF. These nodes range from -3 to 3 and the intervals between nodes are equal. These nodes are also the nodes for the numeric integration to calculate the nonmonotonic area of true and estimated IRF. These 101 evaluation nodes are the same as the evaluation nodes of the nonparametric smooth regression method.

Results Analysis. The results analysis in the simulation study aimed to compare three nonparametric methods. Two criteria were used for comparison: the type I and type II error rate. Also the nonmonotonic area correlation aimed to find the similarity and differences among three nonparametric methods on nonmonotonicity estimation.

Before calculating the type I and II error rates, a null distribution of the statistic, the nonmonotonic area, should be generated to evaluate whether the estimated IRFs are monotonic. This null distribution sets a criterion that if the nonmonotonic area of one IRF is bigger than most of the values in the null distribution (e.g., 95%), this IRF is nonmonotonic. A group of replicated data should be simulated first to generate a null distribution. Since this is a simulated data study and the true parameters are known, these true parameters were used instead of the draws from the posterior distribution of parameter estimates in the PPMC method to simulate replicated data. Five hundred replicated data were simulated using the 3-PL model and 60 groups of true parameters. Each replicated data has the responses of 10,000 examinees on a simulated 60-item test. The 60 IRFs were estimated by each nonparametric method and the areas of nonmonotonic IRF were calculated in every replicated data. Thus, for one IRF estimated by one nonparametric method, there were 500 areas of nonmonotonic IRF, which formed the null distribution.

Once the nonmonotonic area of one IRF was calculated in one replication by one nonparametric method, it was compared with the corresponding null distribution to calculate p-value. P-value equates to the percentage the number of values in the null distribution bigger than the nonmonotonic area. If this area is bigger than 95% values in the null distribution and p-value of this area is smaller than 0.05, then the item with this area was identified with a nonmonotonic IRF. Since the true IRFs are known, this conclusion was compared with the true situation. After the comparison was completed, the type I and II error rates of each nonparametric method can be calculated. For all 500 replications, there are 30,000 IRFs, which were estimated by one nonparametric method, with 60 in one replication. Among these 30,000 items, there are 3,000 items with nonmonotonic true IRFs and 27,000 items with monotonic true IRFs. The type I error rate is equal to the proportion of the number of items whose true IRF is monotonic, but the estimated IRF is nonmonotonic to the number of items with monotonic true IRFs (27,000). The type II error rate is equal to the proportion of the number of items whose true IRF is nonmonotonic, but estimated IRF is monotonic to the number of items with nonmonotonic true IRFs (3,000).

The nonmonotonic area correlation between two nonparametric methods is the correlation of 30,000 the nonmonotonic area calculated from the simulated data by these two methods. Since there are three nonparametric methods, there are three nonmonotonic area correlations. The high correlation indicates the nonmonotonic area estimated by two methods have a similar pattern among items. The low correlation indicates the nonmonotonic area estimated by two methods have a different pattern among items.

Real Data Study

The real data study has three phases: IRF estimation with nonmonotonic area calculation phase, items with nonmonotonic IRF identification phase, and nonmonotonic area correlation among three nonparametric methods. In the first phase, three nonparametric methods were used to estimate the IRFs of items on five end-of-year summative assessments. Then in the second phase the PPMC method was used to identify items with nonmonotonic IRFs. After this was done, the correlation of nonmonotonic area of real data calculated by three methods were calculated which aimed to find the similarity and differences among three nonparametric methods on nonmonotonicity estimation in real data study.

IRF estimation and nonmonotonic area calculation. Three nonparametric methods were used to estimate IRFs and the numeric integration was used to calculate the area of nonmonotonic IRF. In one estimated IRF, there might be several small nonmonotonic IRF parts. Therefore, the sum of nonmonotonicity areas in one IRF is considered as the area of nonmonotonic IRF. For every item, the same IRF estimation and nonmonotonic area calculation described earlier in the simulation study section was used.

Items with nonmonotonic IRF identification. The PPMC method was conducted to identify items with nonmonotonic IRFs through drawing parameters from the posterior distributions, generating replicated data, IRF estimation and nonmonotonic area calculation, and PPP-value calculation. This method started with drawing the item parameters and proficiency parameters from the posterior distributions. Then, the replicated data were generated from the 3-PL model using these parameters. Once this was done, the IRFs were estimated using three nonparametric methods and the nonmonotonic areas were calculated for every replicated data. At last, PPP-value using the nonmonotonic area as the discrepancy measure was calculated as a criterion to identify the items with nonmonotonic IRFs. The following section aims to describe the calculation steps of every part.

Posterior distribution and parameters. The posterior distributions of item and proficiency parameters can be simulated through the MCMC algorithm. WinBUGS was used to apply the MCMC algorithm on the real data to simulate the posterior distributions of 3-PL model item parameters and the proficiency parameters. The prior distributions used for MCMC algorithm were: $\log(a_i) \sim N(0,2),\ b_i \sim N(0,2),\ c_i \sim Beta(5,17), \theta_j \sim N(0,1)$. Sinharay et. al. (2006) stated PPMCs are “robust to reasonable changes to the prior distributions” (p. 304). For each end-of-year summative assessment, two Markov chains with the length of 6,000 were set to apply MCMC algorithm in WinBUGS. After these Markov chains converged, 500 groups of item parameters and proficiency parameters of 3-PL model were drawn from the posterior distributions simulated by the MCMC algorithm. One group of item parameters and proficiency parameters contains the parameters of all items on the assessment and 10,000 proficiency parameters.

Replicated data. For every group of draws, a replicated data were generated from the 3-PL model using these draws. This replicated data includes the responses of 10,000 examinees on all items in one assessment. Thus, 500 groups of draws generate 500 replicated data for every end-of-year summative assessment. These replicated data count the uncertainty of the estimation because they were generated using the draws from the posterior distributions of parameters.

IRF estimation and nonmonotonic area calculation. For every replicated data, three nonparametric methods were used to estimate IRFs, and the numeric integration was used to calculate the area of nonmonotonic IRFs. For every item, the same IRF estimation methods and nonmonotonic area calculations utilized in the simulated study were used. Since there are 500 replicated data for an end-of-year summative assessment, there are 500 estimated IRFs and nonmonotonic area values for one item whose IRF was estimated by one nonparametric method. Since the nonmonotonic area is the discrepancy measure, an item's 500 area values formed a posterior predictive distribution of this discrepancy measure of one nonparametric method.

PPP-value calculation. The posterior predictive distributions of the nonmonotonic area were used to calculate the PPP-value of every item for each of the three nonparametric methods. In order to compute the PPP-value, the nonmonotonic area of an item calculated from the real data was compared to its posterior predictive distribution generated from the replicated data. One item has three PPP-values corresponding to three nonparametric methods. These PPP-values are the criteria to judge whether the item has the nonmonotonic IRF. For example, if the PPP-value of one item whose IRF estimated by the nonparametric smooth regression method is an extreme value (less than 0.05), then this item was identified as the item with nonmonotonic IRF by the nonparametric smooth regression method.

Nonmonotonic area correlation. The nonmonotonic area correlation between two nonparametric methods is the correlation of the nonmonotonic area of all the items on each assessment calculated by these two methods. Since there are three nonparametric methods, there are three nonmonotonic area correlations for each assessment. The high correlation indicates the nonmonotonic area estimated by two methods have a similar pattern among items. The low correlation indicates the nonmonotonic area estimated by two methods have a different pattern among items.

Summary

The method of this research includes two parts: the simulation study and the real data study. The simulation study compared three nonparametric methods through simulated data generation, IRF estimation with nonmonotonic area calculation, and results analysis. The real data study calculated the nonmonotonic area of items on five assessments, identified items with nonmonotonic IRF and compared the nonmonotonicity estimation among three nonparametric methods.

Chapter 4: Results

The purposes of this study were to compare three nonparametric methods-the nonparametric smooth regression method, the item-ability regression method, and the B-spline nonparametric IRF method-in a simulation study and to identify the items with nonmonotonic IRFs on five assessments using these nonparametric methods. This chapter presents the comparison results of the simulation study and the calculation results of IRFs of real items on five assessments focusing on nonmonotonicity. These results were then used to identify the real items with nonmonotonic IRFs.

Simulation Study Results

Mean and standard deviation of p-value for each simulated item. P-value equates to the percentage the number of values in the null distribution bigger than the nonmonotonic area of one simulated item. P-value presents the extent of nonmonotonicity. The low p-value represents the large nonmonotonic area. Table 2 presents the mean and standard deviation of p-value for each simulated item over 500 replications for each method.

Table 2

Mean and Standard Deviation of P-value for Each Simulated Item over 500 replications of Three Nonparametric Methods

	Nonparametric Smooth Regression		Item-Ability Regression		B-Splines Nonparametric IRT	
Item	MeanP	StdP	MeanP	StdP	MeanP	StdP
1	0.496	0.278	0.714	0.383	0.552	0.304
2	0.505	0.297	0.569	0.342	0.546	0.298
3	0.504	0.283	0.588	0.365	0.525	0.297
4	0.501	0.291	0.965	0.179	0.577	0.331
5	0.510	0.293	0.595	0.368	0.542	0.294
6	0.482	0.300	0.708	0.387	0.513	0.299
7	0.494	0.304	0.794	0.3687	0.554	0.340
8	0.491	0.279	0.779	0.3737	0.539	0.307
9	0.491	0.284	0.615	0.3737	0.545	0.295
10	**0.264**	**0.229**	**0.250**	**0.2197**	**0.234**	**0.211**
11	0.494	0.297	0.967	0.176	0.567	0.334
12	0.515	0.293	0.733	0.386	0.570	0.315
13	0.485	0.298	0.633	0.377	0.523	0.292
14	0.492	0.289	0.527	0.309	0.556	0.286
15	0.516	0.313	0.994	0.077	0.620	0.369
16	0.514	0.301	0.992	0.088	0.602	0.353
17	0.499	0.303	0.994	0.077	0.618	0.357
18	0.509	0.284	0.681	0.383	0.564	0.314
19	0.531	0.308	0.923	0.260	0.576	0.330
20	**0.071**	**0.113**	**0.472**	**0.483**	**0.06**	**0.099**
21	0.491	0.299	0.978	0.145	0.618	0.351
22	0.486	0.285	0.517	0.297	0.529	0.287
23	0.476	0.297	0.615	0.373	0.507	0.306
24	0.487	0.291	0.906	0.276	0.560	0.349
25	0.485	0.291	0.831	0.345	0.532	0.320
26	0.501	0.290	0.661	0.381	0.540	0.303
27	0.487	0.284	0.635	0.377	0.529	0.286
28	0.497	0.285	0.815	0.357	0.585	0.336
29	0.517	0.300	0.590	0.351	0.533	0.306
30	**0.170**	**0.193**	**0.073**	**0.197**	**0.174**	**0.177**
31	0.508	0.285	0.612	0.366	0.565	0.317
32	0.481	0.292	0.710	0.394	0.518	0.311
33	0.484	0.281	0.544	0.310	0.538	0.274
34	0.509	0.293	0.988	0.107	0.618	0.349
35	0.492	0.283	0.727	0.386	0.537	0.314
36	0.495	0.304	0.996	0.063	0.590	0.352
37	0.494	0.280	0.537	0.318	0.528	0.280
38	0.501	0.293	0.588	0.363	0.530	0.291
39	0.496	0.299	0.718	0.390	0.552	0.320

40	**0.223**	**0.216**	**0.160**	**0.123**	**0.178**	**0.164**
41	0.518	0.306	0.721	0.387	0.589	0.318
42	0.499	0.288	0.802	0.365	0.544	0.325
43	0.518	0.309	0.883	0.304	0.596	0.352
44	0.507	0.294	0.550	0.331	0.540	0.289
45	0.489	0.288	0.708	0.388	0.522	0.296
46	0.486	0.285	0.622	0.376	0.546	0.294
47	0.523	0.294	0.960	0.193	0.630	0.3418
48	0.497	0.293	0.809	0.354	0.558	0.328
49	0.530	0.294	0.953	0.207	0.598	0.328
50	**0.030**	**0.056**	**0.057**	**0.230**	**0.019**	**0.053**
51	0.503	0.299	0.990	0.098	0.583	0.354
52	0.494	0.275	0.576	0.364	0.542	0.2872
53	0.532	0.288	0.668	0.394	0.575	0.312
54	0.510	0.296	0.708	0.386	0.585	0.328
55	0.513	0.283	0.694	0.385	0.543	0.303
56	0.510	0.296	0.922	0.258	0.561	0.335
57	0.508	0.287	0.751	0.377	0.553	0.315
58	0.510	0.293	0.547	0.339	0.551	0.296
59	0.523	0.310	0.936	0.236	0.603	0.330
60	**0.033**	**0.056**	**0.053**	**0.222**	**0.029**	**0.060**

Table 2 shows the mean p-values of items with true nonmonotonic IRFs are much smaller than the mean p-values of the items with true monotonic IRFs although some of them are bigger than 0.05.

Type I and type II error rate. The simulation study first compared three nonparametric methods using type I error rate and type II error rate as the criteria. For each item with true monotonic IRF, the type I error rate is equal to the proportion of the number of replications that the estimated IRF is nonmonotonic to the number of replications (500). For each item with true nonmonotonic IRF, the type II error rate is equal to the proportion of the number of replications that the estimated IRF is monotonic to the number of replications (500). Table 3 includes the type I error rate of three nonparametric methods for all items with true nonmonotonic IRFs.

Table 3

Type I Rate of Three Nonparametric Methods for Each Item with True Monotonic IRF

Item	Nonparametric Smooth Regression	Item-Ability Regression	B-Splines Nonparametric IRT
1	0.028	0.036	0.028
2	0.05	0.058	0.042
3	0.04	0.052	0.044
4	0.06	0.024	0.04
5	0.05	0.04	0.034
6	0.062	0.042	0.046
7	0.062	0.036	0.052
8	0.048	0.022	0.034
9	0.05	0.046	0.036
11	0.044	0.026	0.034
12	0.05	0.028	0.044
13	0.068	0.048	0.038
14	0.056	0.04	0.044
15	0.034	0.006	0.052
16	0.06	0.008	0.054
17	0.058	0.006	0.05
18	0.042	0.044	0.036
19	0.058	0.044	0.036
21	0.048	0.022	0.044
22	0.058	0.046	0.044
23	0.068	0.038	0.048
24	0.062	0.016	0.068
25	0.062	0.024	0.046
26	0.048	0.042	0.032
27	0.054	0.042	0.06
28	0.038	0.036	0.056
29	0.044	0.038	0.038
31	0.04	0.036	0.038
32	0.062	0.054	0.05
33	0.042	0.038	0.034
34	0.048	0.012	0.036
35	0.044	0.042	0.048
36	0.044	0.004	0.036
37	0.034	0.038	0.026
38	0.05	0.036	0.04
39	0.062	0.032	0.044
41	0.062	0.046	0.028
42	0.038	0.032	0.048
43	0.07	0.03	0.044
44	0.052	0.048	0.042
45	0.056	0.046	0.044

46	0.036	0.056	0.034
47	0.048	0.03	0.028
48	0.046	0.024	0.04
49	0.032	0.02	0.03
51	0.048	0.01	0.058
52	0.048	0.03	0.034
53	0.04	0.036	0.044
54	0.056	0.03	0.044
55	0.074	0.034	0.052
56	0.052	0.024	0.044
57	0.052	0.024	0.046
58	0.054	0.044	0.054
59	0.044	0.028	0.034

Table 3 shows that the type I error rate for each item is low. For every item, the type I error rates of three nonparametric methods are closely alike.

Table 4 includes the type II error rate of three nonparametric methods for all items with true nonmonotonic IRFs.

Table 4
Type II Rate of Three Nonparametric Methods for Each Item with True Nonmonotonic IRF

Item	Nonparametric Smooth Regression	Item-Ability Regression	B-Splines Nonparametric IRT
10	0.816	0.792	0.824
20	0.400	0.590	0.332
30	0.628	0.280	0.750
40	0.746	0.770	0.700
50	0.194	0.056	0.072
60	0.206	0.052	0.166

Table 4 presents that the type II error rates of every nonparametric method are low for items 50 and 60. On the other hand, the type II error rates of every nonparametric method are relatively high for other four items especially items10 and 40. When three nonparametric methods were compared for each item, the nonparametric smooth regression method and the B-splines nonparametric IRT method have similar type II error rates and the type II error rate of these two methods are higher than the type II error rate of the item-ability regression method except items 20 and 40.

Table 5 summaries the average type I and type II error rates of three nonparametric methods.

Table 5
Average Type I and II Rates of Three Nonparametric Methods

	Nonparametric Smooth Regression	Item-Ability Regression	B-Splines Nonparametric IRT
Average Type I error	0.051	0.033	0.042
Average Type II error	0.498	0.423	0.474

The item-ability regression method has the lowest type I and type II error rates (0.033 and 0.432). The nonparametric smooth regression method has the highest type I and type II error rates (0.051 and 0.498). The differences of type I and type II error rates among three methods are not large. To sum up, the type II error rate is much higher than the type I error rate of each method.

Estimated IRFs and true IRFs of items with true nonmonotonic IRFs. Figures 11-16 are the IRFs estimated by three nonparametric methods and the true IRFs of items 10, 20, 30, 40, 50, and 60 from the replication 200.

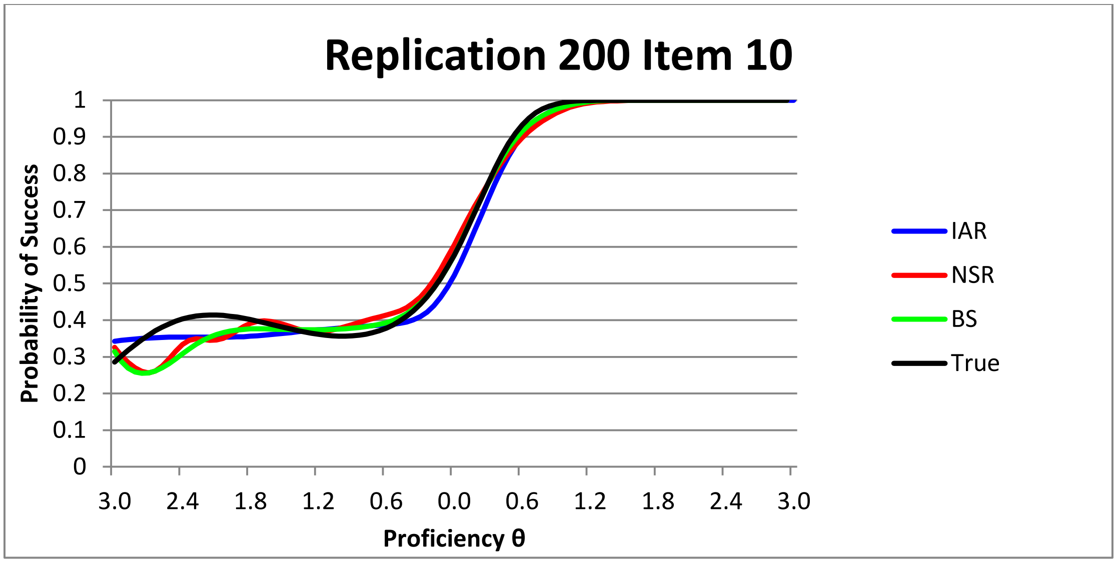

Figure 11. IRFs estimated by three nonparametric methods for simulated item 10 from replication 200. The black line is the true IRF.

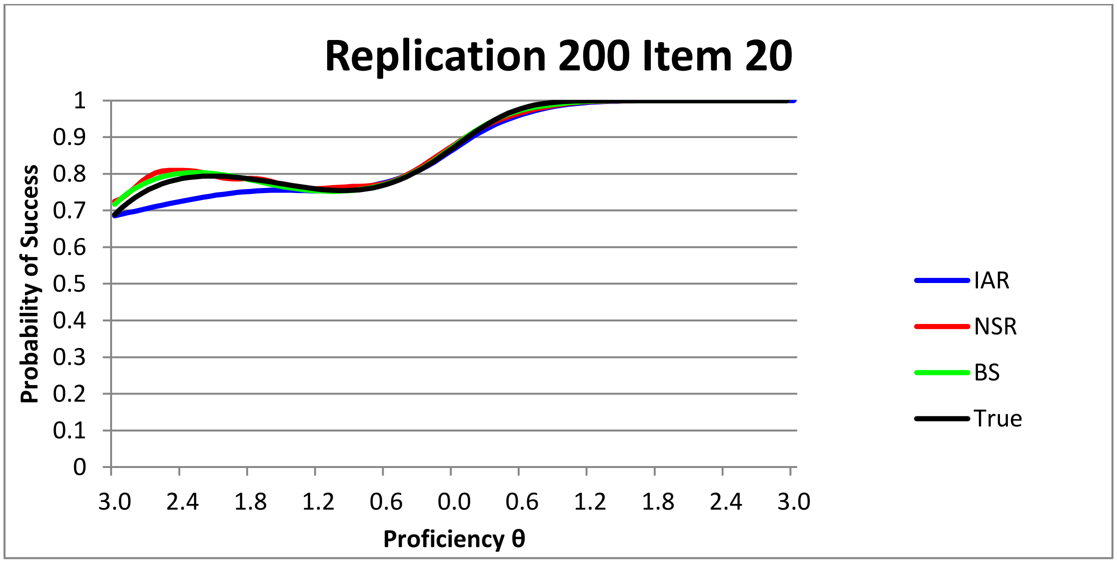

Figure 12. IRFs estimated by three nonparametric methods for simulated item 20 from replication 200. The black line is the true IRF.

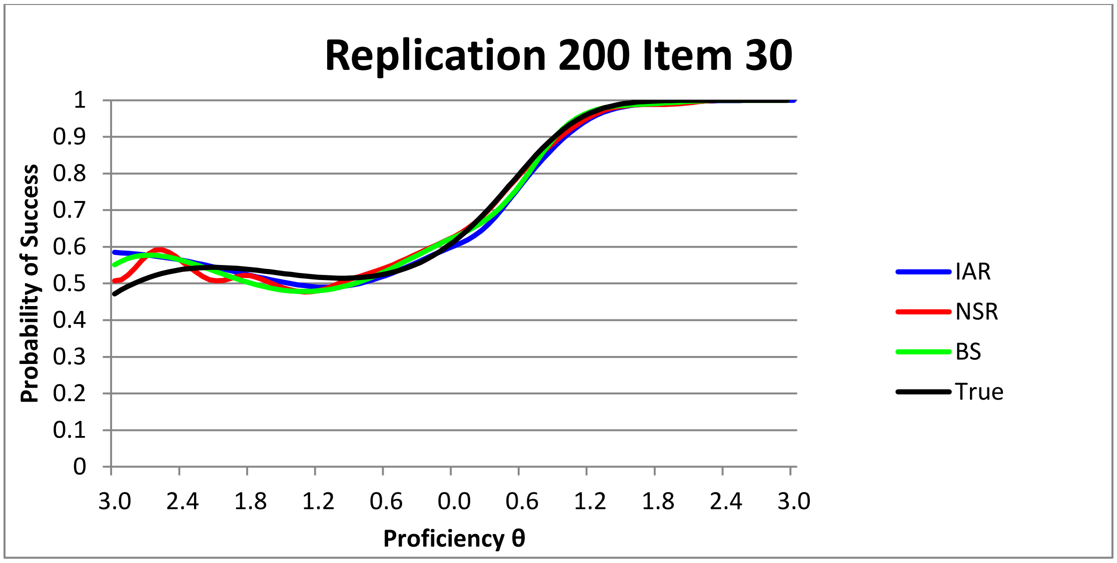

Figure 13. IRFs estimated by three nonparametric methods for simulated item 30from replication 200. The black line is the true IRF.

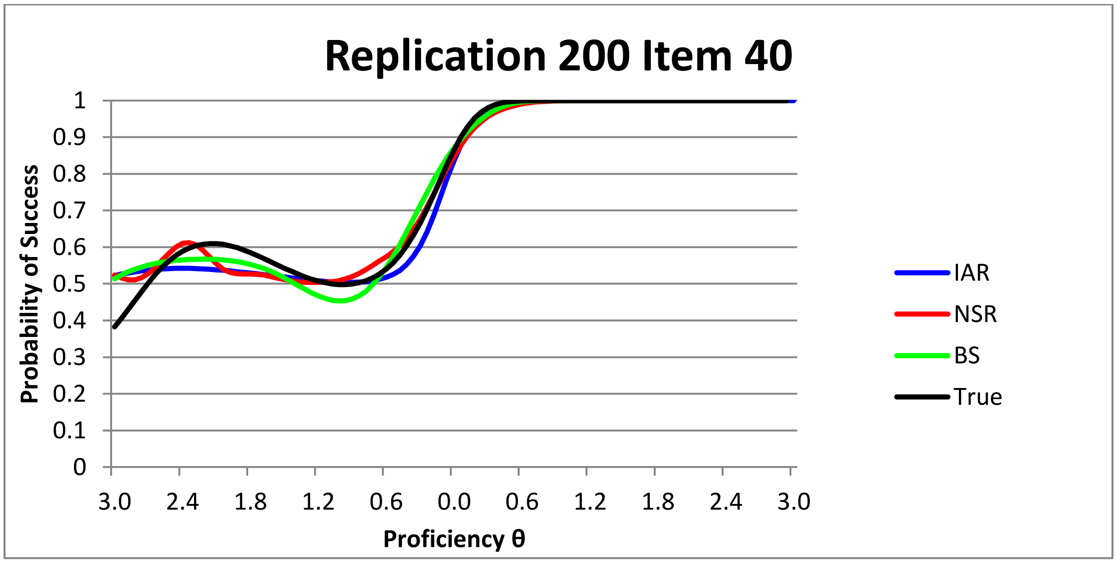

Figure 14. IRFs estimated by three nonparametric methods for simulated item 40from replication 200. The black line is the true IRF.

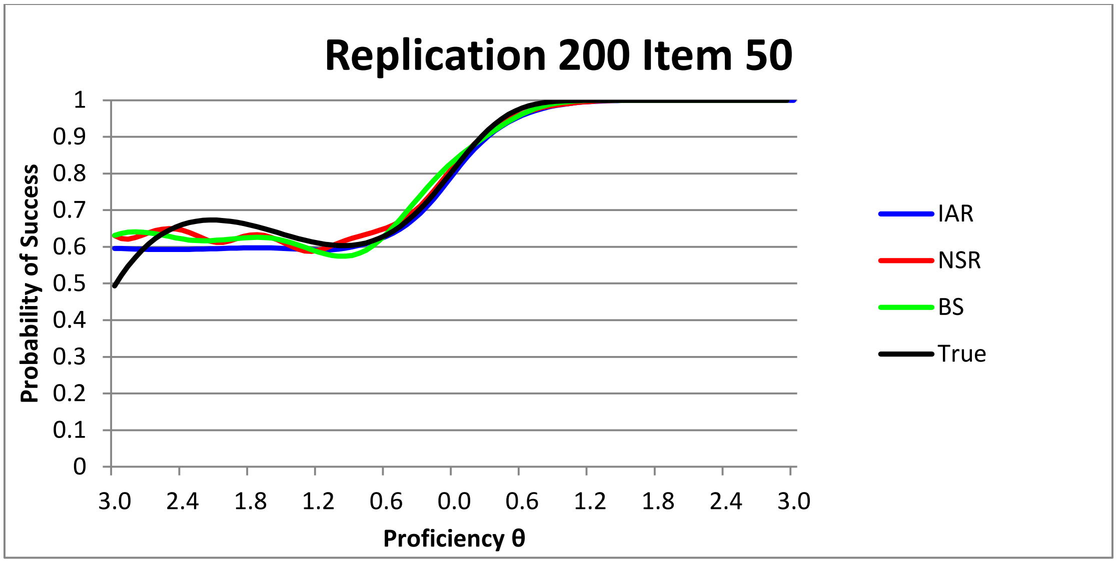

Figure 15. IRFs estimated by three nonparametric methods for simulated item 50from replication 200. The black line is the true IRF.

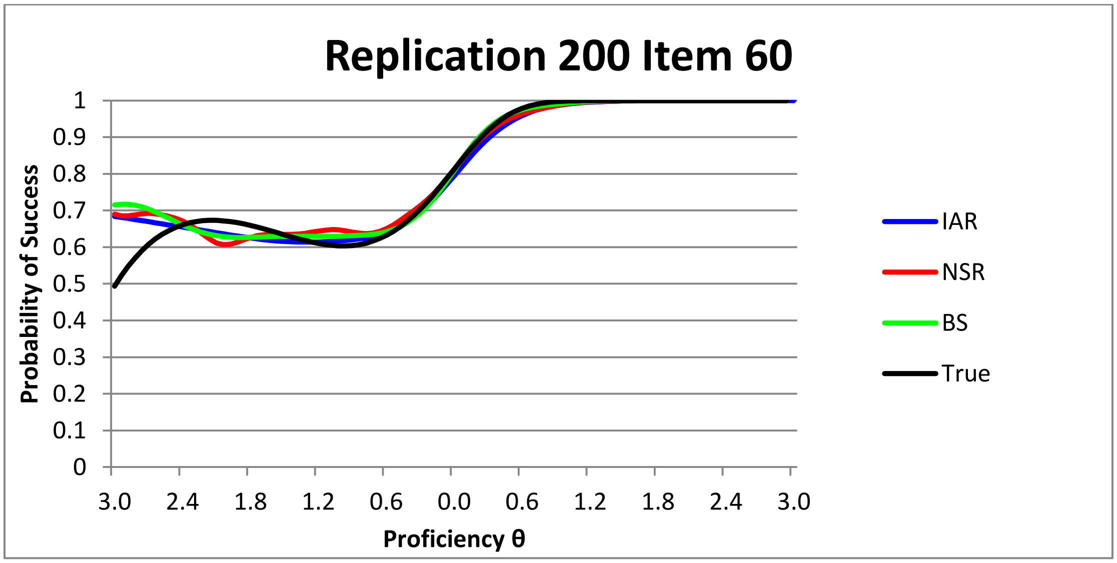

Figure 16. IRFs estimated by three nonparametric methods for simulated item 60from replication 200. The black line is the true IRF.

Nonmonotonic area correlation. Table 6 illustrates the nonmonotonic area correlation among three nonparametric methods. The nonmonotonic area correlation between two nonparametric methods is the correlation of 30,000 the nonmonotonic area calculated from the simulated data by these two methods.

Table 6

Nonmonotonic Area Correlation among Three Nonparametric Methods for the Simulated Data

	Nonparametric Smooth Regression	Item-Ability Regression	B-Splines Nonparametric IRT
Nonparametric Smooth Regression		0.510	0.704
Item-Ability Regression			0.507
B-Splines Nonparametric IRT			

The nonmonotonic area correlation between the nonparametric smooth regression method and the B-splines nonparametric IRT method is high (0.704). In addition, the nonmonotonic area correlation between the nonparametric smooth regression method and the item-ability regression method (0.510) and the nonmonotonic area correlation between the B-splines nonparametric IRT method and the item-ability regression method (0.507) are similar. The high nonmonotonic correlation between the nonparametric smooth regression method and the B-splines nonparametric IRT method indicates the nonmonotonic area estimated by these two methods have a similar pattern among items. The nonmonotonic area pattern among items estimated by the item-ability regression method is not as similar as the nonmonotonic area pattern estimated by the nonparametric smooth regression method and the B-splines nonparametric IRT method.

Real Data Study Results

The real data study results are presented in order of assessments. For each assessment, the PPP-values for every item of three nonparametric methods are presented first, the nonmonotonic area correlation of all the items on the assessment among three nonparametric methods is shown next, and the IRFs of some items with low PPP-value are presented at last.

Math assessment A. Table 7 includes the PPP-values of three nonparametric methods for each item on this assessment.

Table 7
PPP-Value of Three Nonparametric Methods for Each Item on Math Assessment A

	Nonparametric Smooth	Item-Ability	B-Splines Nonparametric
Item	Regression	Regression	IRT
1	0.256	1.000	0.310
2	0.426	1.000	0.462
3	0.776	1.000	0.860
4	**0.030**	0.200	0.166
5	**0.014**	**0.006**	**0.014**
6	**0.044**	1.000	0.142
7	0.384	0.214	0.616
8	0.746	1.000	0.760
9	0.542	1.000	0.752
10	0.302	0.630	0.272
11	0.098	1.000	0.120
12	0.588	1.000	0.862
13	0.296	0.074	0.174
14	0.318	0.276	0.450
15	0.278	0.070	0.402
16	0.834	0.604	0.820
17	0.432	0.672	0.430
18	0.554	0.434	0.800
19	0.866	1.000	0.698
20	0.260	0.140	0.384
21	0.900	1.000	1.000
22	1.000	1.000	1.000
23	0.916	1.000	1.000
24	0.318	1.000	0.122
25	0.722	0.640	0.908
26	0.348	0.556	0.270
27	0.428	0.450	0.362
28	1.000	1.000	0.776
29	0.180	0.100	0.100
30	0.384	0.546	0.310
31	0.914	0.494	1.000
32	0.196	1.000	0.208
33	**0.036**	1.000	**0.018**
34	0.704	1.000	0.608
35	0.158	1.000	0.144
36	0.072	0.100	0.118
37	**0.000**	0.052	**0.000**
38	0.640	0.426	0.758
39	0.862	1.000	0.946
40	0.220	1.000	0.482
41	0.610	0.068	0.448

42	0.430	1.000	0.406
43	0.858	0.154	0.876
44	**0.008**	**0.000**	**0.006**
45	0.326	1.000	0.630
46	0.644	1.000	0.812
47	0.536	1.000	0.434
48	**0.032**	1.000	0.182
49	0.762	0.914	0.820
50	0.440	0.470	0.442
51	**0.012**	1.000	**0.026**
52	0.538	1.000	0.352
53	0.776	1.000	0.810
54	0.732	1.000	0.944
55	0.950	0.534	0.750
56	0.792	1.000	0.544
57	**0.002**	**0.000**	**0.004**
58	0.482	0.234	1.000
59	0.254	1.000	0.168
60	0.280	0.142	0.342
61	0.122	1.000	0.120
62	0.332	1.000	0.572
63	0.200	0.412	0.340
64	0.844	0.056	0.644
65	0.082	0.318	**0.040**
66	0.142	0.332	0.164
67	0.344	1.000	0.628
68	0.834	0.376	0.836
69	0.058	**0.000**	**0.034**
70	0.782	0.544	0.862
71	0.244	0.522	0.304
72	0.846	1.000	1.000
73	0.396	1.000	0.248
74	0.070	0.166	0.098
75	0.400	0.110	0.278
76	0.054	0.474	0.086
77	0.200	0.104	0.208
78	0.088	1.000	0.068
79	0.800	1.000	0.756
80	0.404	0.342	0.332
81	0.204	1.000	0.224
82	0.840	1.000	0.876
83	0.206	0.196	0.132

Nine items are identified with nonmonotonic IRFs by the nonparametric smooth regression method, four items are identified with nonmonotonic IRFs by the item-ability regression method, and eight items are identified with nonmonotonic IRFs by the B-splines nonparametric IRT method. Items 5, 44 and 57 are identified with nonmonotonic IRFs by three methods. Items 37 and 69 are identified with nonmonotonic IRFs by two methods and the PPP values of the third method for these two items are very close to 0.05 (0.052 of the item-ability regression method for item 37 and 0.058 of the nonparametric smooth regression method for item 69).

Table 8 presents the nonmonotonic area correlation among three nonparametric methods. The nonmonotonic area correlation between two nonparametric methods is the correlation of the nonmonotonic area of all items on the math assessment A calculated by these two methods.

Table 8
Nonmonotonic Area Correlation among Three Nonparametric Methods for Math Assessment A

	Nonparametric Smooth Regression	Item-Ability Regression	B-Splines Nonparametric IRT
Nonparametric Smooth Regression		0.857	0.984
Item-Ability Regression			0.853
B-Splines Nonparametric IRT			

The nonmonotonic area correlation between the nonparametric smooth regression method and the B-splines nonparametric IRT method is close to 1 (0.984). Also, the nonmonotonic area correlation between the nonparametric smooth regression method and the item-ability regression method (0.857) and the nonmonotonic area correlation between the B-splines nonparametric IRT method and the item-ability regression (0.853) are high and similar. The very high nonmonotonic correlation between the nonparametric smooth regression method and the B-splines nonparametric IRT method indicates the nonmonotonic area estimated by these two methods have a similar pattern among items on the math assessment A. The nonmonotonic area correlations among these three methods are high which indicate that the nonmonotonic area patterns among items on the math assessment A estimated by these three methods are similar. For example, there are 11 items identified with nonmonotonic IRFs by at least one method and the PPP-values of three nonparametric methods for five items are very similar.

Figures 17-27 present IRFs estimated by three nonparametric methods for items identified with nonmonotonic IRFs by at least one method. Figures 17, 18, and 19 present the IRFs for items 5, 44 and 57 which are identified with nonmonotonic IRFs by three nonparametric methods.

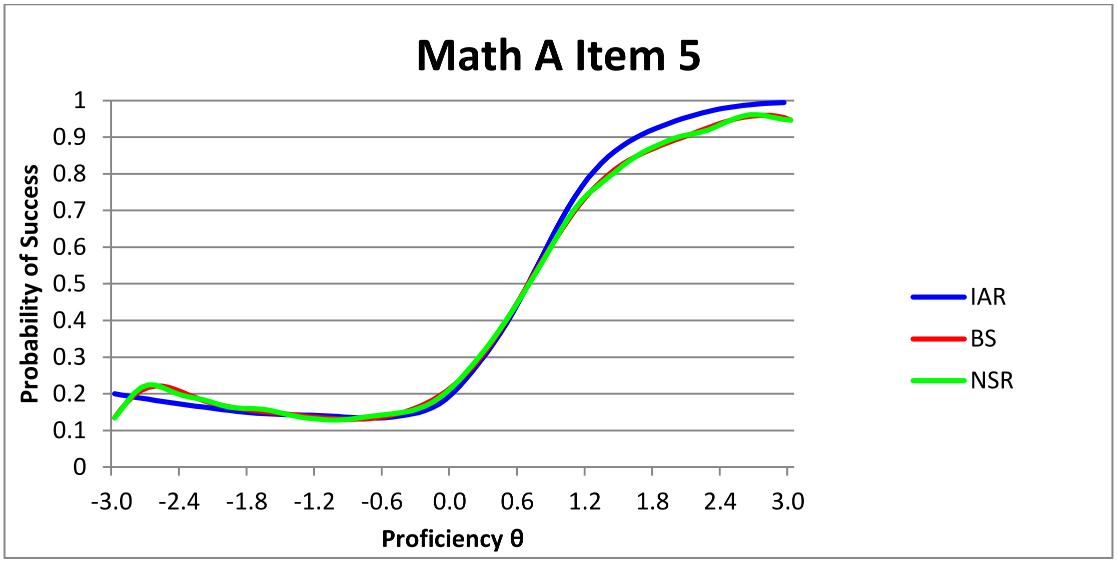

*Figure 17.*IRFs estimated by three nonparametric methods for item 5 on math A. The PPP-value of the item-ability regression method is 0.006. The PPP-value of the B-splines nonparametric IRT method is 0.014. The PPP-value of the nonparametric smooth regression method is 0.014.

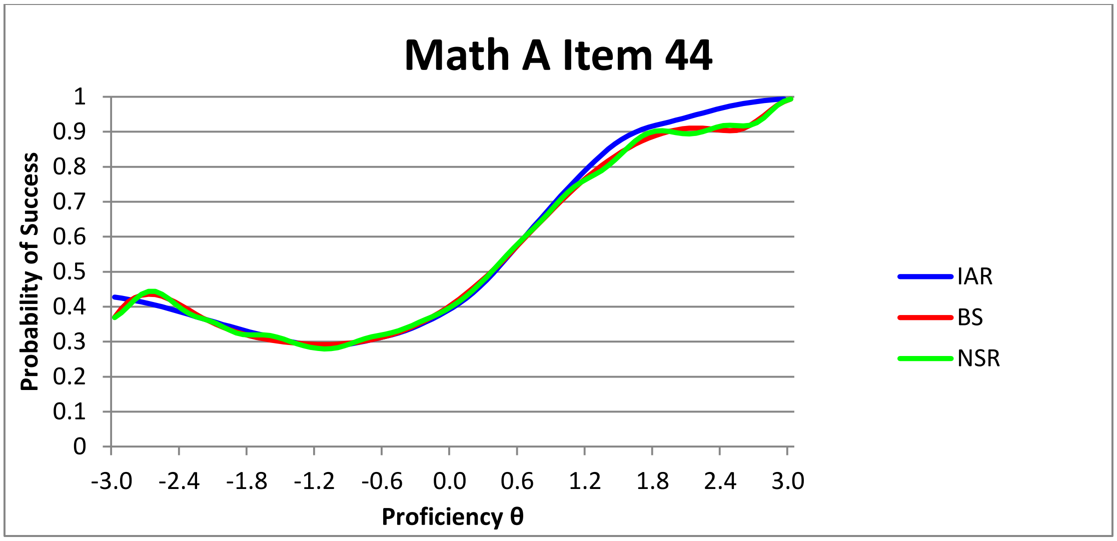

Figure 18. IRFs estimated by three nonparametric methods for item 44 on math A. The PPP-value of the item-ability regression method is 0.000. The PPP-value of the B-splines nonparametric IRT method is 0.006. The PPP-value of the nonparametric smooth regression method is 0.008.

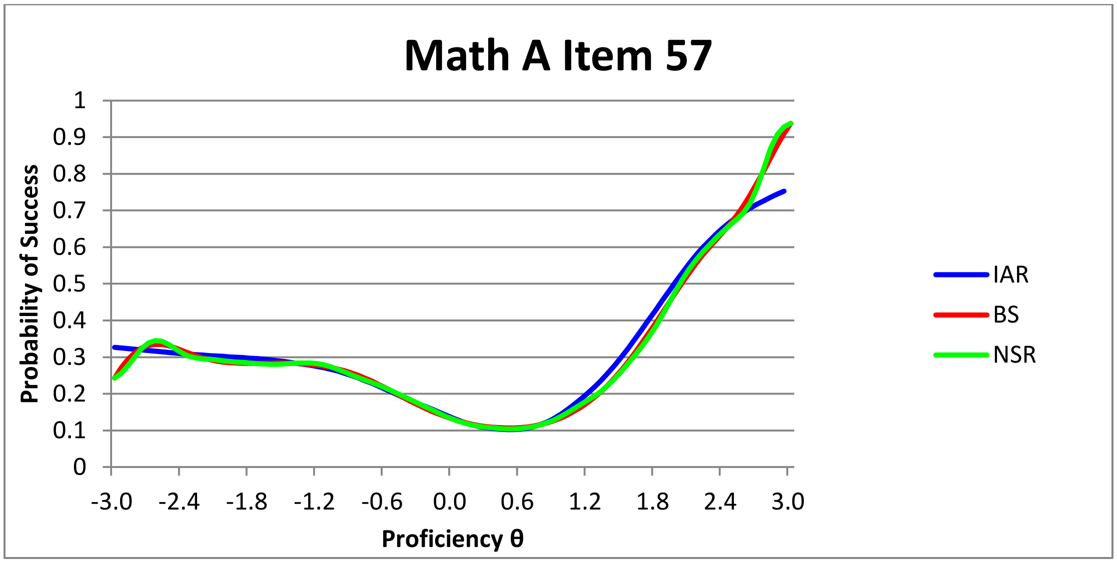

*Figure 19.*IRFs estimated by three nonparametric methods for item 57 on math A. The PPP-value of the item-ability regression method is 0.006. The PPP-value of the B-splines nonparametric IRT method is 0.014. The PPP-value of the nonparametric smooth regression method is 0.014.

For these three items, the IRFs of three nonparametric methods are very similar with the exception of the end parts at the very low and high theta range and the nonmonotonic parts of IRFs of three methods appear at the same theta range.

Figures 20 and 21 present the IRFs for items 37 and 69 which are identified with nonmonotonic IRFs by two methods and the PPP values of the third method for these two items are very close to 0.05.

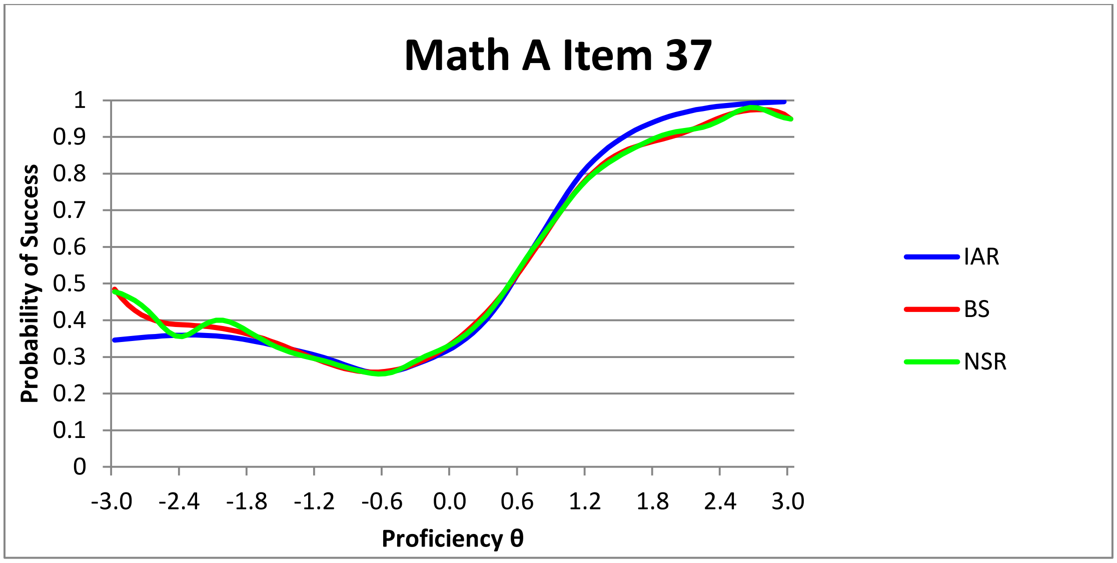

*Figure 20.*IRFs estimated by three nonparametric methods for item 37 on math A. The PPP-value of the item-ability regression method is 0.052. The PPP-value of the B-splines nonparametric IRT method is 0.000. The PPP-value of the nonparametric smooth regression method is 0.000.

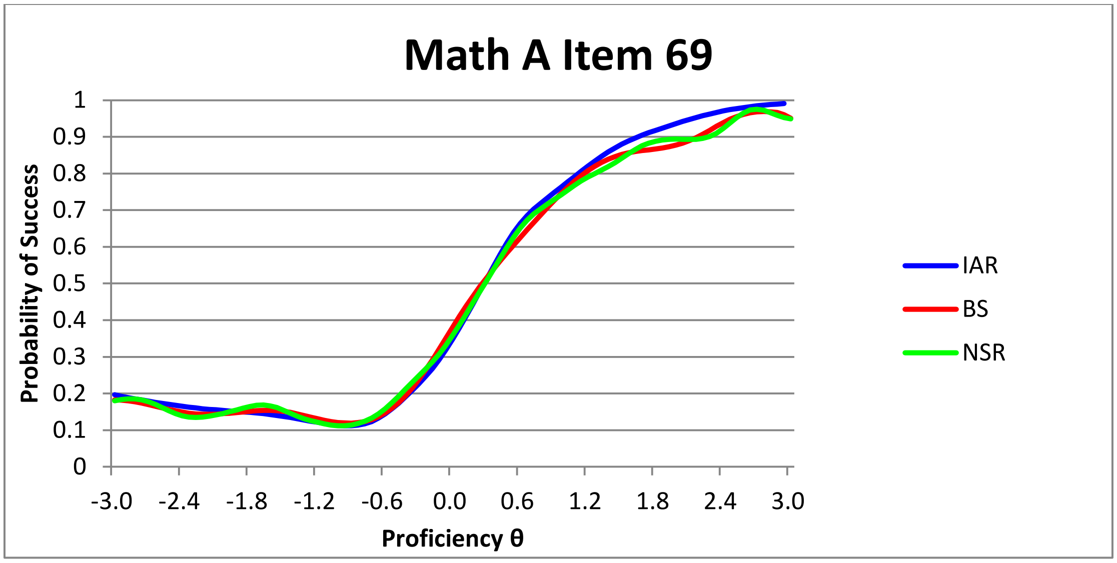

Figure 21.IRFs estimated by three nonparametric methods for item 69 on math A. The PPP-value of the item-ability regression method is 0.000. The PPP-value of the B-splines nonparametric IRT method is 0.034. The PPP-value of the nonparametric smooth regression method is 0.058.

For these two items, the differences among the IRFs of three nonparametric methods are small and the nonmonotonic part appears at the similar theta range.

Figures 22-27 present the IRFs of the other items identified with nonmonotonic IRFs by at least one method. These items are 4, 6, 33, 48, and 51, 65.

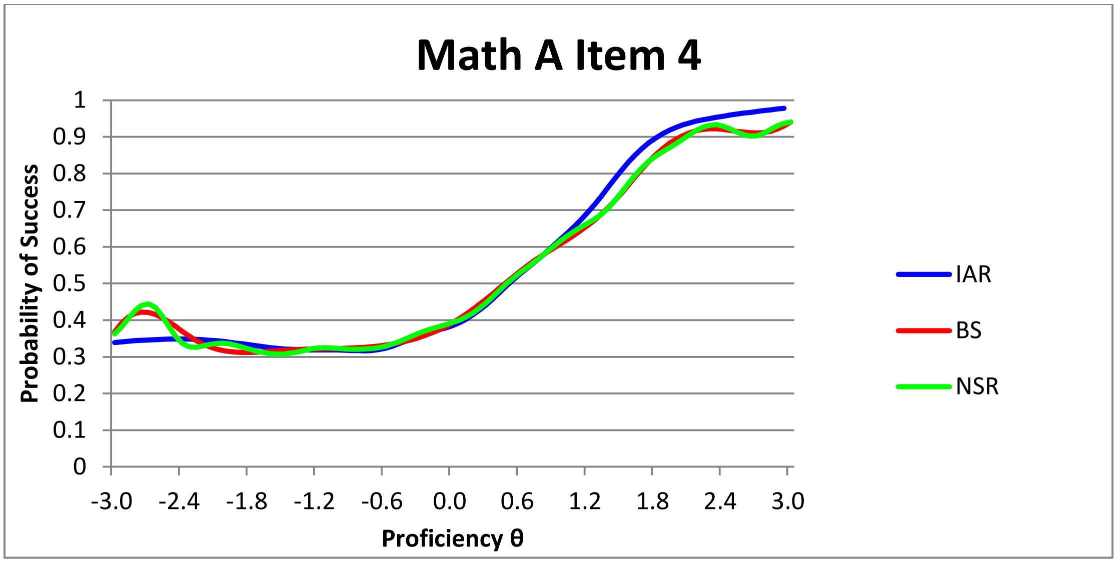

*Figure 22.*IRFs estimated by three nonparametric methods for item 4 on math A. The PPP-value of the item-ability regression method is 0.200. The PPP-value of the B-splines nonparametric IRT method is 0.166. The PPP-value of the nonparametric smooth regression method is 0.030.

For item 4, the IRFs of three nonparametric methods are very similar with the exception of the end parts at the very low and high theta range. The nonmonotonic parts of three IRFs appear at the same theta range but not very obvious.

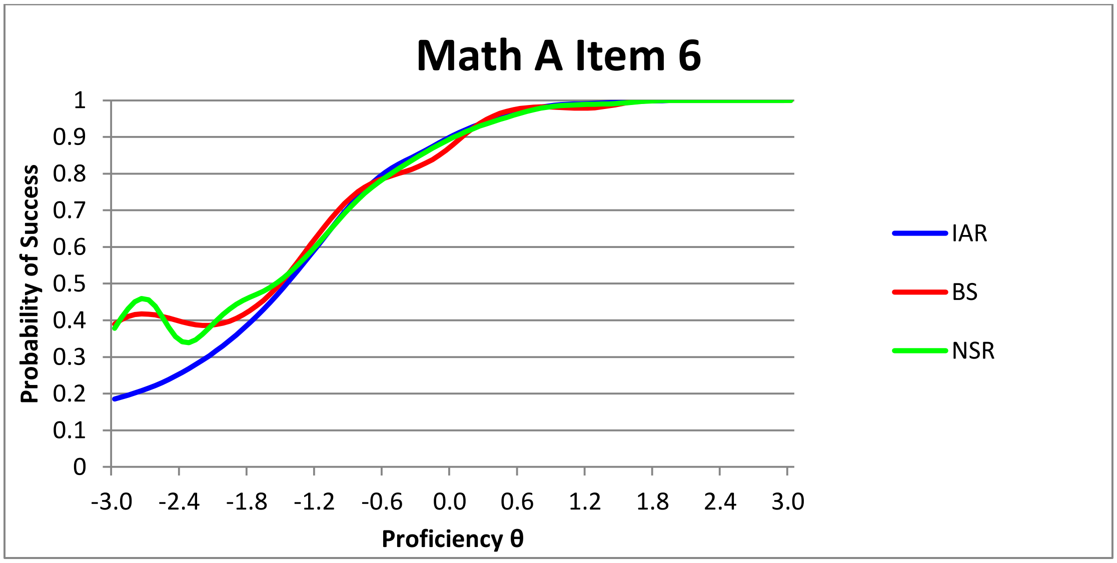

*Figure 23.*IRFs estimated by three nonparametric methods for item 6 on math A. The PPP-value of the item-ability regression method is 1.000. The PPP-value of the B-splines nonparametric IRT method is 0.144. The PPP-value of the nonparametric smooth regression method is 0.044.

For item 6, the IRFs of the nonparametric smooth regression method and the B-splines nonparametric IRT method are similar but are different from the IRF of the item-ability regression method at the low theta range. There are several small nonmonotonic parts in the IRFs of the nonparametric smooth regression method and the B-spline nonparametric IRT method. The IRF of the item-ability regression method is monotonic.

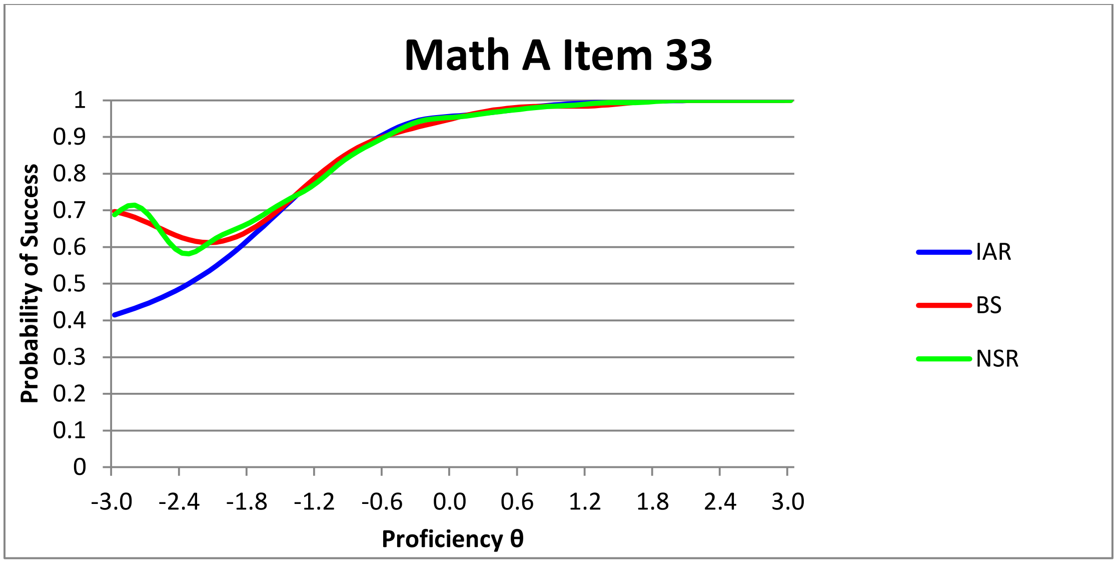

*Figure 24.*IRFs estimated by three nonparametric methods for item 33 on math A. The PPP-value of the item-ability regression method is 1.000. The PPP-value of the B-splines nonparametric IRT method is 0.018. The PPP-value of the nonparametric smooth regression method is 0.036.

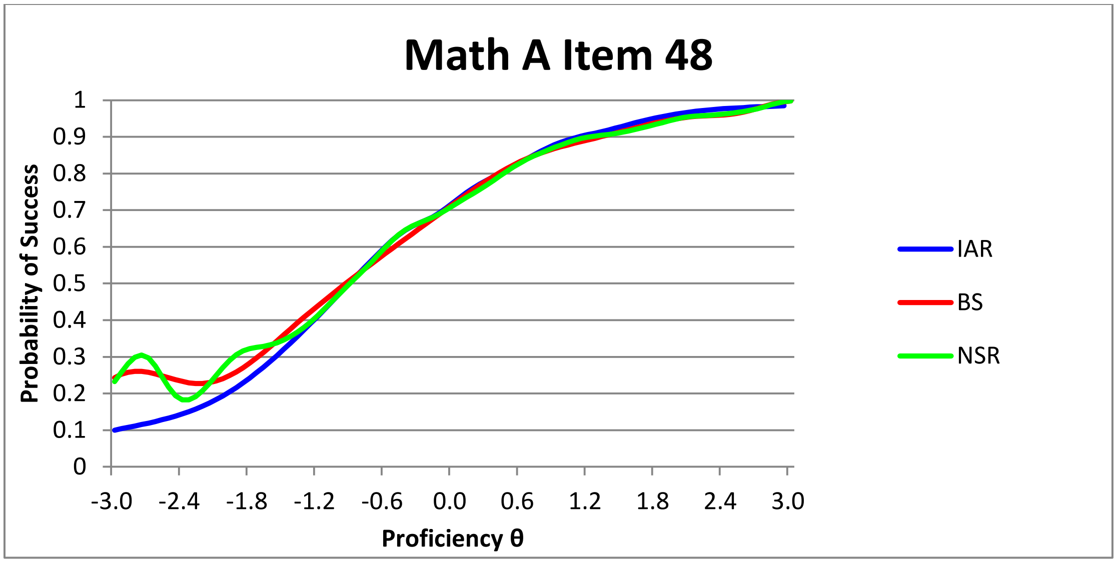

*Figure 25.*IRFs estimated by three nonparametric methods for item 48 on math A. The PPP-value of the item-ability regression method is 1.000. The PPP-value of the B-splines nonparametric IRT method is 0. 182. The PPP-value of the nonparametric smooth regression method is 0.032.

For item 33 and 48, the IRFs of the nonparametric smooth regression method and the B-splines nonparametric IRT method are very similar but are different from the IRF of the item-ability regression method at the low theta range. The IRF of the item-ability regression method is monotonic.

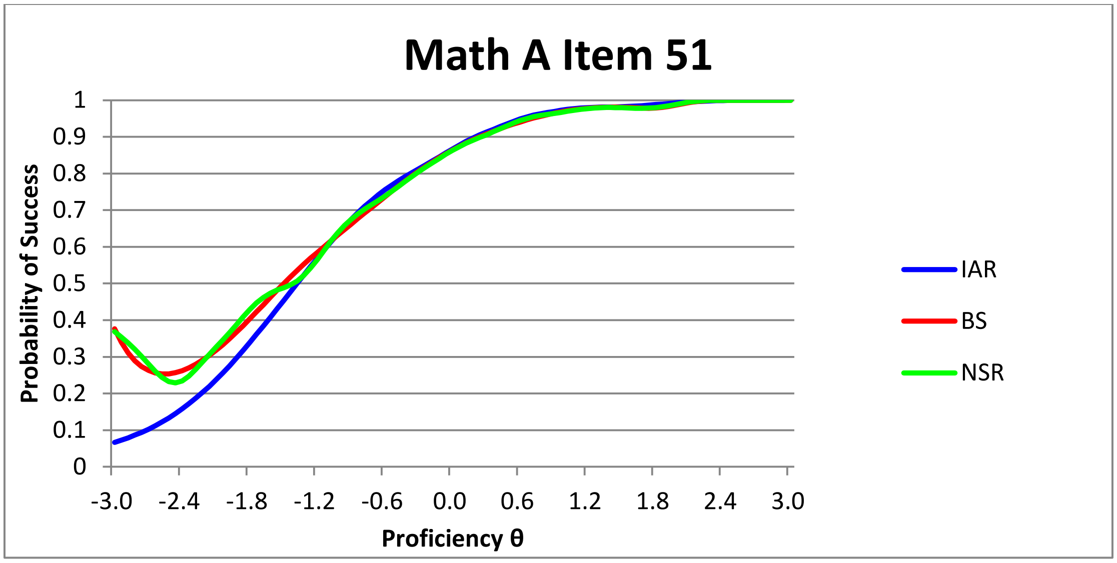

*Figure 26.*IRFs estimated by three nonparametric methods for item 51 on math A. The PPP-value of the item-ability regression method is 1.000. The PPP-value of the B-splines nonparametric IRT method is 0.026. The PPP-value of the nonparametric smooth regression method is 0.012.

For item 51, the IRFs of the nonparametric smooth regression method and the B-splines nonparametric IRT method are very similar but are different from the IRF of the item-ability regression method at the low theta range. The IRF of the item-ability regression method is monotonic.

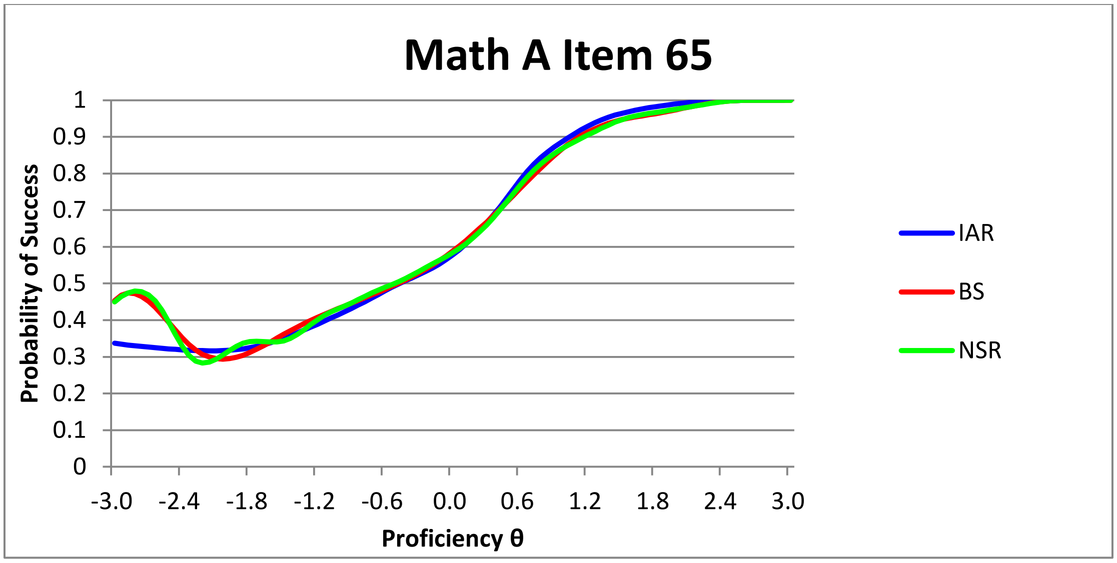

*Figure 27.*IRFs estimated by three nonparametric methods for item 65 on math A. The PPP-value of the item-ability regression method is 0.318. The PPP-value of the B-splines nonparametric IRT method is 0.040. The PPP-value of the nonparametric smooth regression method is 0.082.

For item 65, the IRFs of the nonparametric smooth regression method and the B-splines nonparametric IRT method are very similar but are different from the IRF of the item-ability regression method at the nonmonotonic part which appears at the low theta range. The nonmonotonic part of the IRF of the item-ability regression method is not very obvious.

Math assessment B. Table 9 includes the PPP-values of three nonparametric methods for each item on this assessment.

Table 9

PPP-Value of Three Nonparametric Methods for Each Item on Math Assessment B

Item	Nonparametric Smooth Regression	Item-Ability Regression	B-Splines Nonparametric IRT
1	0.846	1.000	0.670
2	0.828	1.000	0.160
3	0.234	0.140	0.352
4	0.902	1.000	0.842
5	0.660	0.562	0.222
6	0.898	1.000	0.364
7	**0.040**	0.102	**0.014**
8	0.130	0.110	0.066
9	0.114	1.000	0.270
10	0.492	1.000	**0.028**
11	0.796	1.000	0.682
12	0.798	1.000	0.216
13	0.402	0.414	0.230
14	0.376	1.000	0.580
15	0.334	1.000	0.150
16	0.830	1.000	0.332
17	0.518	0.344	0.072
18	0.788	0.270	0.576
19	0.456	0.108	**0.004**
20	0.918	0.390	0.200
21	0.178	0.408	**0.004**
22	0.782	1.000	0.346
23	0.400	**0.042**	0.678
24	0.798	1.000	0.566
25	**0.012**	1.000	0.098
26	0.434	1.000	0.224
27	0.556	**0.010**	**0.016**
28	0.610	0.246	0.098
29	0.338	**0.022**	0.094
30	0.174	1.000	0.066
31	0.282	0.312	0.142
32	0.608	1.000	0.064
33	0.058	0.058	0.090
34	0.456	0.508	0.066
35	0.072	0.140	**0.000**
36	0.912	1.000	0.526
37	0.364	0.212	0.286
38	0.748	0.418	0.236
39	0.760	0.118	0.074
40	0.636	1.000	1.000
41	0.082	**0.000**	**0.018**

42	0.668	0.438	0.220
43	0.646	0.442	0.418
44	0.776	0.430	0.792
45	0.724	0.294	0.476
46	0.430	0.548	0.400
47	0.470	0.628	0.152
48	0.072	**0.002**	**0.000**
49	0.594	1.000	1.000
50	0.970	1.000	0.922
51	0.944	1.000	0.452
52	0.272	1.000	0.114
53	0.694	1.000	0.276
54	0.638	1.000	0.598
55	0.410	0.090	0.300
56	0.902	1.000	0.768
57	0.588	1.000	0.306
58	0.388	0.726	0.214
59	0.352	1.000	0.306
60	0.932	1.000	0.630
61	0.850	0.354	0.114
62	0.746	1.000	**0.020**
63	0.626	1.000	0.204
64	0.504	0.666	0.614
65	0.308	0.098	0.526
66	0.180	1.000	0.068
67	0.964	1.000	0.548
68	0.376	1.000	**0.010**
69	0.252	**0.012**	**0.044**
70	0.918	1.000	0.684
71	0.476	1.000	0.742
72	0.472	1.000	0.358
73	0.408	0.518	0.208
74	0.814	1.000	0.730
75	0.842	1.000	0.550
76	0.338	1.000	0.066
77	0.444	1.000	1.000
78	0.654	1.000	0.648
79	0.782	0.418	0.172
80	0.310	0.170	0.098
81	0.850	0.114	**0.022**
82	0.408	1.000	0.192
83	0.368	1.000	0.658
84	**0.002**	**0.008**	**0.002**

Three items are identified with nonmonotonic IRFs by the nonparametric smooth regression method, seven items are identified with nonmonotonic IRFs by the item-ability regression method, and 13 items are identified with nonmonotonic IRFs by the B-splines nonparametric IRT method. Item 84 is identified with nonmonotonic IRFs by three methods. Item 41 and 48 are identified with nonmonotonic IRFs by two methods and the PPP value of the third method for these two items are close to 0.05 (0.082 of the nonparametric smooth regression method for item 41 and 0.072 of the nonparametric smooth regression method for item 48).

Table 10 presents the nonmonotonic area correlation among three nonparametric methods. The nonmonotonic area correlation between two nonparametric methods is the correlation of the nonmonotonic area of all the items on the math assessment B calculated by these two methods.

Table 10
Nonmonotonic Area Correlation among Three Nonparametric Methods for Math Assessment B

	Nonparametric Smooth Regression	Item-Ability Regression	B-Splines Nonparametric IRT
Nonparametric Smooth Regression		0.720	0.782
Item-Ability Regression			0.670
B-Splines Nonparametric IRT			

The nonmonotonic area correlation between the nonparametric smooth regression method and the B-splines nonparametric IRT method is the highest (0.784). In addition, the nonmonotonic area correlation between the nonparametric smooth regression method and the item-ability regression method (0.720) and the nonmonotonic area correlation between the B-splines nonparametric IRT method and the item-ability regression method (0.670) are not very high and not very similar. Three nonmonotonic correlations of the math assessment B are not as high as three nonmonotonic correlations of the math assessment A, which indicate that the nonmonotonic area patterns among items on the math assessment B estimated by these three methods are not very similar. For example, there is only one item identified with nonmonotonic IRF by three methods out of 16 items identified with nonmonotonic IRFs by at least one method.

Figures 28-43 present IRFs estimated by three nonparametric methods of items identified with nonmonotonic IRFs by at least one method. Figure 28 presents the IRFs of item 84 which is identified with nonmonotonic IRF by three nonparametric methods.

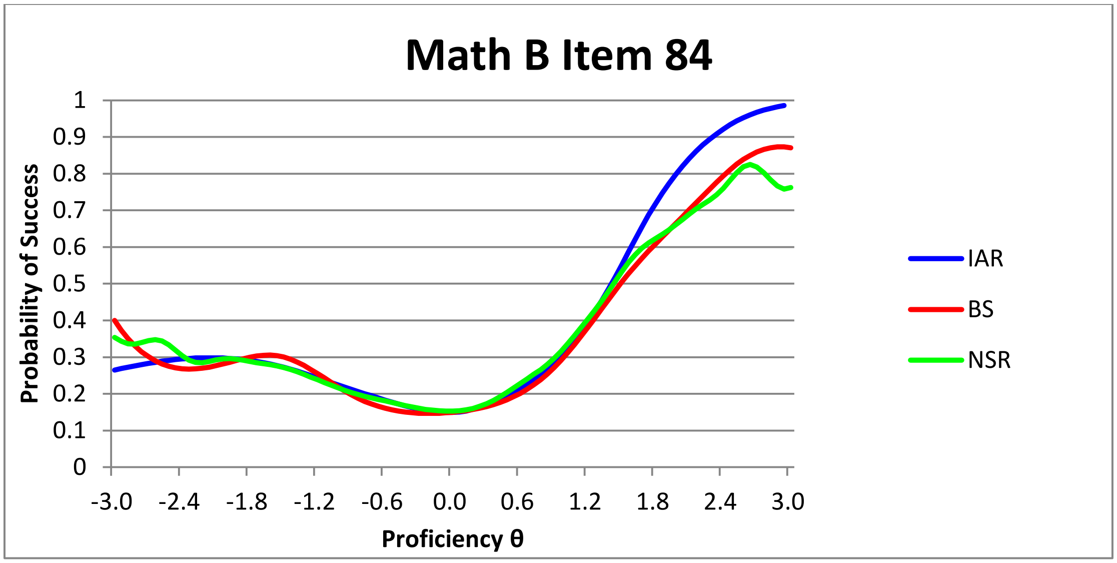

*Figure 28.*IRFs estimated by three nonparametric methods for item 84 on math B. The PPP-value of the item-ability regression method is 0.008. The PPP-value of the B-splines nonparametric IRT method is 0.002. The PPP-value of the nonparametric smooth regression method is 0.002.

For item 84, the IRFs of three nonparametric methods are very similar with the exception of the end parts at the very low and high theta range and the largest nonmonotonic parts of three IRFs appear at the same theta range. There are more curvatures in the IRFs of the nonparametric smooth regression method and the B-splines nonparametric IRT method at the low and high theta range than the IRF of the item-ability regression method.

Figures 29 and 30 present the IRFs for items 41 and 48 which are identified with nonmonotonic IRFs by two methods and the PPP values of the third method for these two items are close to 0.05.

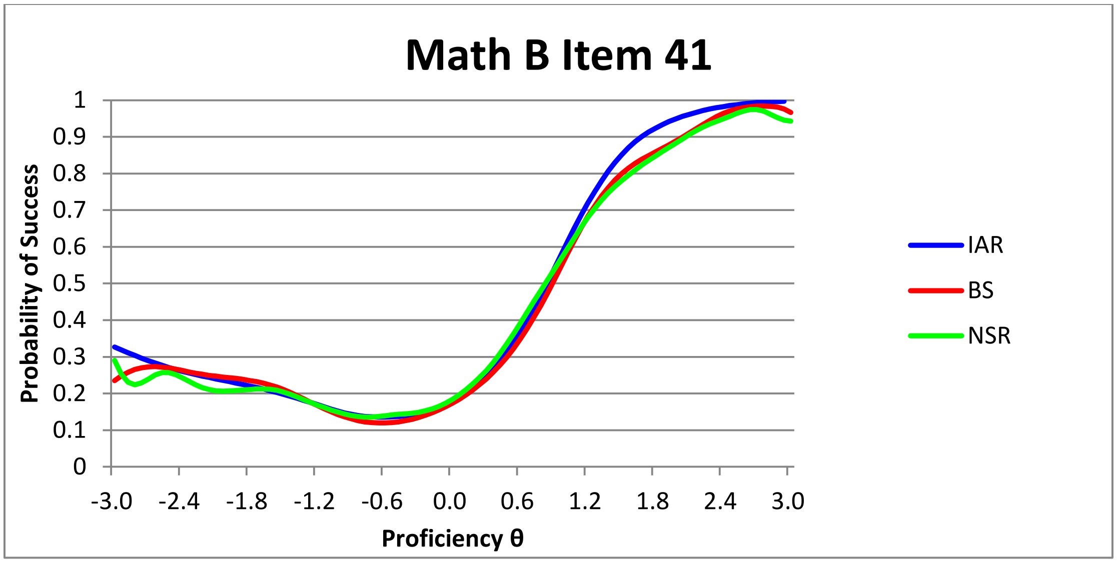

*Figure 29.*IRFs estimated by three nonparametric methods for item 41 on math B. The PPP-value of the item-ability regression method is 0.000. The PPP-value of the B-splines nonparametric IRT method is 0.018. The PPP-value of the nonparametric smooth regression method is 0.082.

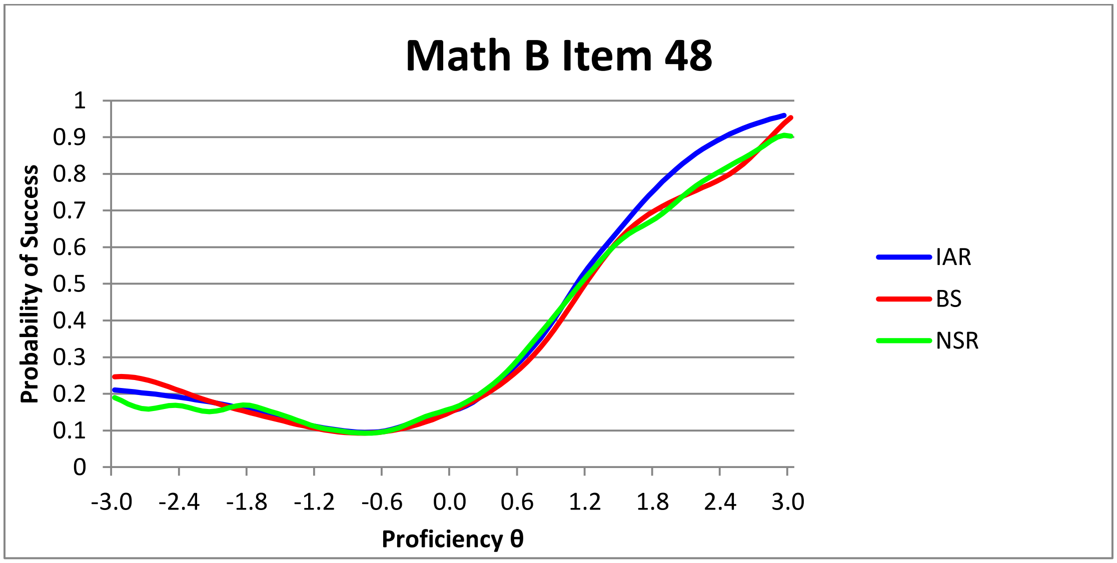

*Figure 30.*IRFs estimated by three nonparametric methods for item 48 on math B. The PPP-value of the item-ability regression method is 0.002. The PPP-value of the B-splines nonparametric IRT method is 0.000. The PPP-value of the nonparametric smooth regression method is 0.072.

For these two items, the differences among the IRFs of three nonparametric methods are small and at the end parts at the very low and high theta range. The nonmonotonic parts of three IRFs appear at the similar theta range.

Figures 31-43 present the IRFs of the other items identified with nonmonotonic IRFs by at least one method. These items are 7, 10, 19, 21, 23, 25, 27, 29, 35, 62, 68, 69, and 81.

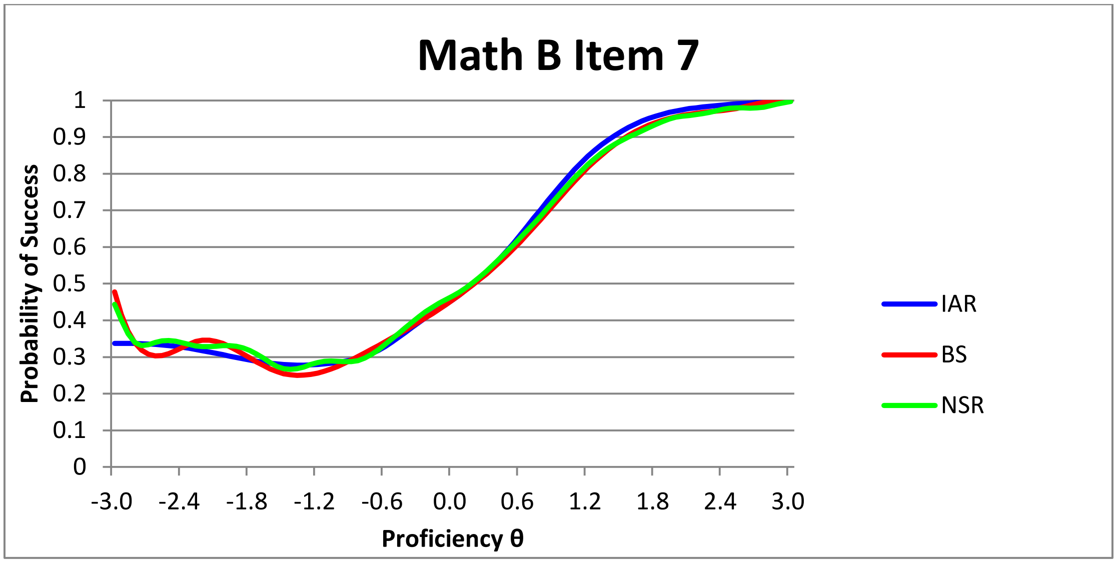

Figure 31.IRFs estimated by three nonparametric methods for item 7 on math B. The PPP-value of the item-ability regression method is 0.102. The PPP-value of the B-splines nonparametric IRT method is 0.014. The PPP-value of the nonparametric smooth regression method is 0.040.

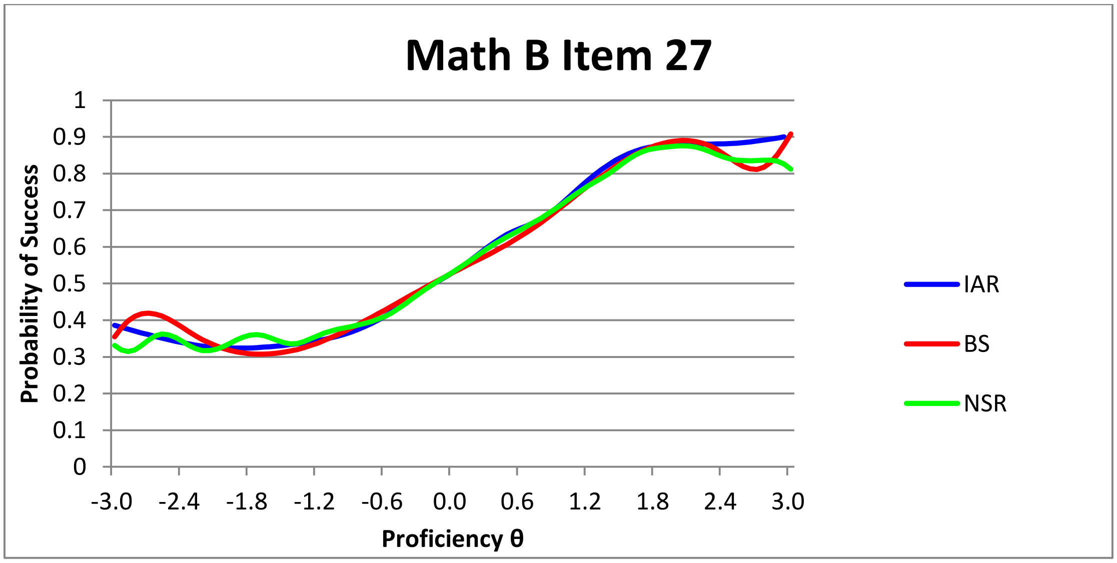

Figure 32.IRFs estimated by three nonparametric methods for item 27 on math B. The PPP-value of the item-ability regression method is 0.010. The PPP-value of the B-splines nonparametric IRT method is 0.016. The PPP-value of the nonparametric smooth regression method is 0.556.

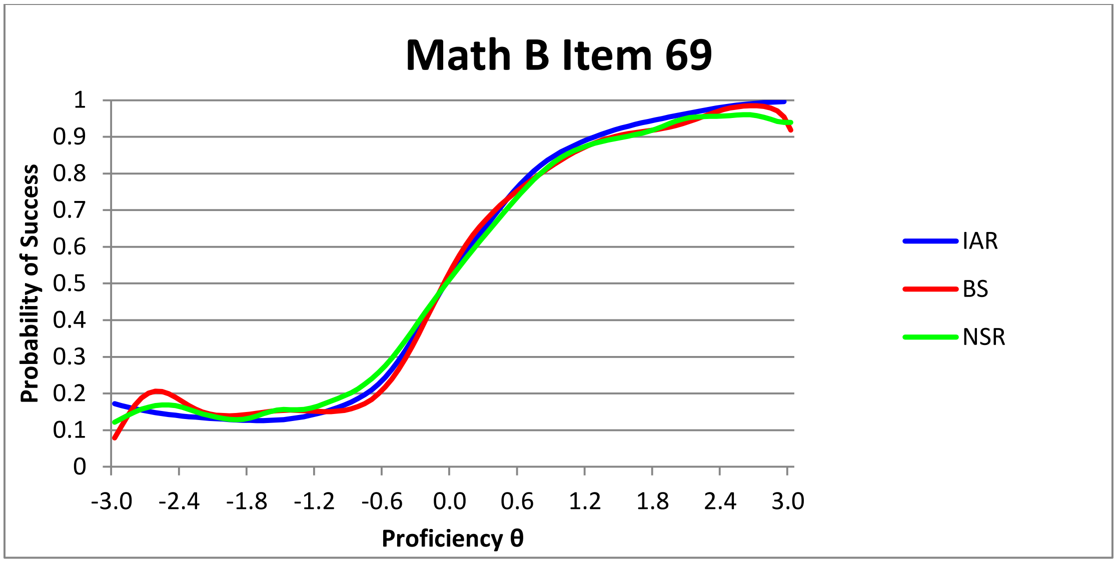

*Figure 33.*IRFs estimated by three nonparametric methods for item 69 on math B. The PPP-value of the item-ability regression method is 0.012. The PPP-value of the B-splines nonparametric IRT method is 0.044. The PPP-value of the nonparametric smooth regression method is 0.252.

For items 7, 27 and 69, the IRFs of three nonparametric methods are very similar with the exception of the end parts at the very low and high theta range and the largest nonmonotonic parts of three IRFs appear at the same theta range, but the nonmonotonicity is not obvious. There are more curvatures in the IRFs of the nonparametric smooth regression method and the B-splines nonparametric IRT method at the very low and high theta range.

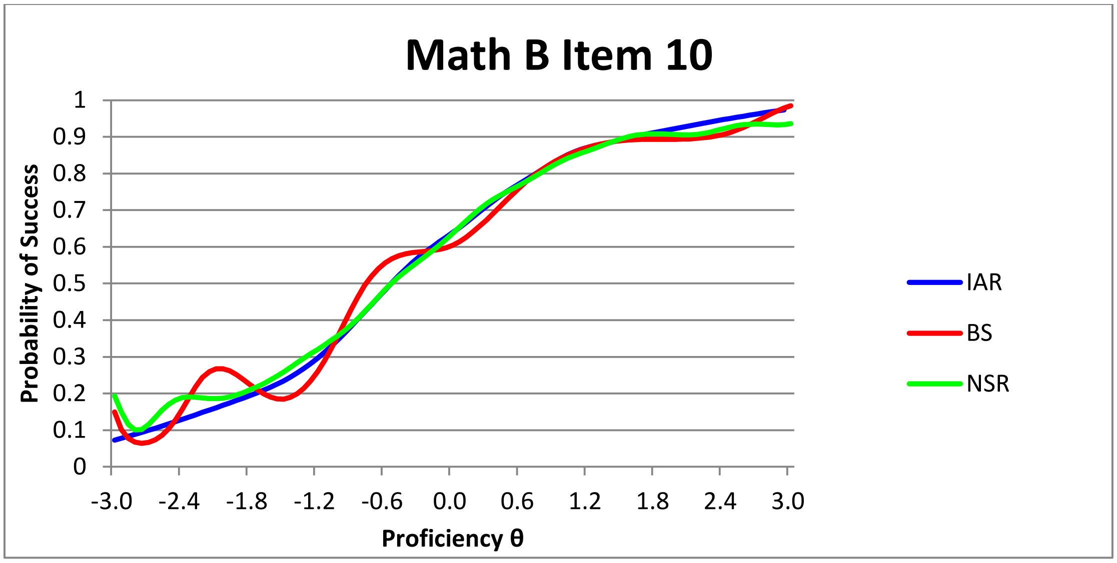

*Figure 34.*IRFs estimated by three nonparametric methods for item 10 on math B. The PPP-value of the item-ability regression method is 1.000. The PPP-value of the B-splines nonparametric IRT method is 0.028. The PPP-value of the nonparametric smooth regression method is 0.492.

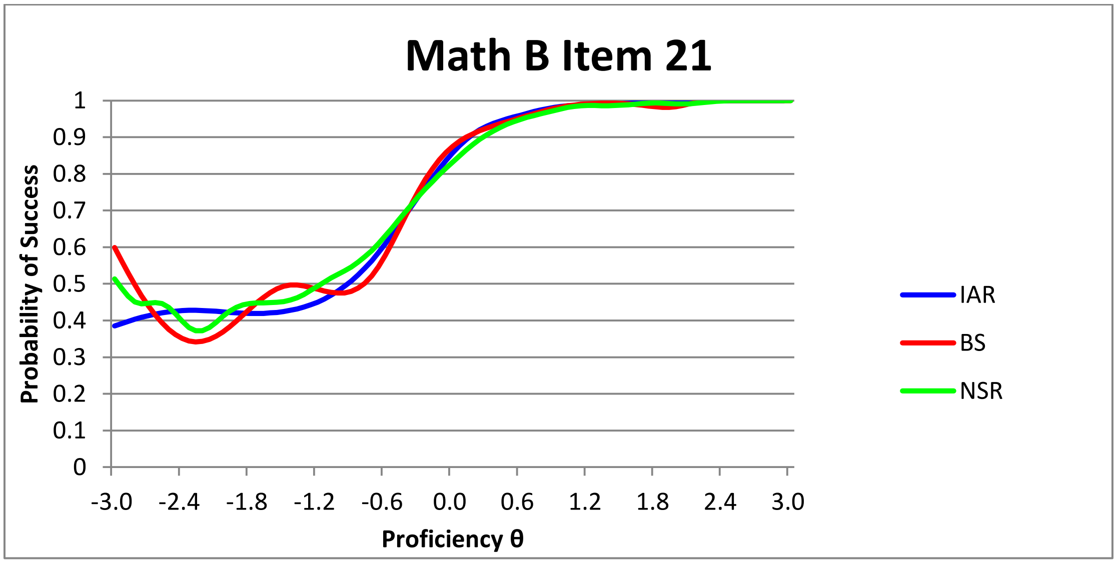

*Figure 35.*IRFs estimated by three nonparametric methods for item 21 on math B. The PPP-value of the item-ability regression method is 0.408. The PPP-value of the B-splines nonparametric IRT method is 0.004. The PPP-value of the nonparametric smooth regression method is 0.178.

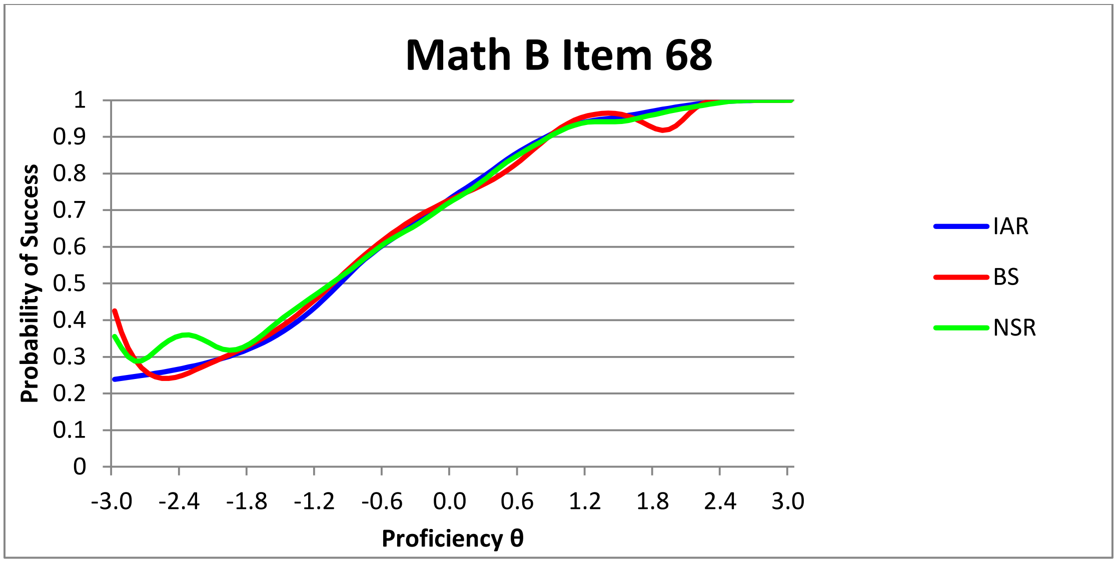

*Figure 36.*IRFs estimated by three nonparametric methods for item 68 on math B. The PPP-value of the item-ability regression method is 1.000. The PPP-value of the B-splines nonparametric IRT method is 0.010. The PPP-value of the nonparametric smooth regression method is 0.376.

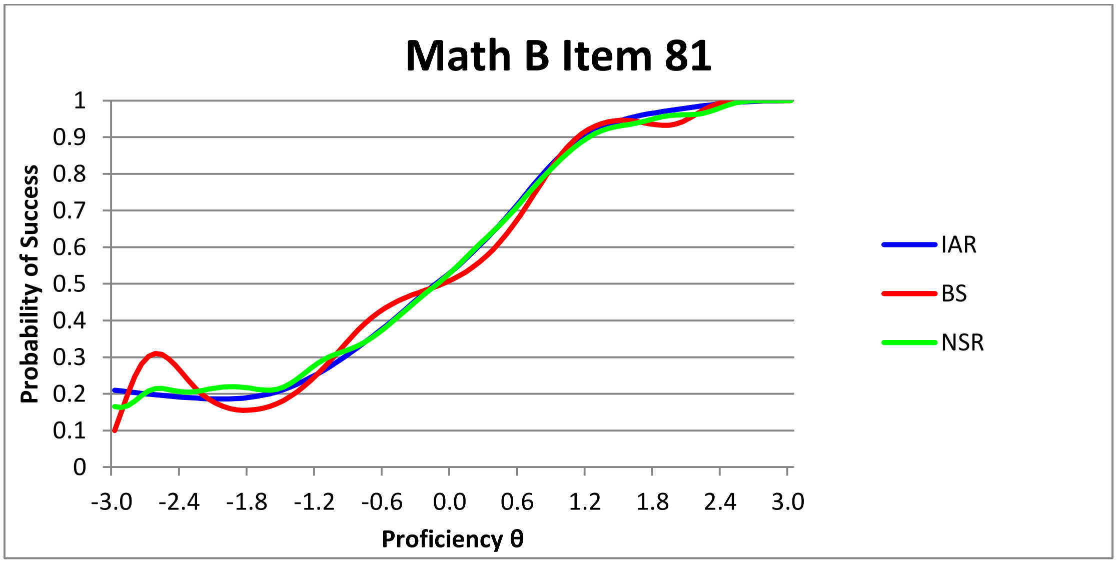

*Figure 37.*IRFs estimated by three nonparametric methods for item 81 on math B. The PPP-value of the item-ability regression method is 0.114. The PPP-value of the B-splines nonparametric IRT method is 0.022. The PPP-value of the nonparametric smooth regression method is 0.850.

For items 10, 21, 68, and 81, the IRFs of the nonparametric smooth regression method and the item-ability regression method are very similar but are different from the IRF of the B-splines nonparametric IRT method. The IRFs of the B-splines nonparametric IRT methods for these items have curvatures which lead to the low PPP-value.

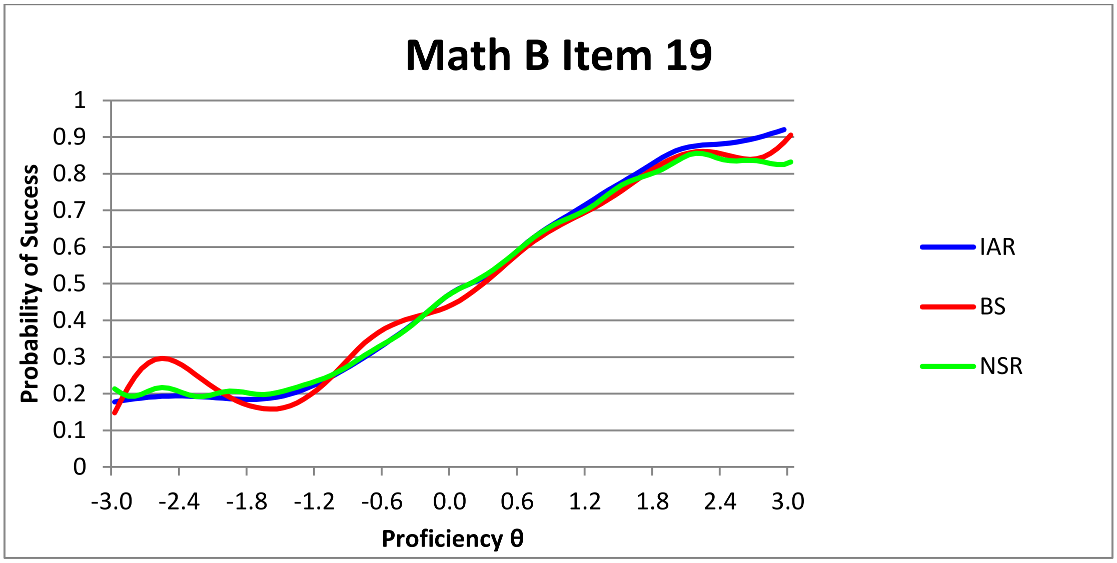

*Figure 38.*IRFs estimated by three nonparametric methods for item 19 on math B. The PPP-value of the item-ability regression method is 0.108. The PPP-value of the B-splines nonparametric IRT method is 0.004. The PPP-value of the nonparametric smooth regression method is 0.456.

For item 19, the IRFs of the nonparametric smooth regression method and the item-ability regression method are very similar but are different from the IRF of the B-splines nonparametric IRT method at the low theta range which is the range where nonmonotonicity of the IRF of the B-splines nonparametric IRT method appears.

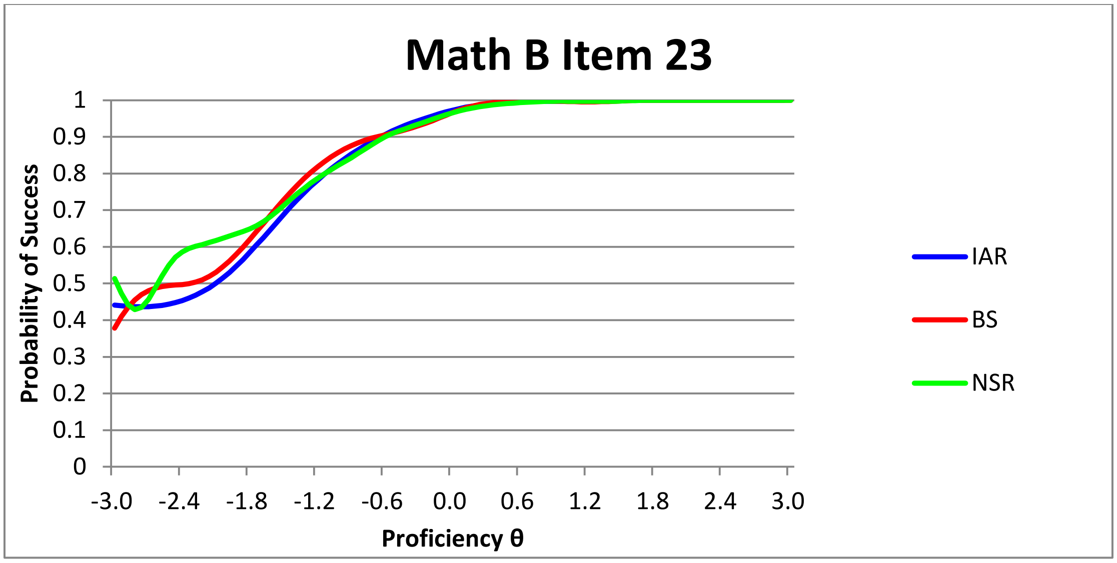

*Figure 39.*IRFs estimated by three nonparametric methods for item 23 on math B. The PPP-value of the item-ability regression method is 0.042. The PPP-value of the B-splines nonparametric IRT method is 0.678. The PPP-value of the nonparametric smooth regression method is 0.400.

For item 23, the IRFs of three nonparametric methods are similar except at the theta range from -3.0 to -0.6. The nonmonotonic parts of the IRFs of the item-ability regression method is not obvious. There are more curvatures in the IRF of the nonparametric smooth regression method at the low theta range.

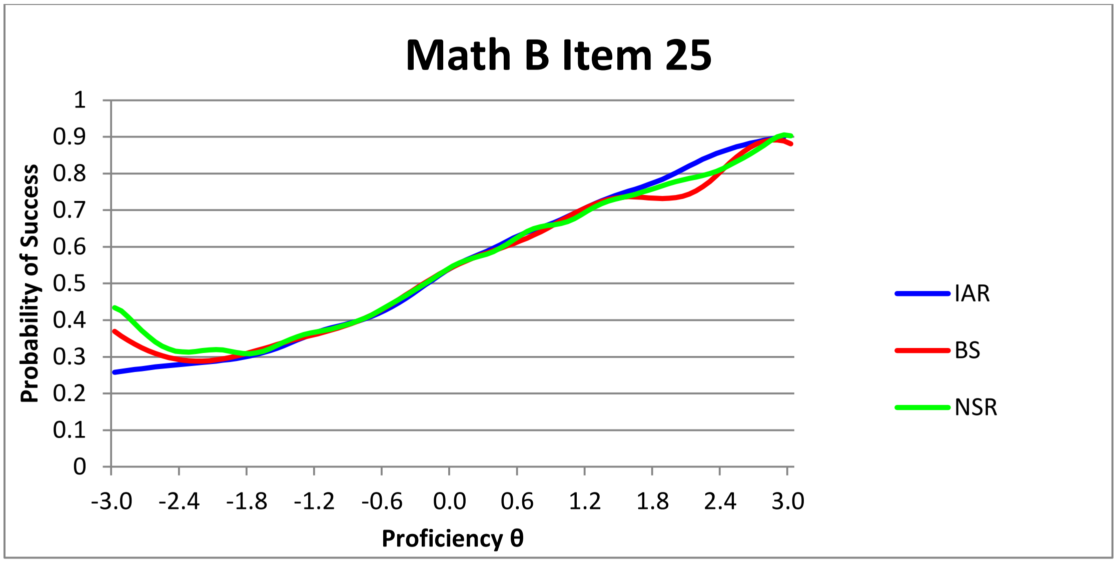

*Figure 40.*IRFs estimated by three nonparametric methods for item 25 on math B. The PPP-value of the item-ability regression method is 1.000. The PPP-value of the B-splines nonparametric IRT method is 0.098. The PPP-value of the nonparametric smooth regression method is 0.012.

For item 25, the IRFs of the nonparametric smooth regression method and the B-splines nonparametric IRT method are very similar but are different from the IRF of item-ability regression method at the very low and high theta range.

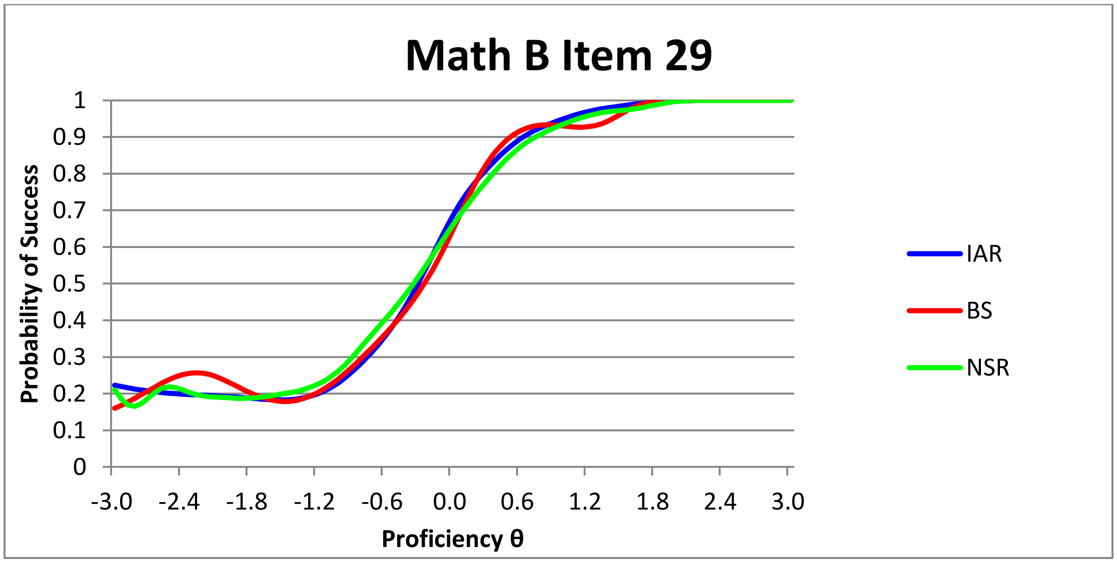

*Figure 41.*IRFs estimated by three nonparametric methods for item 29 on math B. The PPP-value of the item-ability regression method is 0.022. The PPP-value of the B-splines nonparametric IRT method is 0.094. The PPP-value of the nonparametric smooth regression method is 0.338.

For item 29, the IRFs of the nonparametric smooth regression method and the item-ability regression method are very similar but are different from the IRF of the B-splines nonparametric IRT method. The nonmonotonic parts of the IRFs of the nonparametric smooth regression method and the item-ability regression method are not obvious. The IRF of the B-splines nonparametric IRT method has some curvatures.

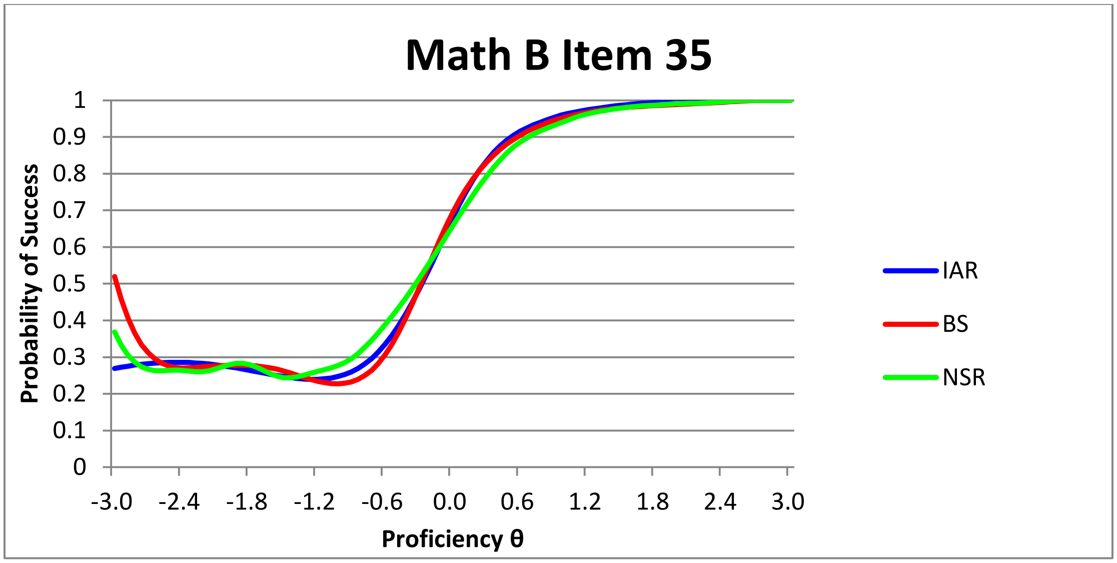

*Figure 42.*IRFs estimated by three nonparametric methods for item 35 on math B. The PPP-value of the item-ability regression method is 0.140. The PPP-value of the B-splines nonparametric IRT method is 0.000. The PPP-value of the nonparametric smooth regression method is 0.072.

For item 35, the IRFs of the nonparametric smooth regression method and the B-splines nonparametric IRT method are very similar but are different from the IRF of the item-ability regression method at the very low theta range.

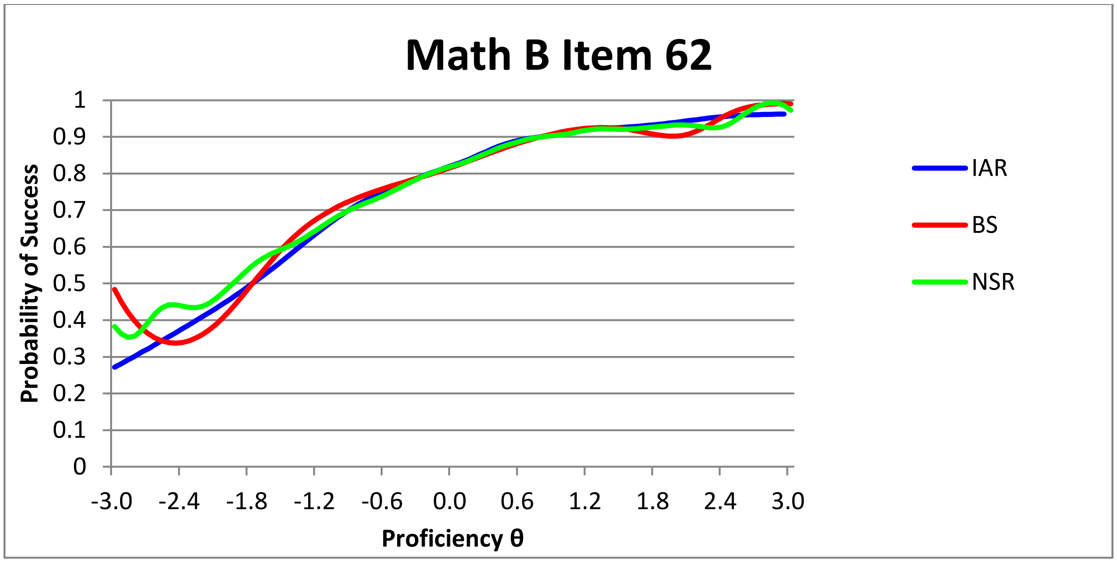

*Figure 43.*IRFs estimated by three nonparametric methods for item 62 on math B. The PPP-value of the item-ability regression method is 1.000. The PPP-value of the B-splines nonparametric IRT method is 0.020. The PPP-value of the nonparametric smooth regression method is 0.746.

For item 25, the IRFs of three methods are similar except at the low theta range.

Math assessment C. Table 11 includes the PPP-values of three nonparametric methods for each item on this assessment.

Table 11

PPP-Value of Three Nonparametric Methods for Each Item on Math Assessment C

Item	Nonparametric Smooth Regression	Item-Ability Regression	B-Splines Nonparametric IRT
1	0.970	1.000	1.000
2	**0.012**	**0.000**	**0.000**
3	0.694	1.000	1.000
4	0.378	**0.042**	0.336
5	0.906	1.000	0.844
6	0.848	0.188	0.468
7	**0.044**	**0.002**	**0.002**
8	0.990	0.580	0.766
9	**0.034**	**0.008**	0.102
10	0.572	**0.004**	0.744
11	0.190	1.000	0.882
12	0.736	0.200	0.820
13	0.200	1.000	0.346
14	0.970	1.000	0.902
15	0.696	1.000	0.710
16	0.736	1.000	0.418
17	0.964	0.074	0.744
18	0.236	0.096	0.096
19	0.864	**0.022**	0.766
20	**0.022**	0.182	**0.046**
21	0.568	0.172	0.764
22	0.176	1.000	0.322
23	0.380	0.056	0.444
24	**0.046**	**0.004**	**0.030**
25	**0.038**	**0.000**	**0.028**
26	0.068	1.000	0.076
27	0.614	**0.016**	0.292
28	0.552	1.000	0.628
29	**0.018**	**0.006**	**0.008**
30	0.368	**0.008**	0.172
31	**0.022**	1.000	0.150
32	0.784	1.000	0.632
33	0.212	0.368	0.182
34	**0.000**	**0.000**	**0.004**
35	0.428	1.000	0.406
36	0.362	1.000	0.266
37	0.278	1.000	0.350
38	0.964	1.000	0.762
39	0.964	1.000	0.760
40	0.074	0.066	0.256
41	0.570	1.000	0.684

42	0.346	0.336	0.628
43	0.182	0.686	0.690
44	0.336	**0.000**	0.074
45	0.446	0.566	0.734
46	0.324	0.274	0.392
47	0.558	1.000	0.514
48	0.718	**0.042**	0.718
49	**0.020**	**0.014**	0.234
50	0.076	**0.000**	**0.018**
51	0.722	1.000	0.862
52	0.414	1.000	0.498
53	**0.002**	**0.014**	**0.000**
54	0.506	**0.022**	0.414
55	0.100	0.070	0.256
56	**0.038**	**0.022**	0.090
57	0.824	**0.026**	0.540
58	0.992	1.000	0.832
59	0.912	1.000	0.662
60	0.166	**0.004**	0.056
61	0.148	1.000	0.074
62	0.572	0.208	0.552
63	0.224	**0.016**	0.206
64	0.756	1.000	1.000
65	0.790	1.000	0.656
66	0.438	1.000	0.596
67	0.788	1.000	0.708
68	0.086	0.194	0.260
69	0.528	1.000	0.374
70	0.692	1.000	0.538
71	0.768	0.052	0.678
72	0.228	1.000	0.532
73	0.200	**0.000**	0.154
74	0.164	**0.022**	0.192
75	0.264	0.064	0.568
76	0.056	0.508	0.080
77	0.110	1.000	0.402
78	0.298	1.000	0.486
79	0.712	1.000	0.746
80	**0.010**	**0.000**	**0.000**
81	**0.006**	0.230	**0.006**
82	0.346	1.000	0.568
83	0.422	0.386	0.776
84	0.652	1.000	0.708

Fourteen items are identified with nonmonotonic IRFs by the nonparametric smooth regression method, 25 items are identified with nonmonotonic IRFs by the item-ability regression method, and 11 items are identified with nonmonotonic IRFs by the B-splines nonparametric IRT method. Items 2, 7 24, 25, 29, 34, 53, and 80 are identified with nonmonotonic IRFs by three methods. Also item 50 is identified with nonmonotonic IRF by two methods and the PPP value of the third method is very close to 0.05 (0.076 of the nonparametric smooth regression method).

Table 12 presents the nonmonotonic area correlation among three nonparametric methods. The nonmonotonic area correlation between two nonparametric methods is the correlation of the nonmonotonic area of all items on the math assessment C calculated by these two methods.

Table 12
Nonmonotonic Area Correlation among Three Nonparametric Methods for Math Assessment C

	Nonparametric Smooth Regression	Item-Ability Regression	B-Splines Nonparametric IRT
Nonparametric Smooth Regression		0.614	0.789
Item-Ability Regression			0.816
B-Splines Nonparametric IRT			

The nonmonotonic area correlation between the item-ability regression method and the B-splines nonparametric IRT method is highest (0.816). Also, the nonmonotonic area correlation between the nonparametric smooth regression method and the B-splines nonparametric IRT method is high (0.789). But, the nonmonotonic area correlation between the nonparametric smooth regression method and the item-ability regression method (0.614) is not very high which indicates that the nonmonotonic area estimated by these two methods do not have a pattern among items on the math assessment C as similar as the nonmonotonic area estimated by the item-ability regression method and the B-splines nonparametric IRT method.

Figures 44-71 present IRFs estimated by three nonparametric methods of items identified with nonmonotonic IRFs by at least one method. Figures 44-51 present the IRFs of items2, 7, 24, 25, 29, 34, 53, and 80 which are identified with nonmonotonic IRFs by three nonparametric methods.

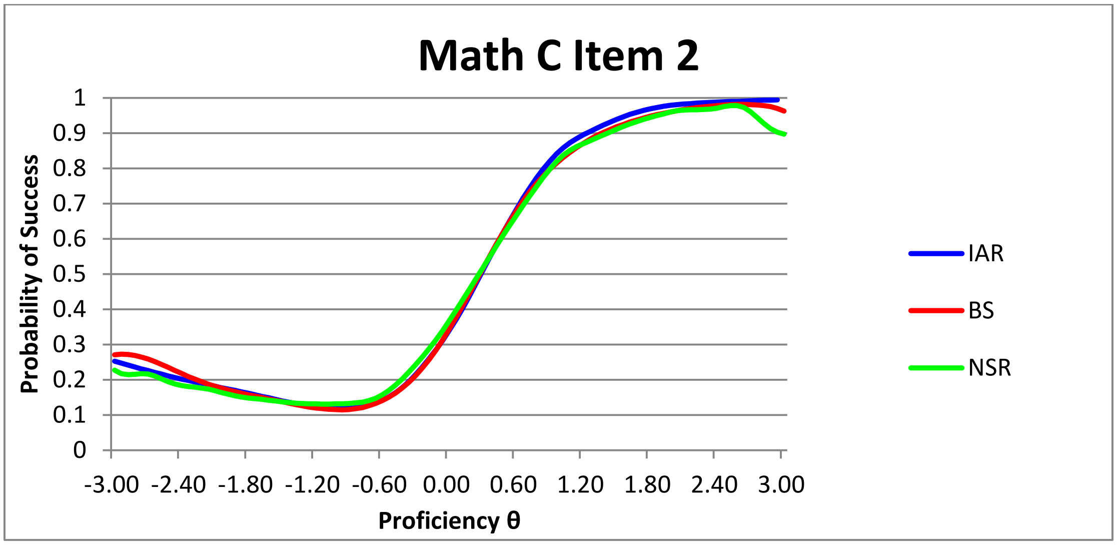

*Figure 44.*IRFs estimated by three nonparametric methods for item 2 on math C. The PPP-value of the item-ability regression method is 0.000. The PPP-value of the B-splines nonparametric IRT method is 0.000. The PPP-value of the nonparametric smooth regression method is 0.012.

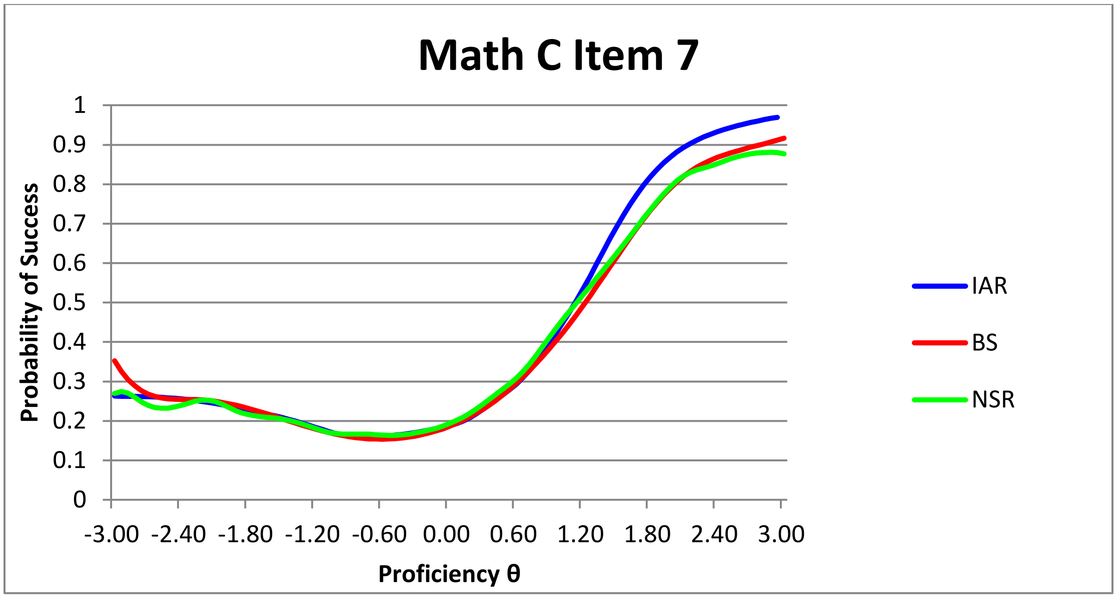

*Figure 45.*IRFs estimated by three nonparametric methods for item 7 on math C. The PPP-value of the item-ability regression method is 0.002. The PPP-value of the B-splines nonparametric IRT method is 0.002. The PPP-value of the nonparametric smooth regression method is 0.044.

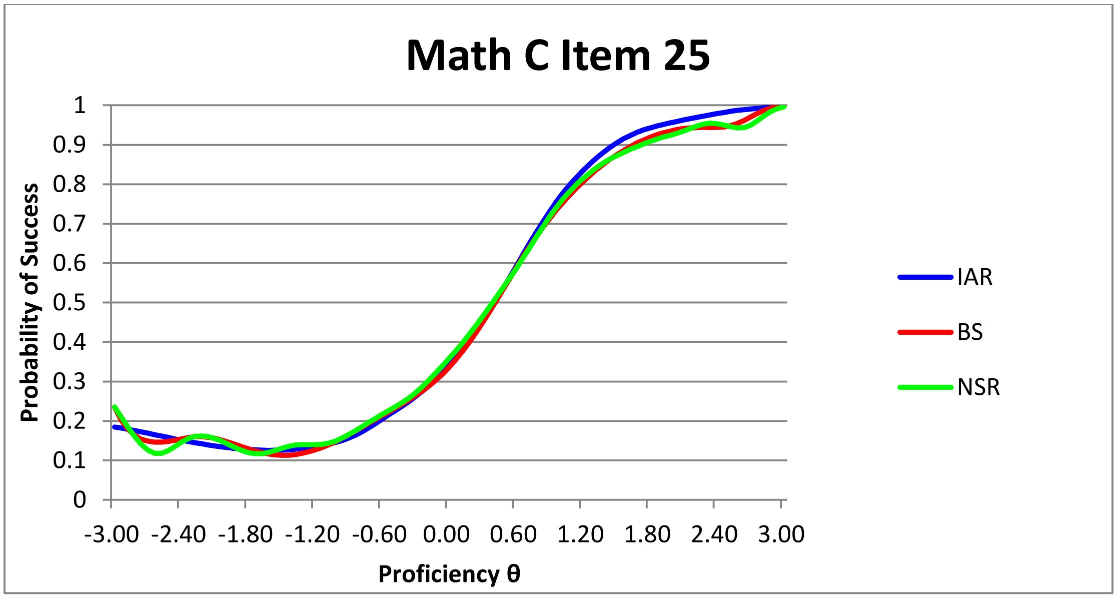

*Figure 46.*IRFs estimated by three nonparametric methods for item 25 on math C. The PPP-value of the item-ability regression method is 0.000. The PPP-value of the B-splines nonparametric IRT method is 0.028. The PPP-value of the nonparametric smooth regression method is 0.038.

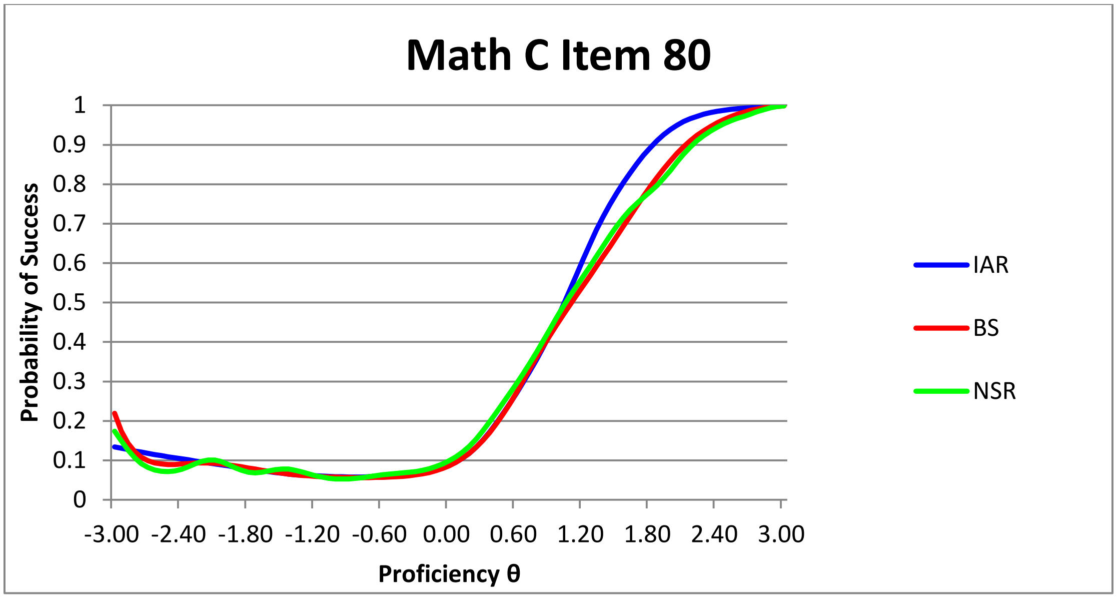

*Figure 47.*IRFs estimated by three nonparametric methods for item 80 on math C. The PPP-value of the item-ability regression method is 0.000. The PPP-value of the B-splines nonparametric IRT method is 0.000. The PPP-value of the nonparametric smooth regression method is 0.010.

For these four items, the IRFs of three nonparametric methods are very similar, especially the IRFs of the nonparametric smooth regression method and the B-splines nonparametric IRT method, with the exception of the ends parts at the very low and high theta range. The nonmonotonic parts of three IRFs appear at the same theta range.

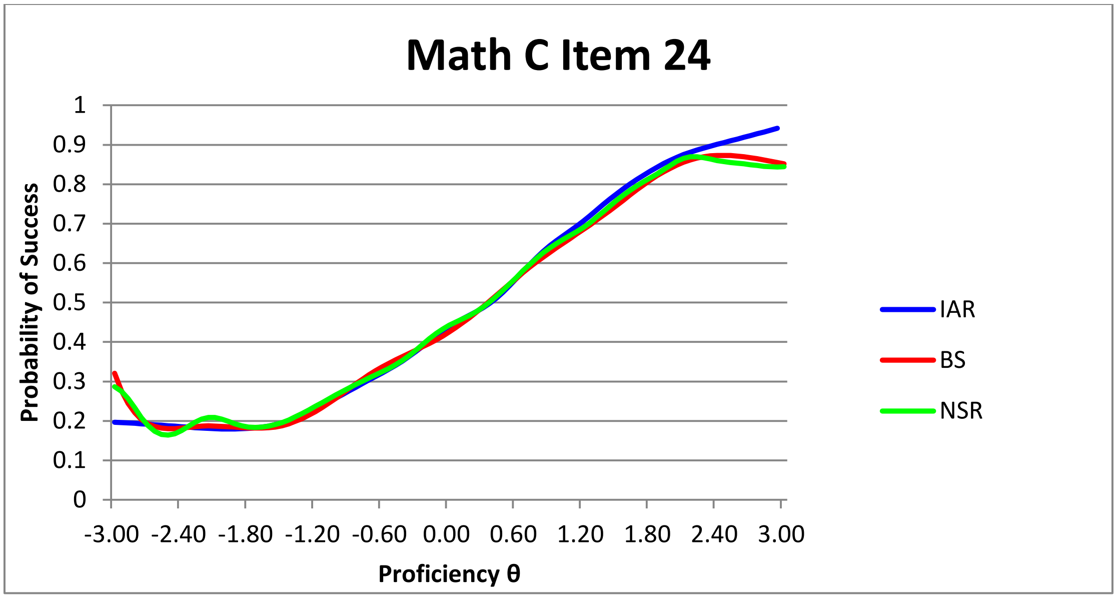

*Figure 48.*IRFs estimated by three nonparametric methods for item 24 on math C. The PPP-value of the item-ability regression method is 0.004. The PPP-value of the B-splines nonparametric IRT method is 0.030. The PPP-value of the nonparametric smooth regression method is 0.046.

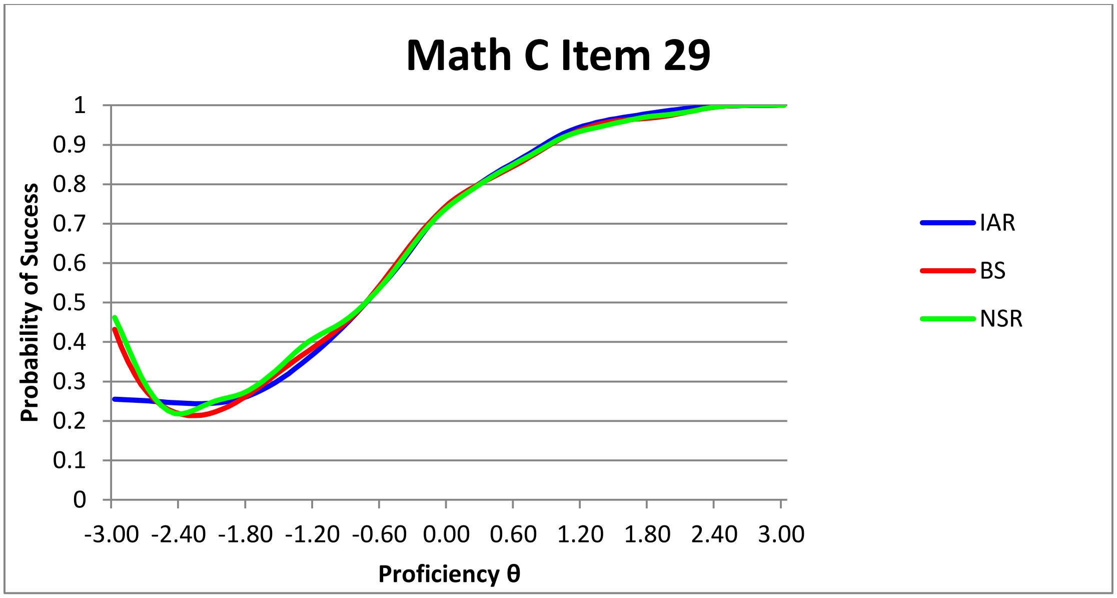

*Figure 49.*IRFs estimated by three nonparametric methods for item 29 on math C. The PPP-value of the item-ability regression method is 0.006. The PPP-value of the B-splines nonparametric IRT method is 0.008. The PPP-value of the nonparametric smooth regression method is 0.018.

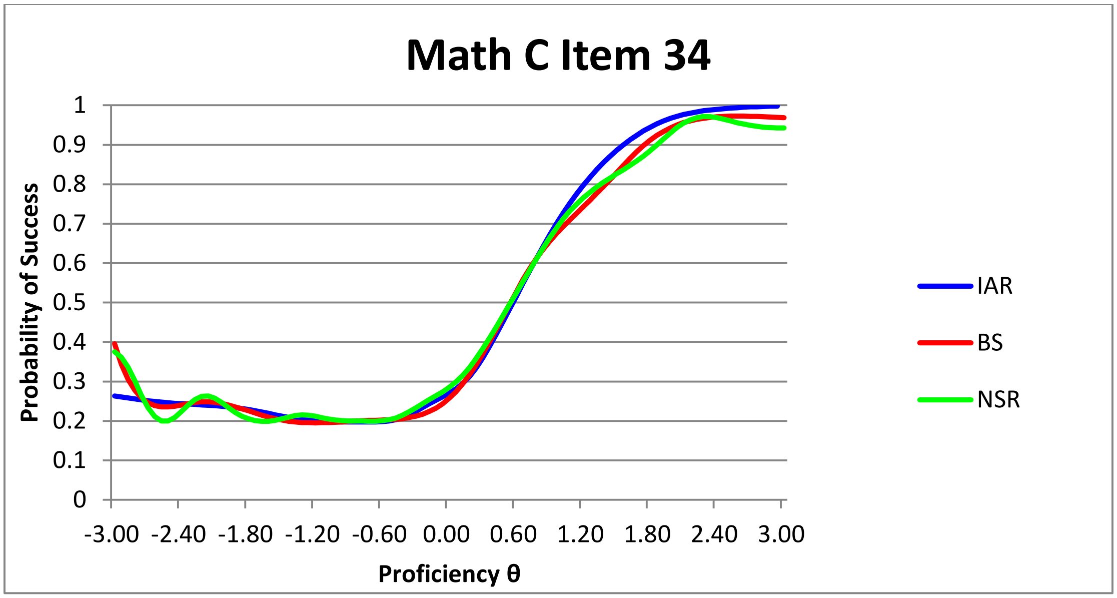

*Figure 50.*IRFs estimated by three nonparametric methods for item 34 on math C. The PPP-value of the item-ability regression method is 0.000. The PPP-value of the B-splines nonparametric IRT method is 0.004. The PPP-value of the nonparametric smooth regression method is 0.000.

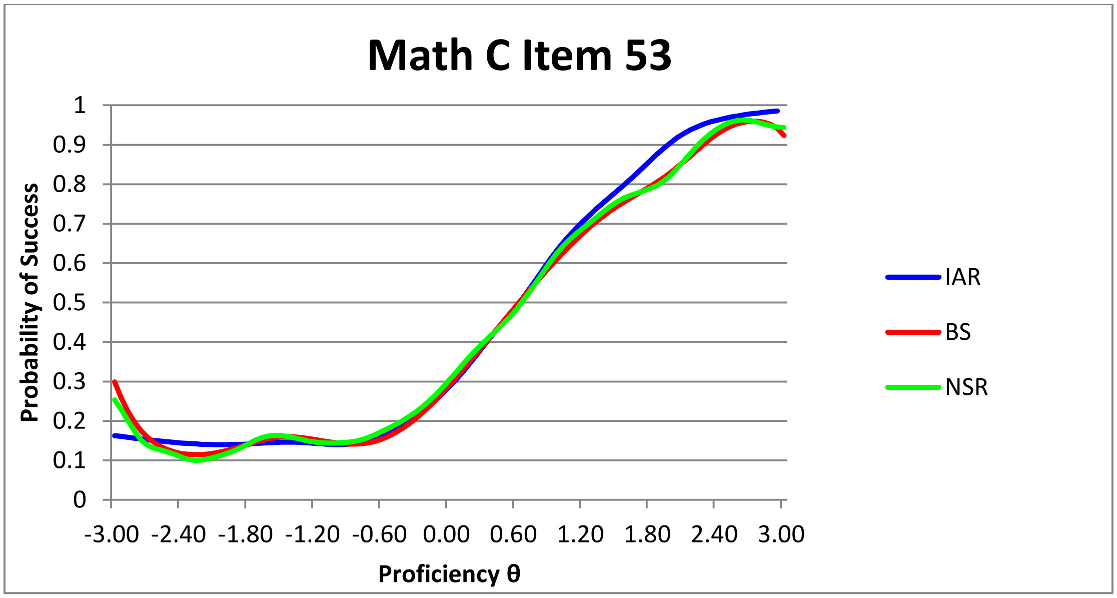

*Figure 51.*IRFs estimated by three nonparametric methods for item 53 on math C. The PPP-value of the item-ability regression method is 0.004. The PPP-value of the B-splines nonparametric IRT method is 0.030. The PPP-value of the nonparametric smooth regression method is 0.046.

For these four items, the IRFs of the nonparametric smooth regression method and the B-splines nonparametric IRT method are very similar but are different from the IRF of the item-ability regression method at the very low and high theta range. The nonmonotonic parts of three IRFs appear at the same theta range but the nonmonotonic part of the IRF of the item ability regression method is smaller than the nonmonotonic parts of the other two IRFs.

Figure 52 presents the IRFs for item 50 which is identified as an item with nonmonotonic IRF by two methods and the PPP value of the third method is very close to 0.05.

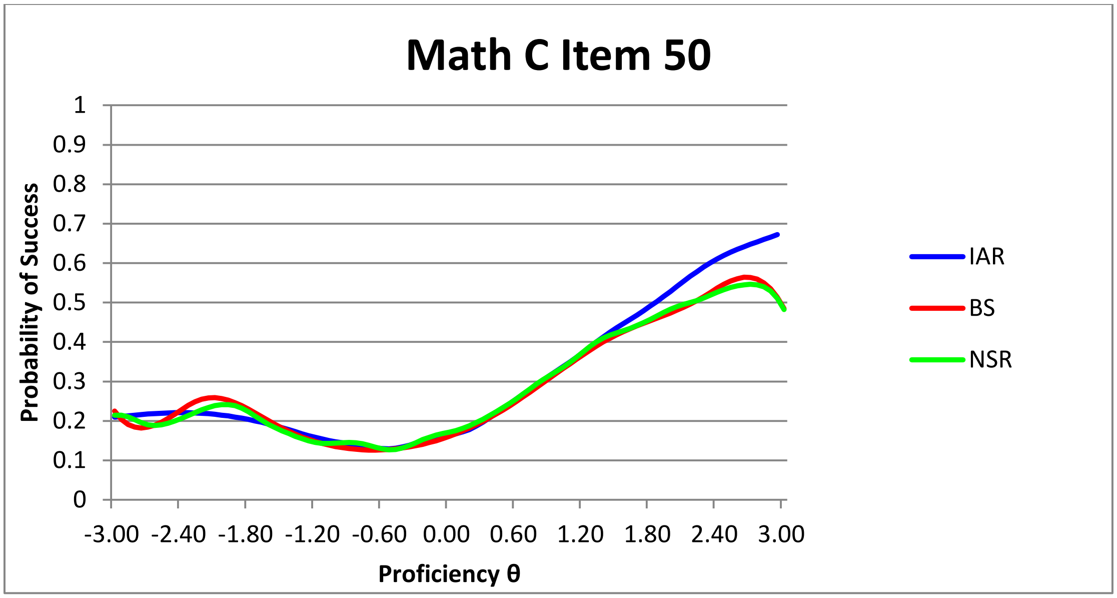

*Figure 52.*IRFs estimated by three nonparametric methods for item 50 on math C. The PPP-value of the item-ability regression method is 0.000. The PPP-value of the B-splines nonparametric IRT method is 0.018. The PPP-value of the nonparametric smooth regression method is 0.076.

For item 50, the IRFs of three nonparametric methods are very similar, especially the IRFs of the nonparametric smooth regression method and the B-splines nonparametric IRT method, with the exception of the end parts at the very low and high theta range. The nonmonotonic parts of three IRFs appear at the same theta range.

Figures 53-71 present the IRFs of the other items identified with nonmonotonic IRFs by at least one method. These items are 4, 9, 10, 19, 20, 27, 30, 31, 44, 48, 49, 54, 56, 57, 60, 63, 73, 74, and 81.

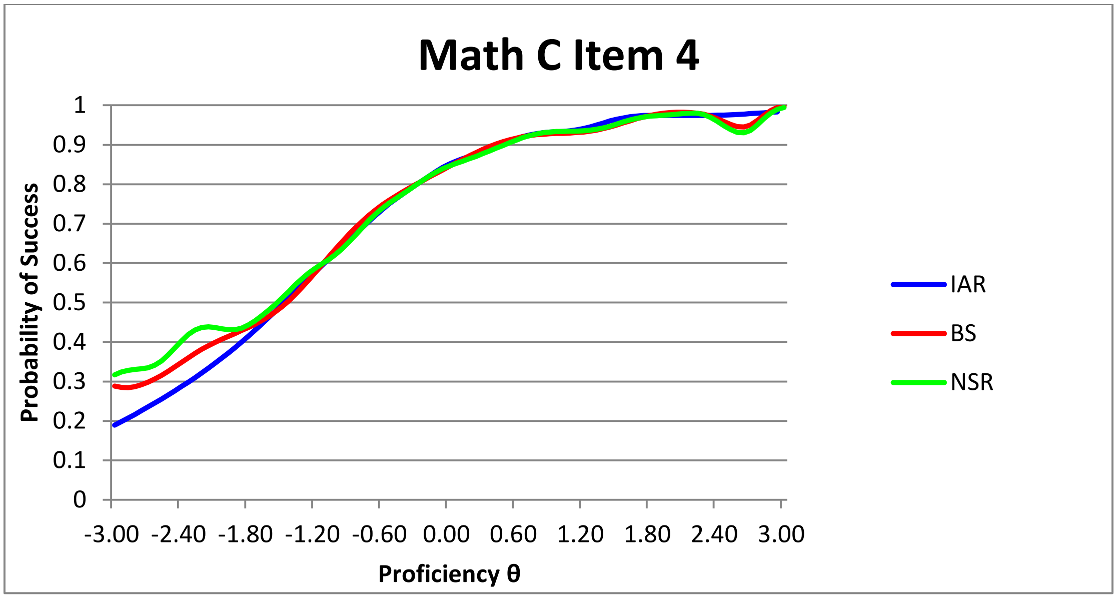

*Figure 53.*IRFs estimated by three nonparametric methods for item 4 on math C. The PPP-value of the item-ability regression method is 0.042. The PPP-value of the B-splines nonparametric IRT method is 0.336. The PPP-value of the nonparametric smooth regression method is 0.378.

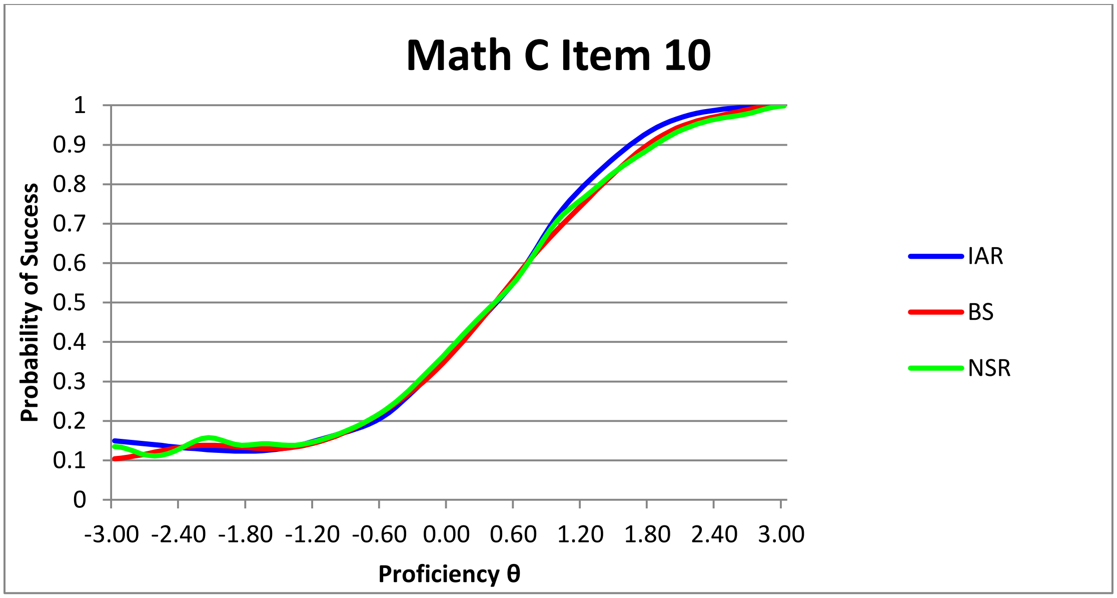

*Figure 54.*IRFs estimated by three nonparametric methods for item 10 on math C. The PPP-value of the item-ability regression method is 0.004. The PPP-value of the B-splines nonparametric IRT method is 0.744. The PPP-value of the nonparametric smooth regression method is 0.572.

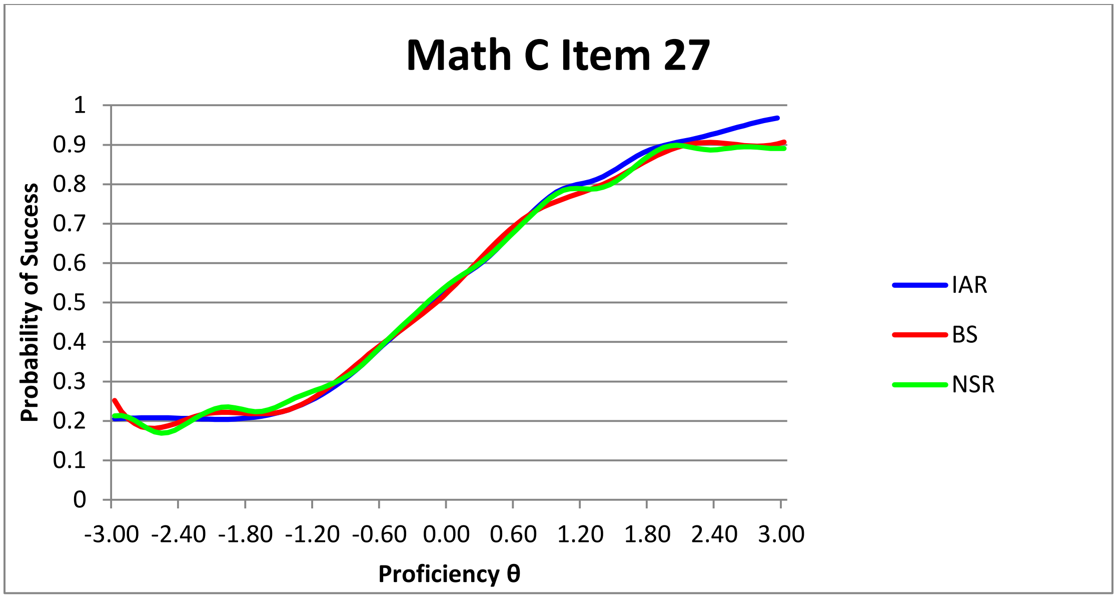

Figure 55.IRFs estimated by three nonparametric methods for item 27 on math C. The PPP-value of the item-ability regression method is 0.016. The PPP-value of the B-splines nonparametric IRT method is 0.292. The PPP-value of the nonparametric smooth regression method is 0.614.

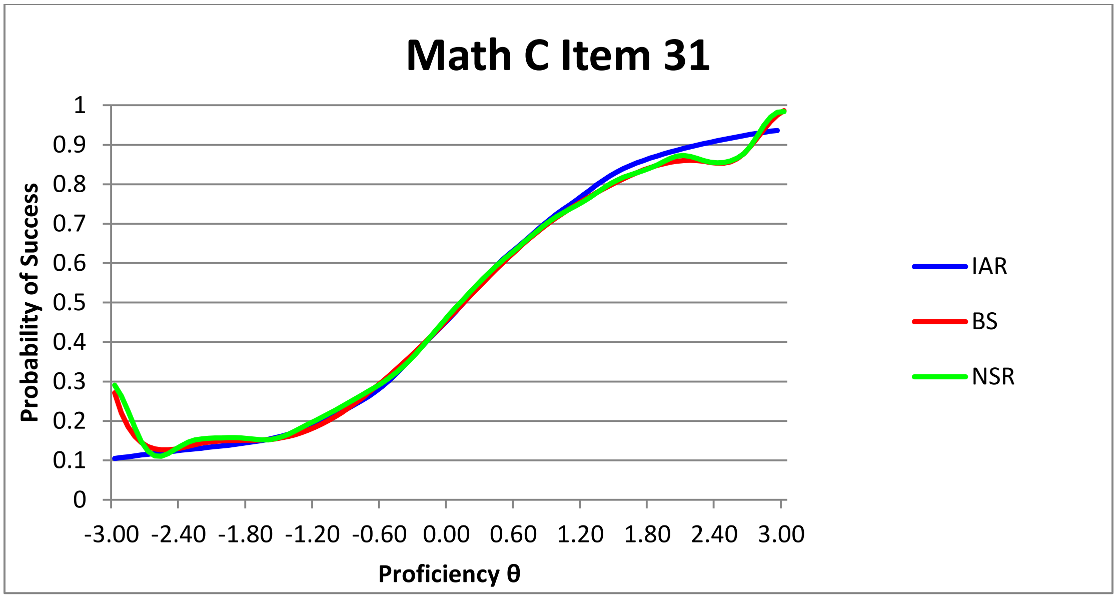

Figure 56.IRFs estimated by three nonparametric methods for item 31 on math C. The PPP-value of the item-ability regression method is 1.000. The PPP-value of the B-splines nonparametric IRT method is 0.150. The PPP-value of the nonparametric smooth regression method is 0.022.

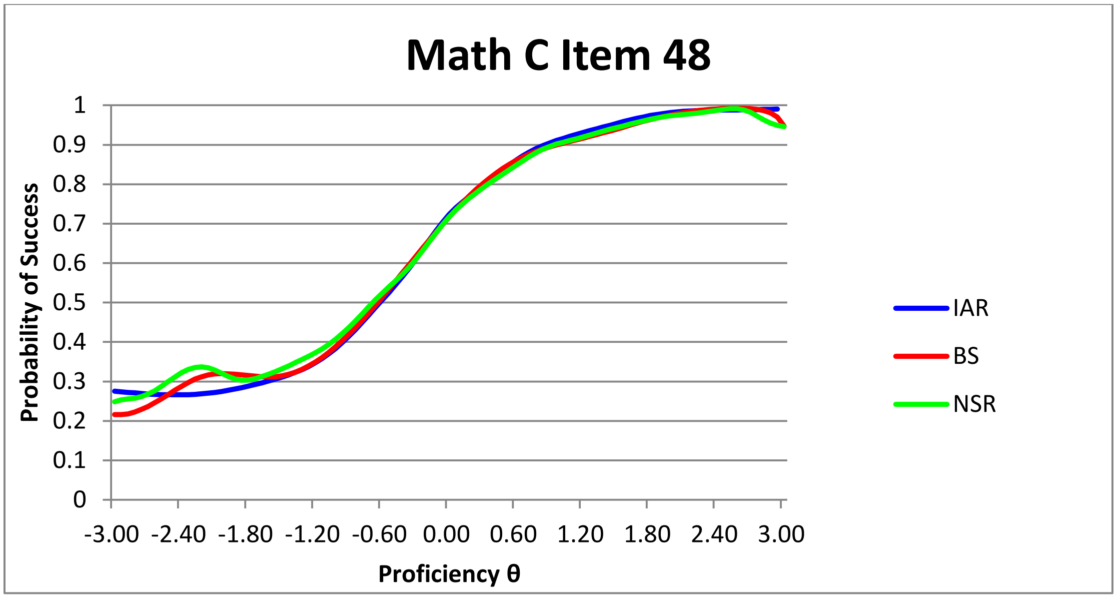

*Figure 57.*IRFs estimated by three nonparametric methods for item 48 on math C. The PPP-value of the item-ability regression method is 0.042. The PPP-value of the B-splines nonparametric IRT method is 0.718. The PPP-value of the nonparametric smooth regression method is 0.718.

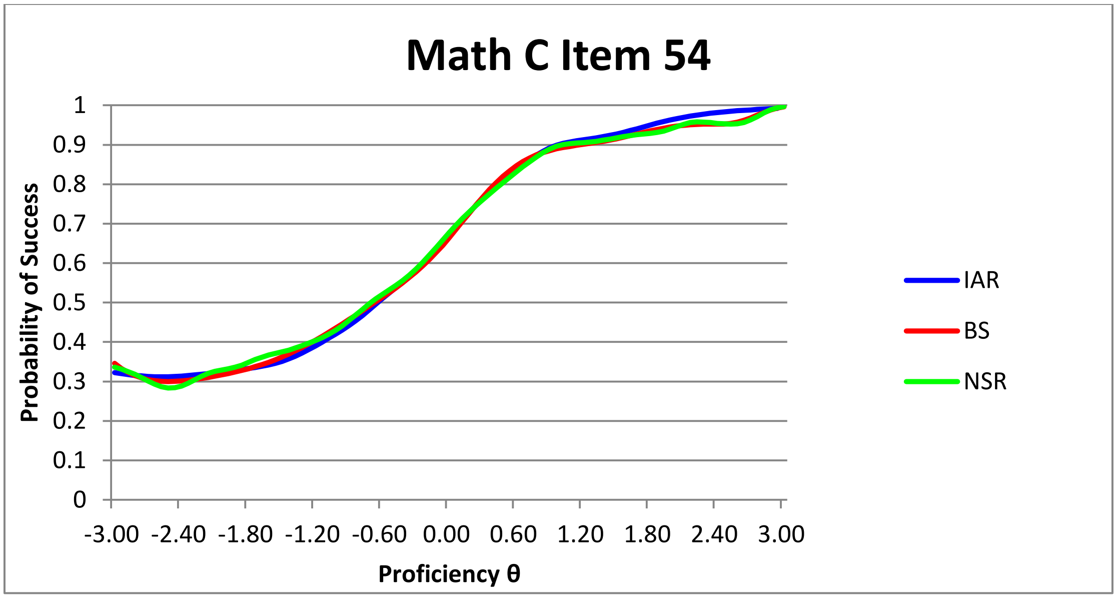

*Figure 58.*IRFs estimated by three nonparametric methods for item 54 on math C. The PPP-value of the item-ability regression method is 0.022. The PPP-value of the B-splines nonparametric IRT method is 0.414. The PPP-value of the nonparametric smooth regression method is 0.506.

For these items, the differences among the IRFs of three nonparametric methods are small. The IRFs of the B-splines nonparametric IRT method and the nonparametric smooth regression method are similar and have more curvatures. The nonmonotonic parts of these IRFs are not very obvious.

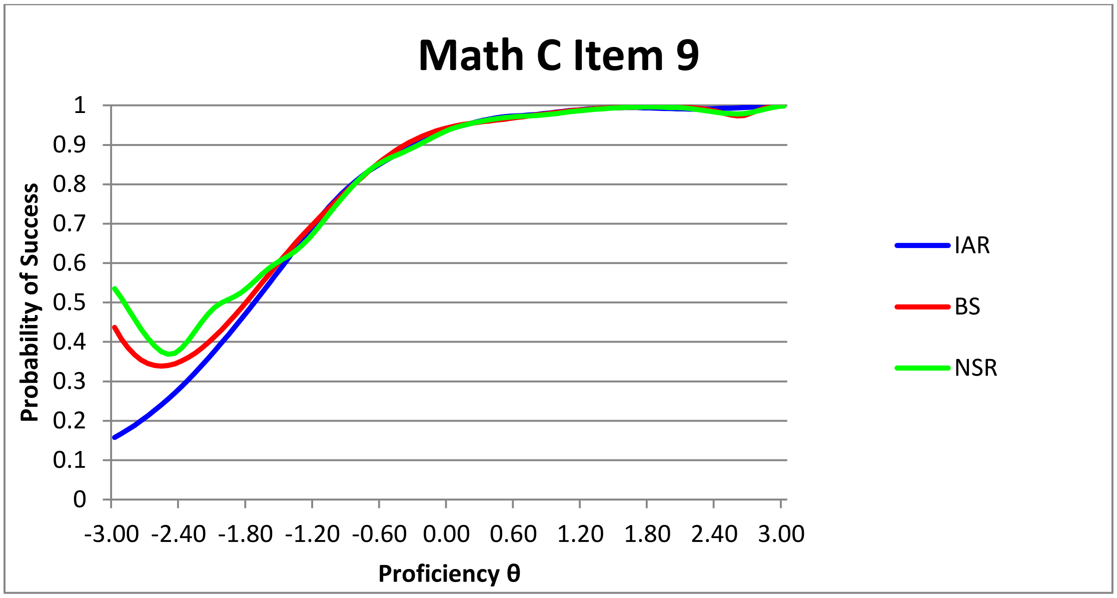

*Figure 59.*IRFs estimated by three nonparametric methods for item 9 on math C. The PPP-value of the item-ability regression method is 0.008. The PPP-value of the B-splines nonparametric IRT method is 0.102. The PPP-value of the nonparametric smooth regression method is 0.034.

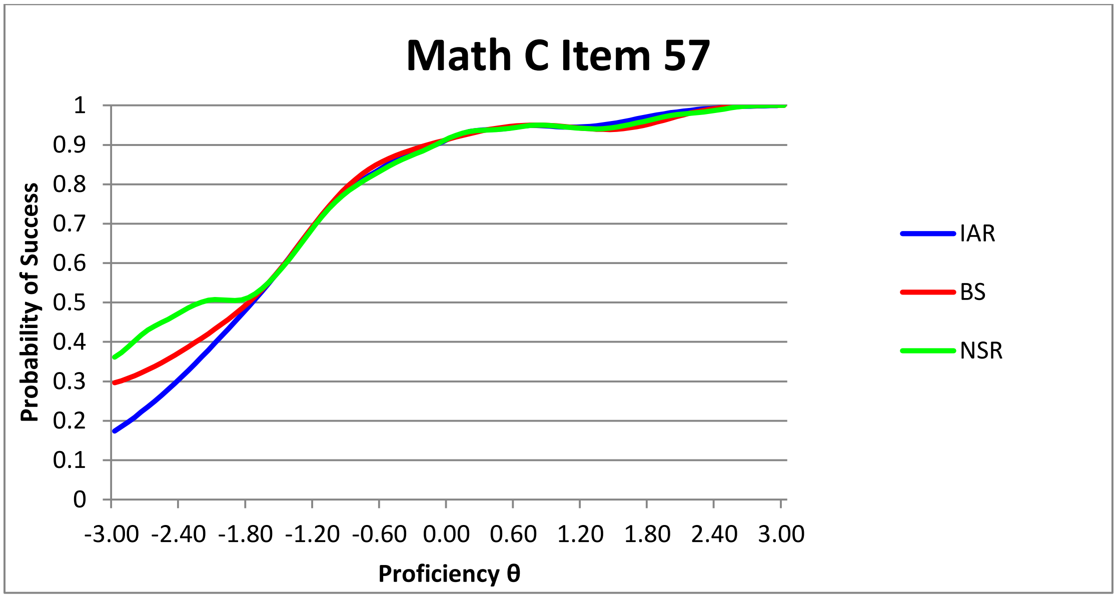

*Figure 60.*IRFs estimated by three nonparametric methods for item 57 on math C. The PPP-value of the item-ability regression method is 0.026. The PPP-value of the B-splines nonparametric IRT method is 0.540. The PPP-value of the nonparametric smooth regression method is 0.824.

For items 10 and 57, the differences among the IRFs of three nonparametric methods are at the low theta range. The IRFs of three nonparametric methods all have a nonmonotonic part which appears at the similar high theta range.

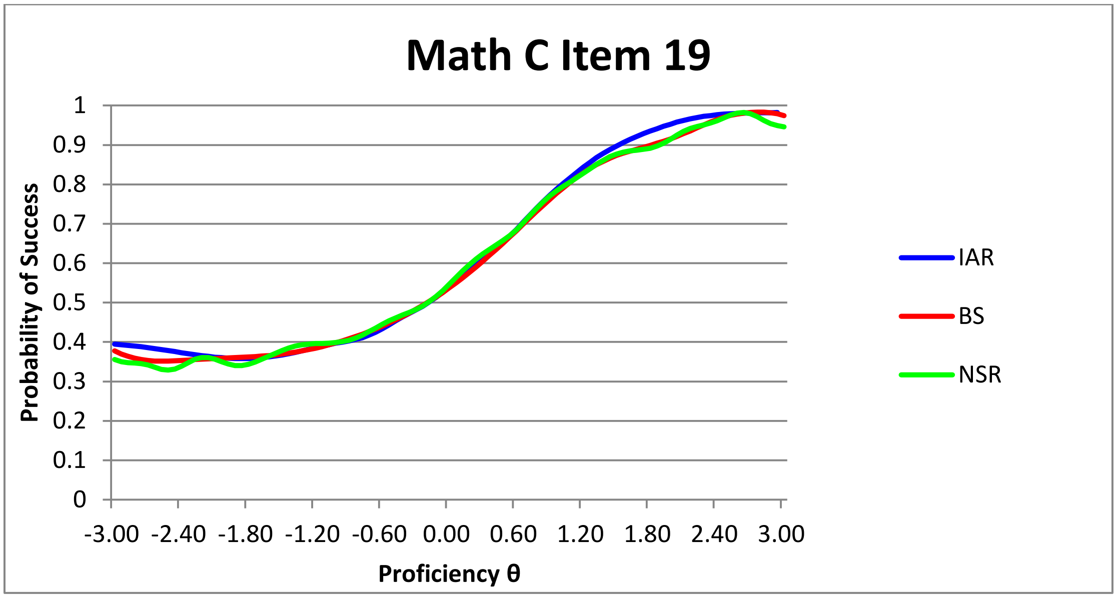

Figure 61.IRFs estimated by three nonparametric methods for item 19 on math C. The PPP-value of the item-ability regression method is 0.022. The PPP-value of the B-splines nonparametric IRT method is 0.766. The PPP-value of the nonparametric smooth regression method is 0.864.

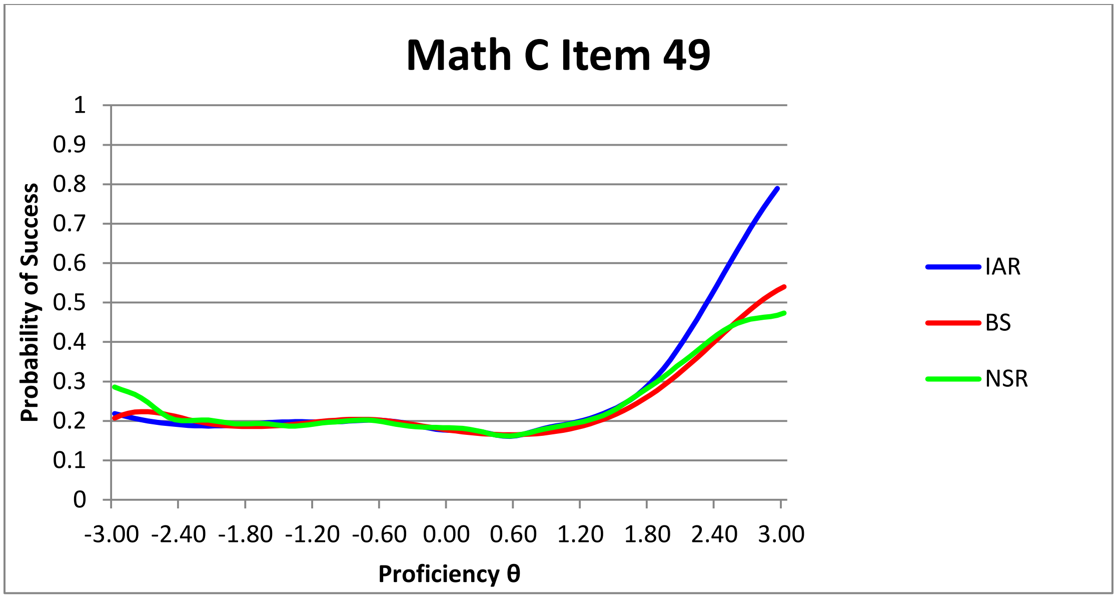

*Figure 62.*IRFs estimated by three nonparametric methods for item 49 on math C. The PPP-value of the item-ability regression method is 0.014. The PPP-value of the B-splines nonparametric IRT method is 0.234. The PPP-value of the nonparametric smooth regression method is 0.020.

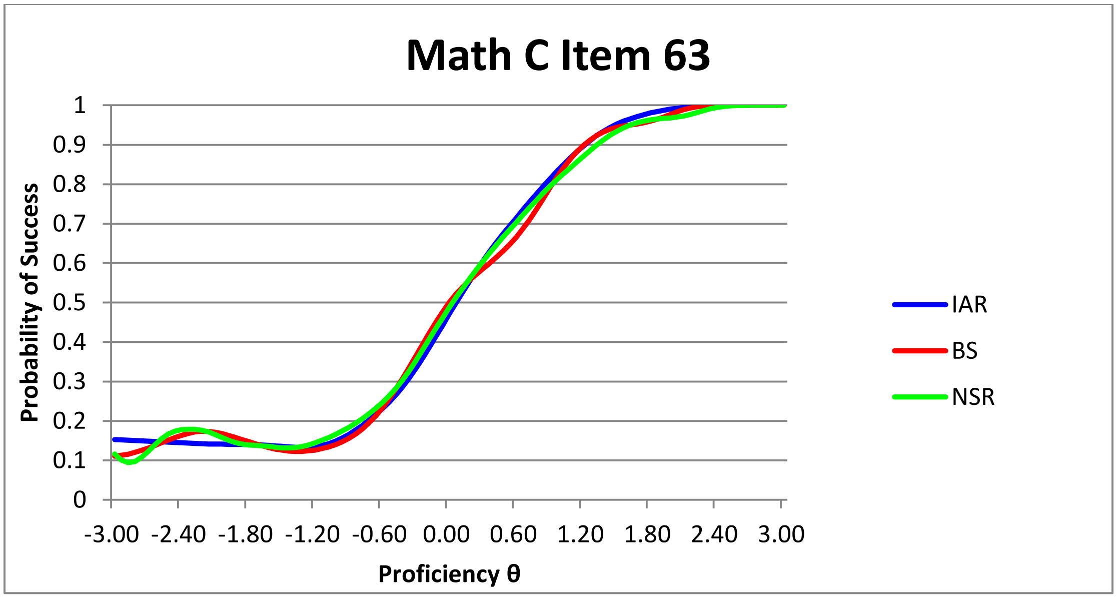

*Figure 63.*IRFs estimated by three nonparametric methods for item 63 on math C. The PPP-value of the item-ability regression method is 0.016. The PPP-value of the B-splines nonparametric IRT method is 0.206. The PPP-value of the nonparametric smooth regression method is 0.224.

For these three items, the differences among the IRFs of three nonparametric methods are small and appear at the very low and high theta range. The nonmonotonic parts of three IRFs appear at the similar theta range. For item 49, the nonmonotonic part appears at the middle theta range (about -0.6 to 0.6) and for item 19 and 63 the nonmonotonic parts appear at the low theta range.

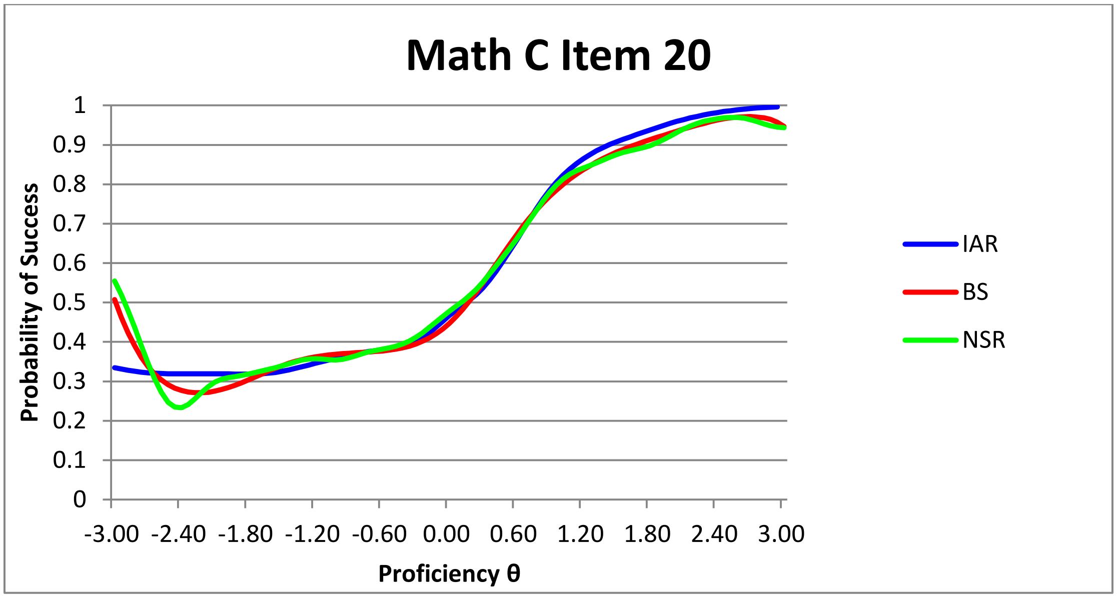

*Figure 64.*IRFs estimated by three nonparametric methods for item 20 on math C. The PPP-value of the item-ability regression method is 0.182. The PPP-value of the B-splines nonparametric IRT method is 0.046. The PPP-value of the nonparametric smooth regression method is 0.022.

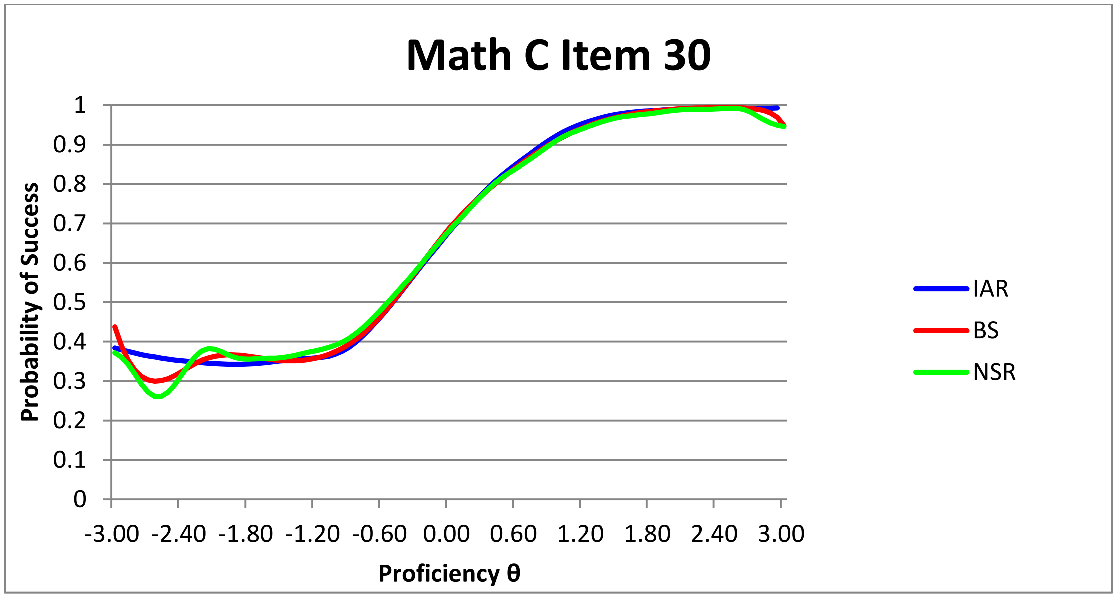

*Figure 65.*IRFs estimated by three nonparametric methods for item 30 on math C. The PPP-value of the item-ability regression method is 0.008. The PPP-value of the B-splines nonparametric IRT method is 0.172. The PPP-value of the nonparametric smooth regression method is 0. 368.

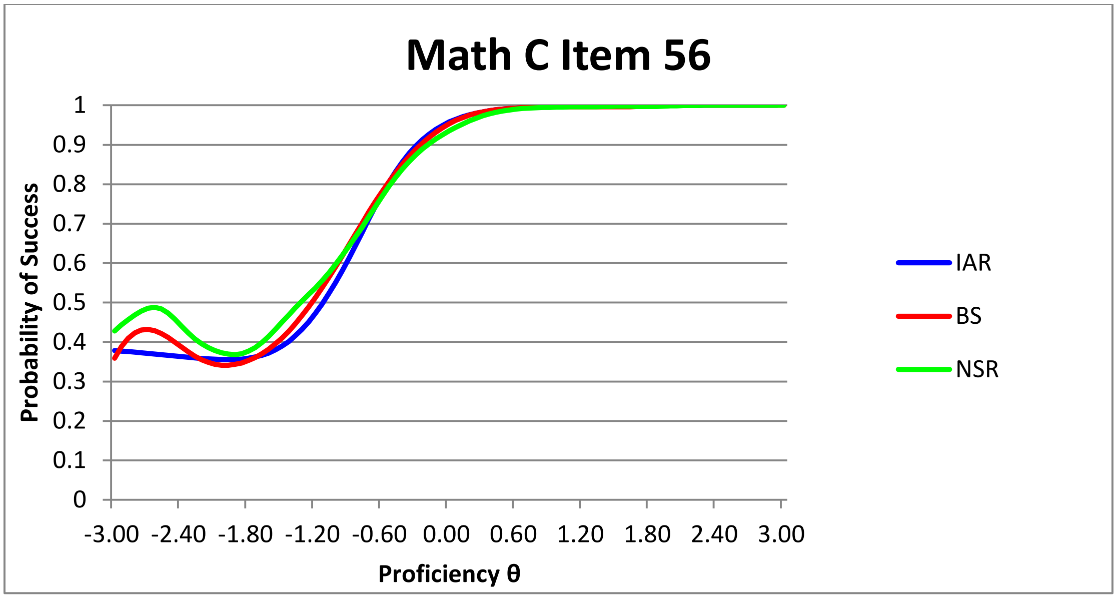

*Figure 66.*IRFs estimated by three nonparametric methods for item 56 on math C. The PPP-value of the item-ability regression method is 0.022. The PPP-value of the B-splines nonparametric IRT method is 0.090. The PPP-value of the nonparametric smooth regression method is 0.038.

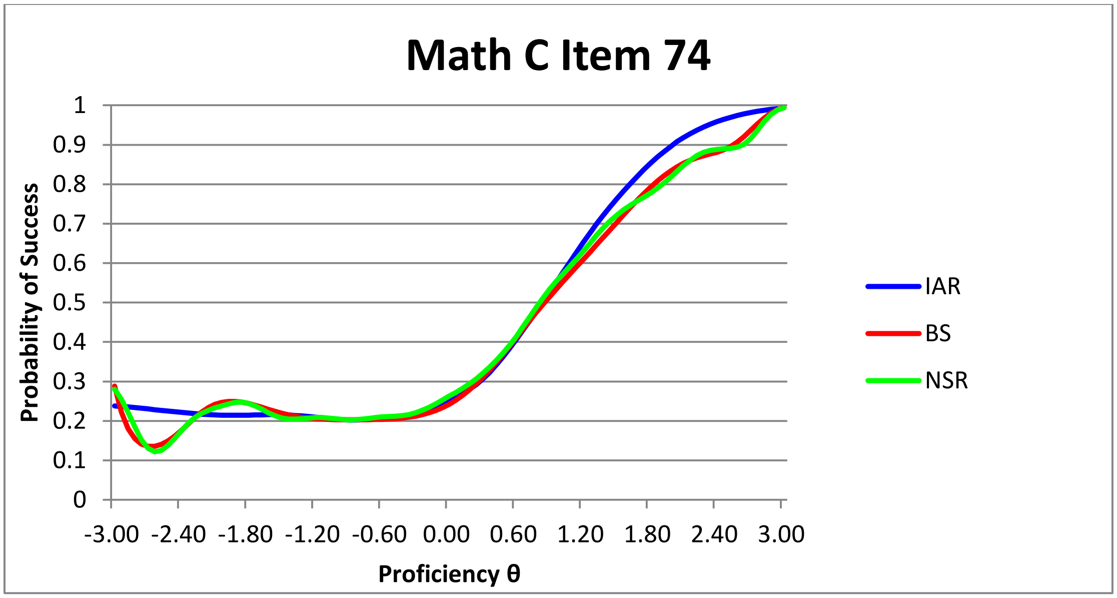

*Figure 67.*IRFs estimated by three nonparametric methods for item 74 on math C. The PPP-value of the item-ability regression method is 0.022. The PPP-value of the B-splines nonparametric IRT method is 0.192. The PPP-value of the nonparametric smooth regression method is 0.164.

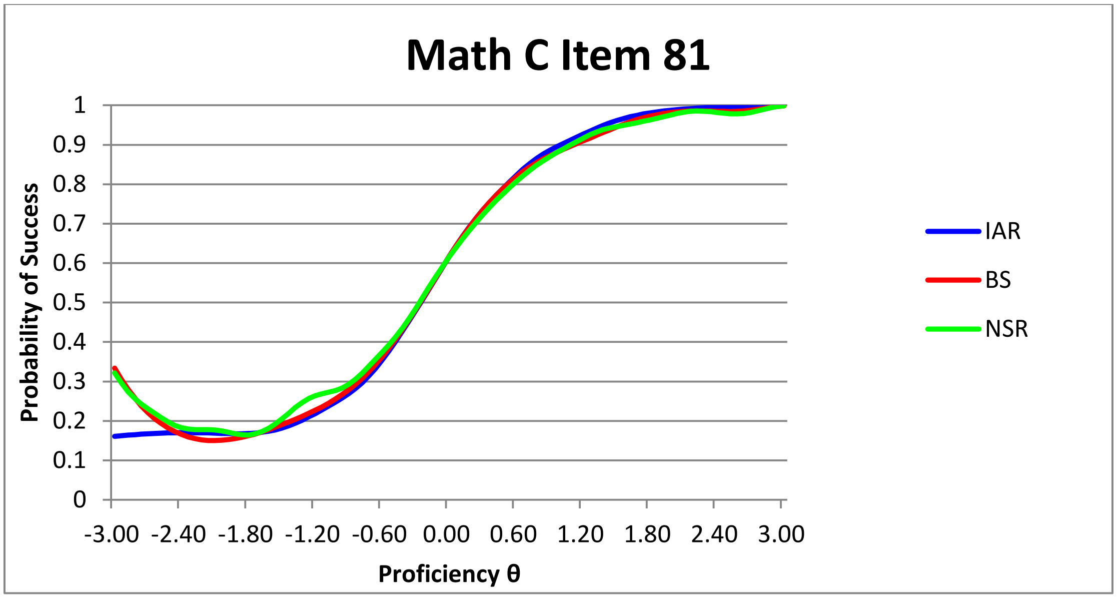

*Figure 68.*IRFs estimated by three nonparametric methods for item 81 on math C. The PPP-value of the item-ability regression method is 0.230. The PPP-value of the B-splines nonparametric IRT method is 0.006. The PPP-value of the nonparametric smooth regression method is 0.006.

For these items, the IRFs of the nonparametric smooth regression method and the B-splines nonparametric IRT method are very similar but are different from the IRF of the item-ability regression method at the low theta range. The IRFs of the B-splines nonparametric IRT method and the nonparametric smooth regression method have more curvatures. The nonmonotonic part of the IRF of the item-ability regression method is smaller than the nonmonotonic parts of the IRFs of the other two methods.

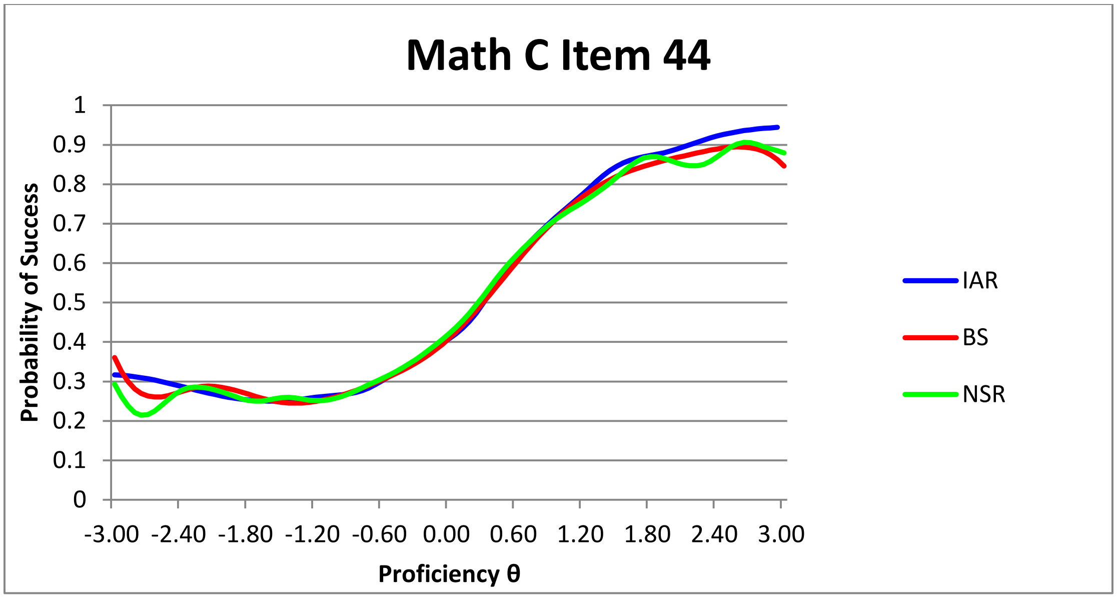

*Figure 69.*IRFs estimated by three nonparametric methods for item 44 on math C. The PPP-value of the item-ability regression method is 0.000. The PPP-value of the B-splines nonparametric IRT method is 0.074. The PPP-value of the nonparametric smooth regression method is 0.336.

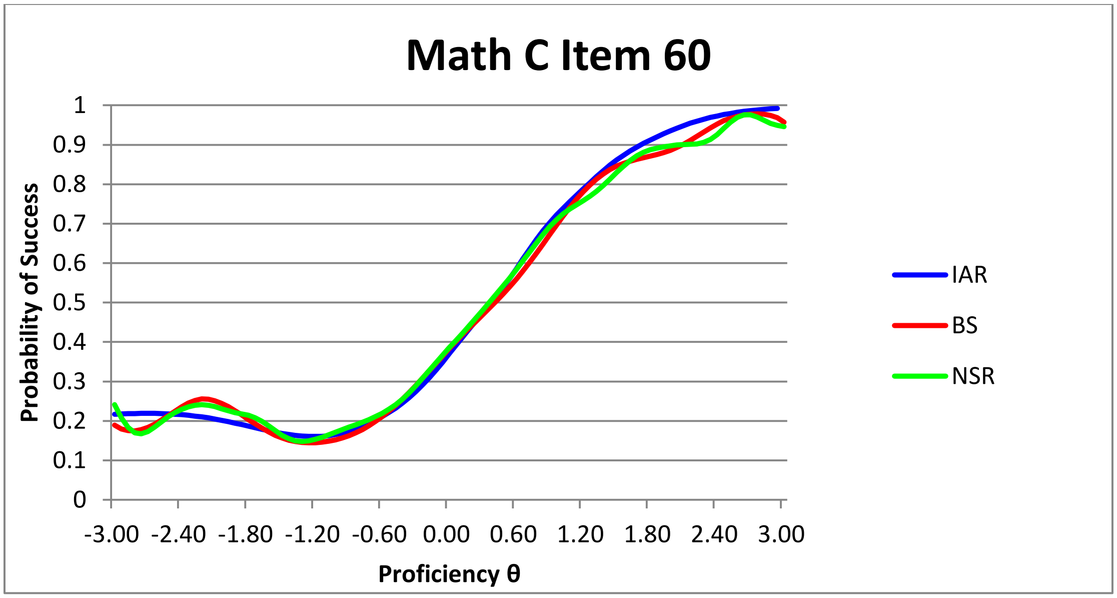

*Figure 70.*IRFs estimated by three nonparametric methods for item 60 on math C. The PPP-value of the item-ability regression method is 0.004. The PPP-value of the B-splines nonparametric IRT method is 0.056. The PPP-value of the nonparametric smooth regression method is 0.166.

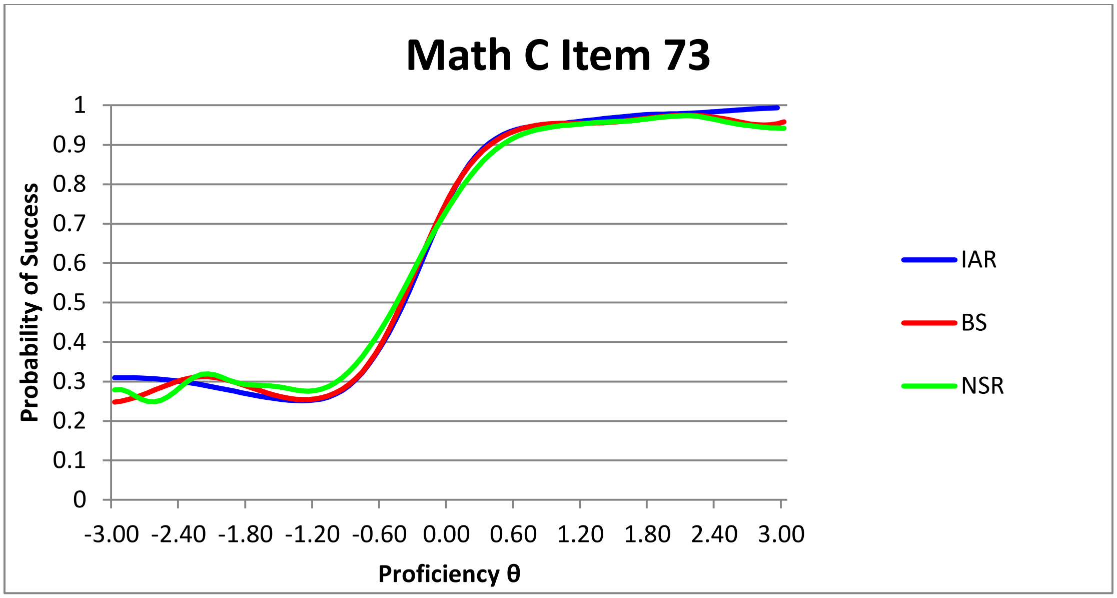

*Figure 71.*IRFs estimated by three nonparametric methods for item 73 on math C. The PPP-value of the item-ability regression method is 0.000. The PPP-value of the B-splines nonparametric IRT method is 0.154. The PPP-value of the nonparametric smooth regression method is 0.200.

For these items, the differences among the IRFs of three nonparametric methods are small. The IRFs of the B-splines nonparametric IRT method and the nonparametric smooth regression method are similar and have more curvatures. The nonmonotonic parts appear at the similar theta range which is the low theta range.

Reading assessment A. Table 13 includes the PPP-values of three nonparametric methods for each item on this assessment.

Table 13

PPP-Value of Three Nonparametric Methods for Each Item on Reading Assessment A

Item	Nonparametric Smooth Regression	Item-Ability Regression	B-Splines Nonparametric IRT
1	0.922	0.508	0.708
2	0.584	0.294	0.738
3	0.088	**0.032**	**0.024**
4	0.820	1.000	0.584
5	0.664	1.000	0.416
6	0.376	1.000	0.594
7	0.122	1.000	0.310
8	0.762	1.000	0.486
9	0.188	1.000	0.268
10	0.686	0.520	0.672
11	0.874	0.368	0.732
12	0.478	0.330	0.698
13	0.306	0.136	0.134
14	0.204	0.296	0.246
15	0.794	1.000	1.000
16	0.796	0.408	0.822
17	0.826	1.000	0.792
18	0.516	0.326	0.646
19	0.422	1.000	0.580
20	0.516	1.000	0.854
21	0.928	1.000	1.000
22	0.698	1.000	1.000
23	0.764	0.218	0.298
24	0.166	1.000	0.338
25	**0.042**	1.000	0.180
26	0.728	1.000	0.732
27	0.220	0.360	0.496
28	0.736	1.000	1.000
29	0.982	0.398	1.000
30	0.932	1.000	1.000
31	0.556	0.586	0.240
32	**0.002**	1.000	**0.038**
33	0.652	1.000	0.556
34	0.946	1.000	0.610
35	0.962	1.000	0.784
36	0.744	0.560	0.724
37	0.674	0.696	0.556
38	0.668	0.174	0.386
39	0.352	1.000	0.326
40	1.000	1.000	1.000
41	0.750	1.000	0.310

42	0.358	1.000	0.452
43	0.726	1.000	0.668
44	0.408	1.000	0.574
45	0.828	0.448	0.858
46	0.098	0.090	**0.026**
47	0.670	0.870	0.548
48	0.500	0.146	0.742
49	0.172	0.448	0.130
50	0.356	1.000	0.180
51	0.614	1.000	0.374
52	0.928	0.154	1.000
53	0.646	0.074	0.352
54	0.586	1.000	0.788
55	0.394	1.000	0.818
56	0.314	0.484	0.460
57	0.112	1.000	0.334
58	0.694	1.000	1.000
59	0.760	1.000	1.000
60	0.802	0.586	0.560
61	0.158	0.228	0.104
62	0.292	0.590	0.176
63	0.232	1.000	0.220
64	0.080	**0.050**	0.256
65	1.000	1.000	1.000
66	0.962	0.108	1.000
67	0.570	0.370	0.572
68	0.760	1.000	1.000
69	0.306	0.714	0.352
70	0.346	1.000	0.512
71	**0.000**	**0.002**	**0.004**
72	0.322	0.552	0.290
73	0.450	1.000	0.436
74	0.208	0.322	0.190
75	0.774	1.000	1.000
76	0.814	0.128	0.354
77	0.266	0.252	0.138
78	0.846	1.000	1.000
79	0.658	0.632	0.624
80	0.162	0.336	0.174

Three items are identified as items with nonmonotonic IRFs by the nonparametric smooth regression method, three items are identified with nonmonotonic IRFs by the item-ability regression method, and four items are identified with nonmonotonic IRFs by the B-splines nonparametric IRT method. Item 71 is identified with nonmonotonic IRF by three methods. Also item 3 is identified with nonmonotonic IRF by two methods and the PPP value of the third method is very close to 0.05 (0.088 of the nonparametric smooth regression method).

Table 14 presents the nonmonotonic area correlation among three nonparametric methods. The nonmonotonic area correlation between two nonparametric methods is the correlation of the nonmonotonic area of all items on the reading assessment A calculated by these two methods.

Table 14
Nonmonotonic Area Correlation among Three Nonparametric Methods for Reading Assessment A

	Nonparametric Smooth Regression	Item-Ability Regression	B-Splines Nonparametric IRT
Nonparametric Smooth Regression		0.886	0.952
Item-Ability Regression			0.897
B-Splines Nonparametric IRT			

The nonmonotonic area correlation between the nonparametric smooth regression method and the B-splines nonparametric IRT method is close to 1 (0.952). In addition, the nonmonotonic area correlation between the nonparametric smooth regression method and the item-ability regression method (0.886) and the nonmonotonic area correlation between the B-splines nonparametric IRT method and the item-ability regression (0.897) are high and similar. The very high nonmonotonic correlation between the nonparametric smooth regression method and the B-splines nonparametric IRT method indicates the nonmonotonic area estimated by two methods have a similar pattern among items on the reading assessment A. The nonmonotonic area correlations among these three methods are high which indicate the nonmonotonic area patterns among items on reading assessment A estimated by these three methods are similar.

Figures 72-77 present the IRFs estimated by three nonparametric methods of items identified as items with nonmonotonic IRFs by at least one method. Figure 72 presents the IRFs of item 71 which is identified with nonmonotonic IRF by three nonparametric methods.

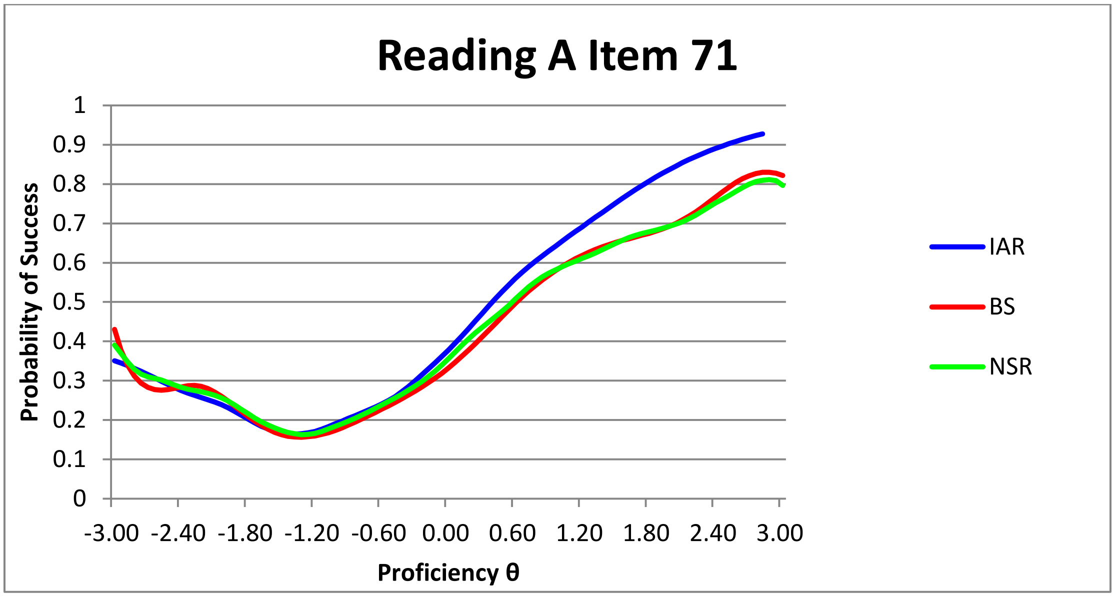

*Figure 72.*IRFs estimated by three nonparametric methods for item 71 on reading A. The PPP-value of the item-ability regression method is 0.002. The PPP-value of the B-splines nonparametric IRT method is 0.004. The PPP-value of the nonparametric smooth regression method is 0.000.

For item 71, the IRFs of three nonparametric methods are very similar with the exception of the parts at the low and high theta range and the nonmonotonic part appears at the same theta range.

Figure 73 presents the IRFs of item 3 which is identified as an item with nonmonotonic IRF by two methods and the PPP value of the third method is very close to 0.05.

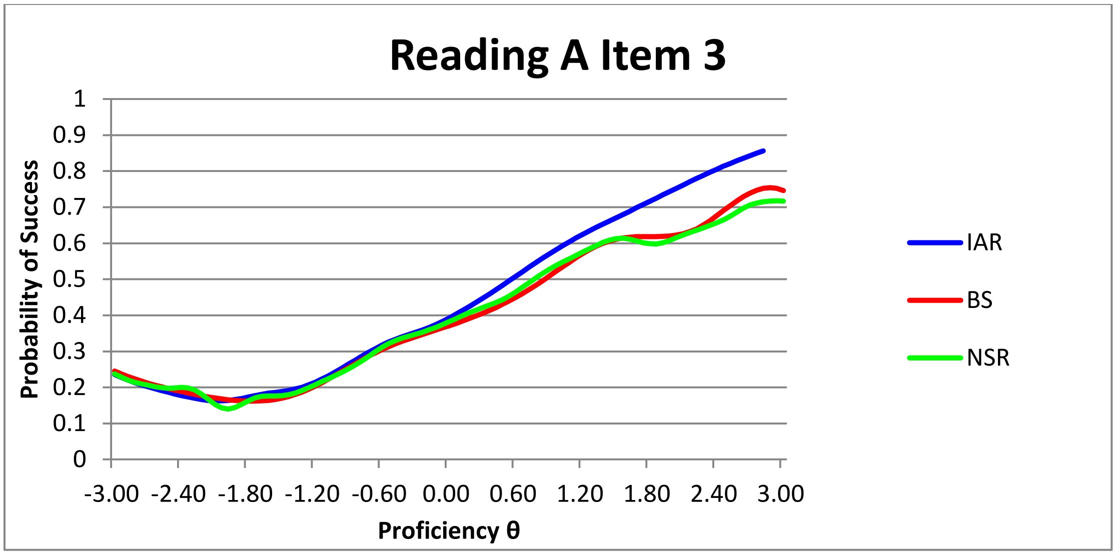

*Figure 73.*IRFs estimated by three nonparametric methods for item 3 on reading A. The PPP-value of the item-ability regression method is 0.032. The PPP-value of the B-splines nonparametric IRT method is 0. 024. The PPP-value of the nonparametric smooth regression method is 0.088.

For this item, the differences among the IRFs of three nonparametric methods are at the high theta range. The IRFs of the nonparametric smooth regression method and the B-splines nonparametric IRT method are very similar. The nonmonotonic parts of three IRFs appear at the similar theta range.

Figures 74-77 present the IRFs of the other items identified with nonmonotonic IRFs by at least one method. These items are 25, 32, 46, and 64.

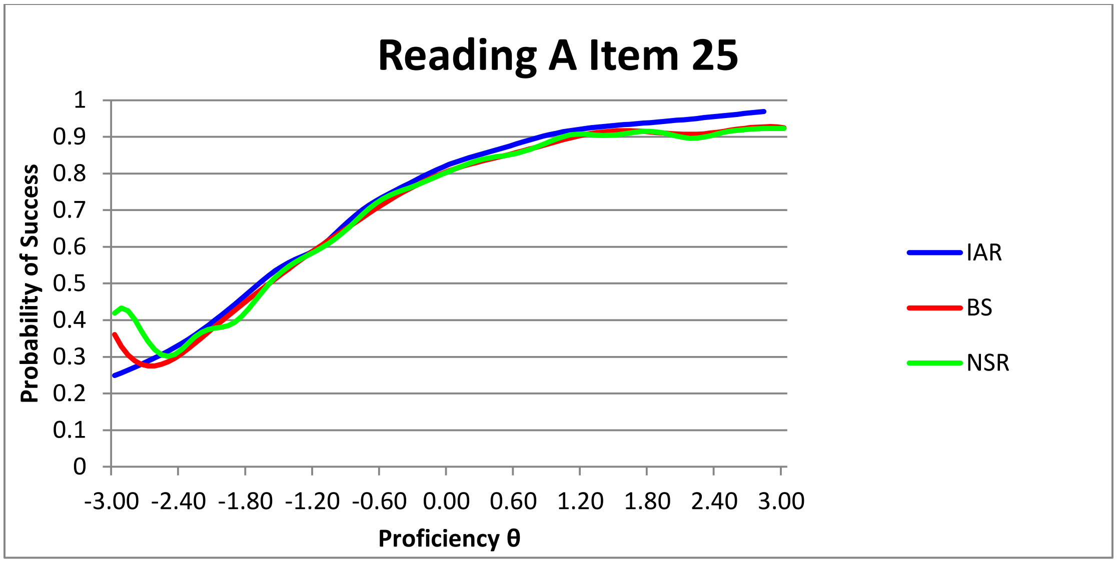

*Figure 74.*IRFs estimated by three nonparametric methods for item 25 on reading A. The PPP-value of the item-ability regression method is 1.000. The PPP-value of the B-splines nonparametric IRT method is 0. 180. The PPP-value of the nonparametric smooth regression method is 0.042.

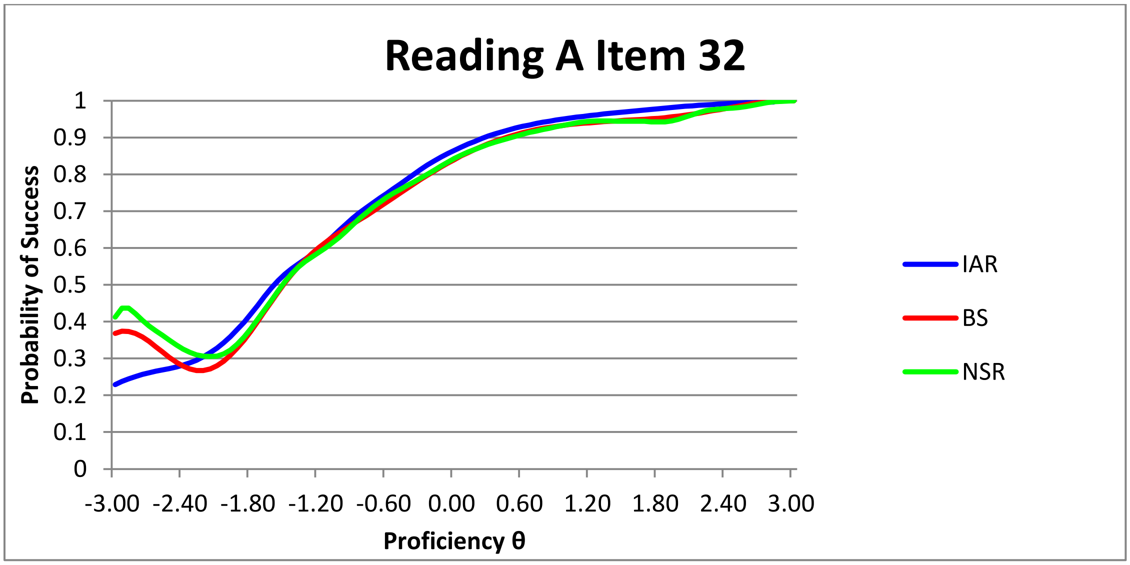

*Figure 75.*IRFs estimated by three nonparametric methods for item 32 on reading A. The PPP-value of the item-ability regression method is 1.000. The PPP-value of the B-splines nonparametric IRT method is 0.038. The PPP-value of the nonparametric smooth regression method is 0.002.

For item 25 and 32, the IRFs of the nonparametric smooth regression method and the B-splines nonparametric IRT method are very similar but are different from the IRF of the item-ability regression method especially at the low theta range. The IRFs of the B-splines nonparametric IRT method and the nonparametric smooth regression method have more curvatures. The IRFs of the item-ability regression method are monotonic.

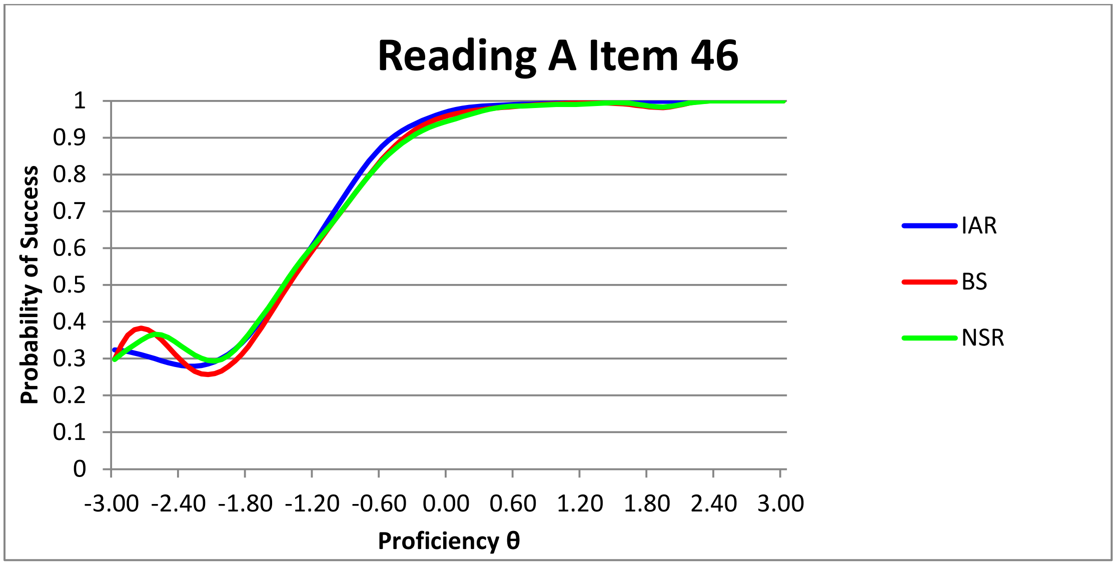

*Figure 76.*IRFs estimated by three nonparametric methods for item 46 on reading A. The PPP-value of the item-ability regression method is 0.090. The PPP-value of the B-splines nonparametric IRT method is 0.026. The PPP-value of the nonparametric smooth regression method is 0.098.

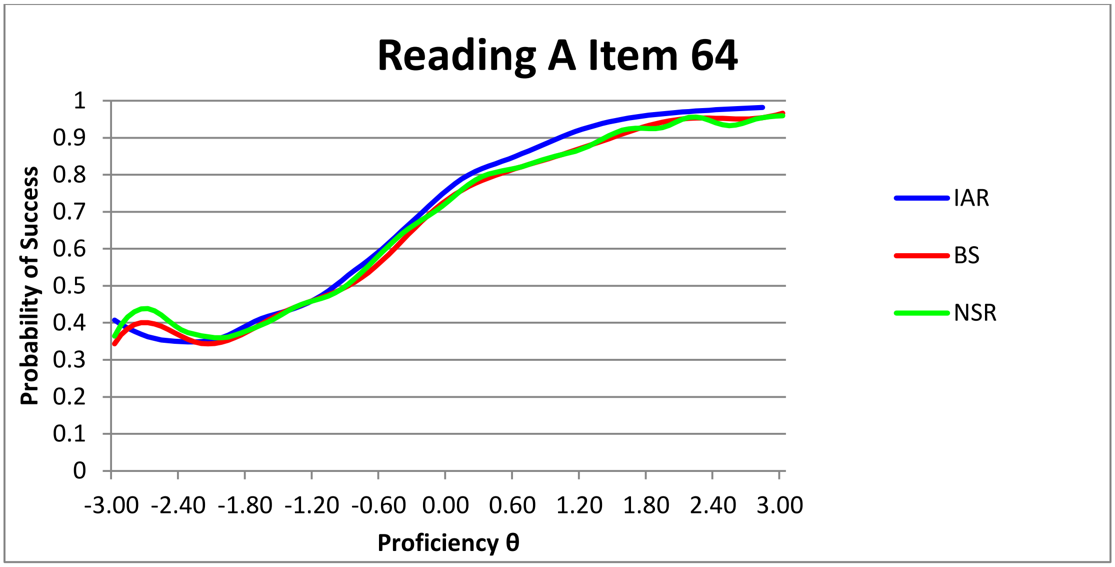

*Figure 77.*IRFs estimated by three nonparametric methods for item 64 on reading A. The PPP-value of the item-ability regression method is 0.050. The PPP-value of the B-splines nonparametric IRT method is 0. 256. The PPP-value of the nonparametric smooth regression method is 0.080.

For item 46 and 64, the IRFs of the nonparametric smooth regression method and the B-splines nonparametric IRT method are very similar but are different from the IRF of the item-ability regression method at the low theta range. The IRFs of the B-splines nonparametric IRT method and the nonparametric smooth regression method have more curvatures. The nonmonotonic parts of the IRFs of three methods appear at the similar theta range.

Reading assessment B. Table 15 includes the PPP-values of three nonparametric methods for each item on this assessment.

Table 15

PPP-Value of Three Nonparametric Methods for Each Item on Reading Assessment B

Item	Nonparametric Smooth Regression	Item-Ability Regression	B-Splines Nonparametric IRT
1	0.092	0.138	0.220
2	0.774	1.000	0.850
3	0.076	1.000	0.344
4	0.934	1.000	0.794
5	0.356	1.000	0.448
6	0.510	0.064	**0.016**
7	1.000	1.000	1.000
8	0.908	1.000	1.000
9	0.516	0.276	0.326
10	0.952	1.000	0.696
11	0.792	0.128	0.760
12	0.758	1.000	0.684
13	0.760	1.000	1.000
14	0.156	0.776	0.198
15	0.322	0.058	0.222
16	0.080	1.000	0.118
17	0.960	0.550	0.808
18	0.736	1.000	0.416
19	0.798	1.000	1.000
20	0.778	1.000	1.000
21	**0.038**	1.000	0.094
22	0.728	1.000	0.576
23	0.772	1.000	1.000
24	0.580	0.372	0.354
25	0.112	0.262	0.068
26	0.614	1.000	1.000
27	0.406	0.242	0.210
28	0.486	0.202	0.220
29	0.108	0.258	0.202
30	0.446	0.136	0.316
31	0.988	1.000	1.000
32	0.644	1.000	1.000
33	0.638	0.706	0.464
34	0.140	1.000	0.126
35	0.118	0.326	0.378
36	0.428	0.192	0.272
37	0.814	0.598	0.840
38	0.248	1.000	0.242
39	0.352	1.000	0.334
40	0.260	1.000	0.478
41	0.854	1.000	0.662

42	0.116	1.000	0.086
43	0.960	0.292	1.000
44	0.532	1.000	0.424
45	1.000	1.000	1.000
46	1.000	1.000	1.000
47	0.248	1.000	0.324
48	**0.020**	1.000	**0.028**
49	0.168	0.524	0.114
50	1.000	1.000	1.000
51	0.282	0.160	0.282
52	0.134	0.440	0.238
53	0.560	1.000	0.640
54	0.340	0.262	0.200
55	0.708	1.000	0.794
56	0.834	0.784	0.774
57	0.128	1.000	0.110
58	0.356	1.000	0.452
59	0.818	1.000	0.778
60	0.072	0.084	0.198
61	0.470	0.152	0.214
62	0.542	1.000	0.806
63	0.442	1.000	0.112
64	**0.038**	**0.010**	**0.006**
65	0.812	1.000	0.624
66	0.650	**0.012**	**0.046**
67	0.722	0.660	0.556
68	0.564	1.000	0.546
69	0.788	0.156	0.300
70	0.806	0.136	0.674
71	0.084	0.234	0.052
72	0.196	0.198	0.122
73	0.612	1.000	0.512
74	0.070	1.000	0.364
75	0.926	1.000	0.888
76	0.404	1.000	**0.038**
77	0.580	1.000	1.000
78	0.846	1.000	1.000
79	0.946	1.000	1.000
80	1.000	1.000	0.932
81	0.794	1.000	0.582

Three items are identified with nonmonotonic IRFs by the nonparametric smooth regression method, two items are identified with nonmonotonic IRFs by the item-ability regression method, and five items are identified with nonmonotonic IRFs by the B-splines nonparametric IRT method. Item 64 is identified with nonmonotonic IRF by three methods.

Table 16 presents the nonmonotonic area correlation among three nonparametric methods. The nonmonotonic area correlation between two nonparametric methods is the correlation of the nonmonotonic area of all items on the reading assessment B calculated by these two methods.

Table 16
Nonmonotonic Area Correlation among Three Nonparametric Methods for Reading Assessment B

	Nonparametric Smooth Regression	Item-Ability Regression	B-Splines Nonparametric IRT
Nonparametric Smooth Regression		0.441	0.782
Item-Ability Regression			0.761
B-Splines Nonparametric IRT			

The nonmonotonic area correlation between the nonparametric smooth regression method and the B-splines nonparametric IRT method is highest (0.782). In addition, the nonmonotonic area correlation between the B-splines nonparametric IRT method and the item-ability regression method (0.761) is high but the nonmonotonic area correlation between the nonparametric smooth regression method and the item-ability regression method (0.441) is not very high. Three nonmonotonic correlations of the reading assessment B are not as high as three nonmonotonic correlations of the reading assessment A, which indicate that the nonmonotonic area patterns among items on the reading assessment B estimated by these three methods are not very similar.

Figures 78-83 IRFs estimated by three nonparametric methods of items identified with nonmonotonic IRFs by at least one method. Figure 78 presents the IRFs of item 64 which is identified with nonmonotonic IRF by three nonparametric methods.

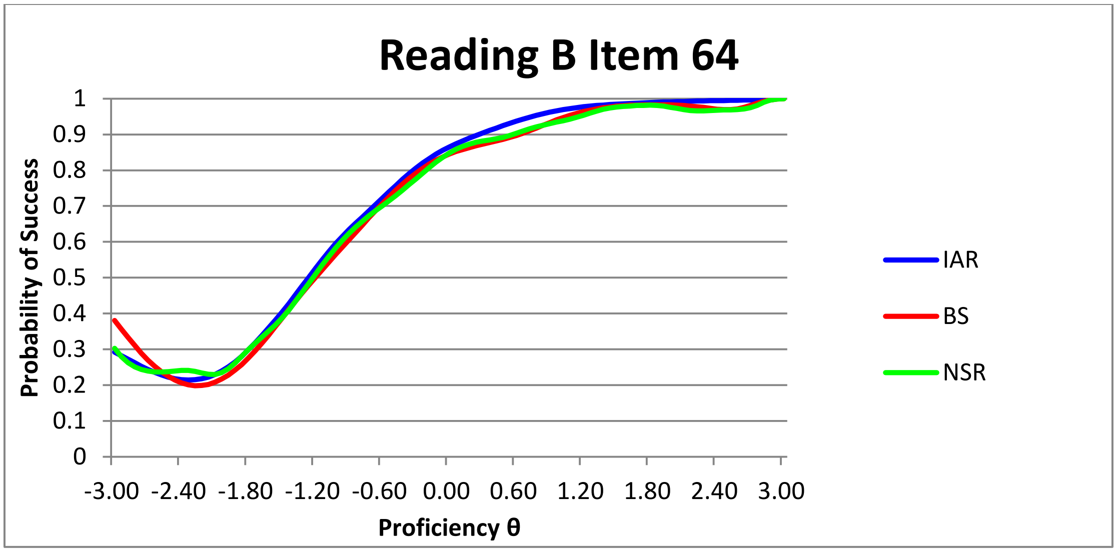

*Figure 78.*IRFs estimated by three nonparametric methods for item 64 on reading B. The PPP-value of the item-ability regression method is 0.010. The PPP-value of the B-splines nonparametric IRT method is 0.006. The PPP-value of the nonparametric smooth regression method is 0.038.

For item 64, the IRFs of three nonparametric methods are very similar with the exception of the end parts at the very low and high theta range and the nonmonotonic parts of three IRFs appear at the same theta range.

Figures 79-83 present the IRFs of the other items identified with nonmonotonic IRFs by at least one method. These items are 6, 21, 48, 66, and 76.

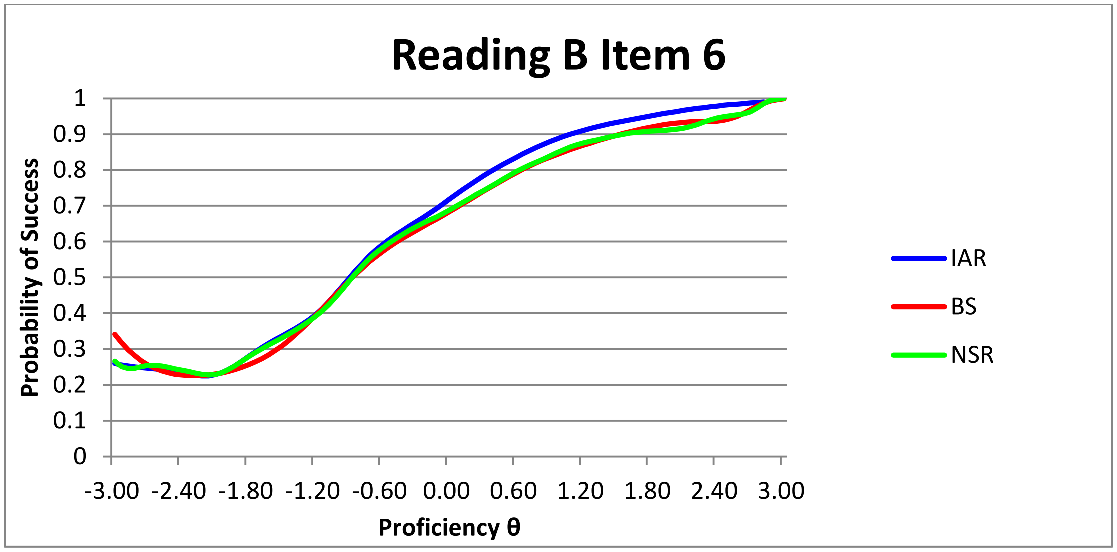

*Figure 79.*IRFs estimated by three nonparametric methods for item 6 on reading B. The PPP-value of the item-ability regression method is 0.064. The PPP-value of the B-splines nonparametric IRT method is 0.016. The PPP-value of the nonparametric smooth regression method is 0.510.

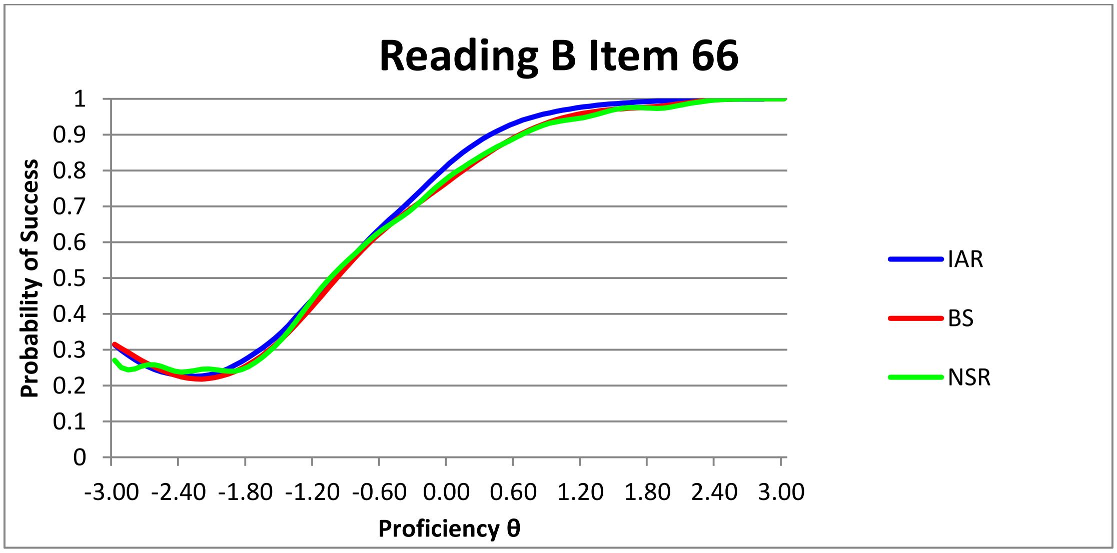

*Figure 80.*IRFs estimated by three nonparametric methods for item 66 on reading B. The PPP-value of the item-ability regression method is 0.012. The PPP-value of the B-splines nonparametric IRT method is 0.046. The PPP-value of the nonparametric smooth regression method is 0.650.

For items 6 and 66, the IRFs of three nonparametric methods are very similar with the exception of the end parts at the very low and high theta range and the nonmonotonic parts of three IRFs appear at the same theta range.

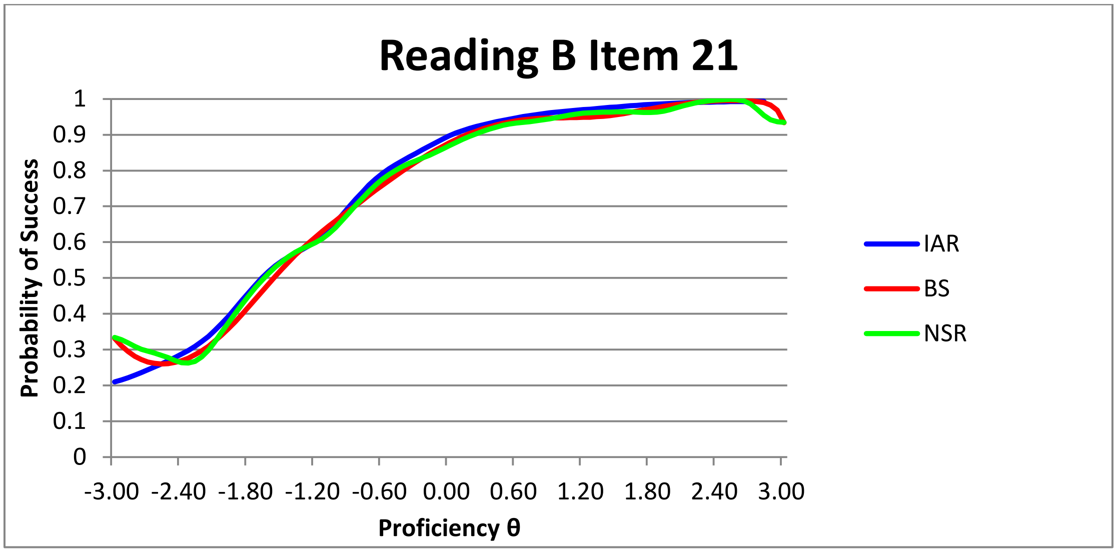

Figure 81.IRFs estimated by three nonparametric methods for item 21 on reading B. The PPP-value of the item-ability regression method is 1.000. The PPP-value of the B-splines nonparametric IRT method is 0.038. The PPP-value of the nonparametric smooth regression method is 0.094.

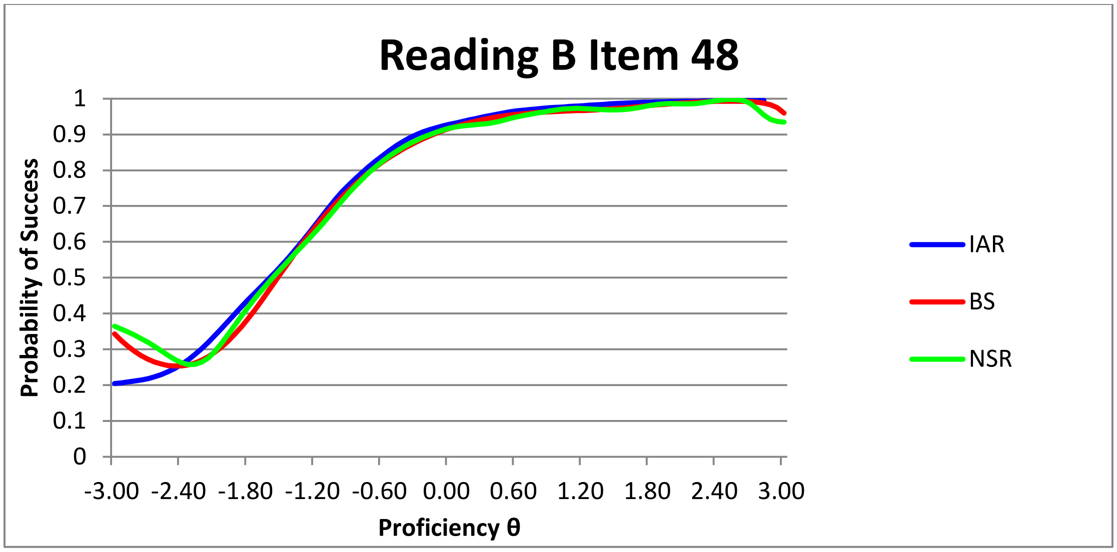

*Figure 82.*IRFs estimated by three nonparametric methods for item 48 on reading B. The PPP-value of the item-ability regression method is 1.000. The PPP-value of the B-splines nonparametric IRT method is 0.028. The PPP-value of the nonparametric smooth regression method is 0.020.

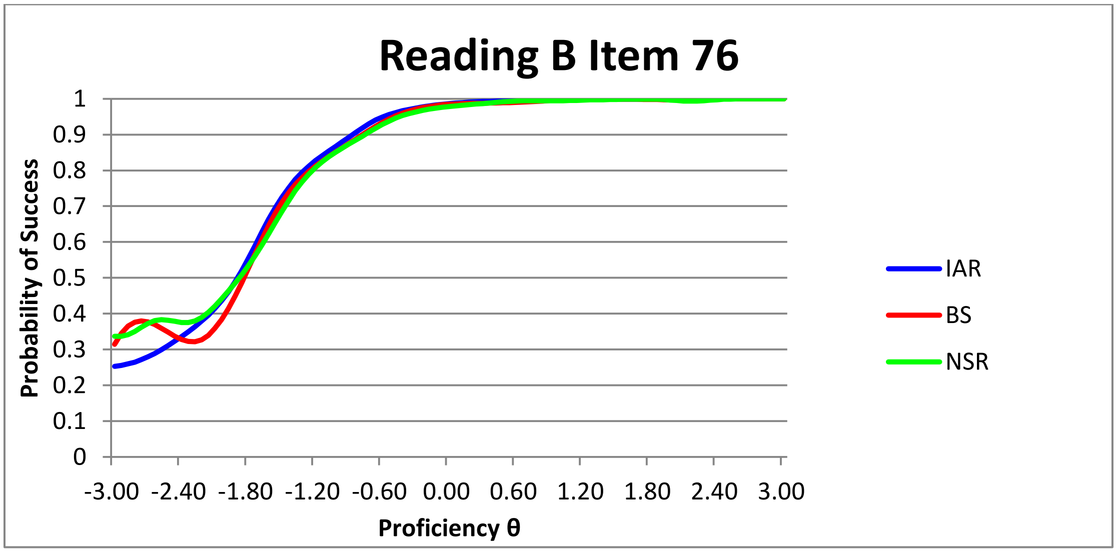

*Figure 83.*IRFs estimated by three nonparametric methods for item 76 on reading B. The PPP-value of the item-ability regression method is 1.000. The PPP-value of the B-splines nonparametric IRT method is 0.038. The PPP-value of the nonparametric smooth regression method is 0.404.

For items 21, 48, and 76, the IRFs of the nonparametric smooth regression method and the B-splines nonparametric IRT method are very similar but are different from the IRF of the item-ability regression method especially at the low theta range. The IRFs of the B-splines nonparametric IRT method and the nonparametric smooth regression method have more curvatures. The IRF of the item-ability regression method is monotonic.

Summary

In the simulation study, the type I and type II error rates of three methods are similar. The type I error rate is lower than the type II error rate. The item-ability regression method has the lowest type I and type II error rates. The nonmonotonic area correlation between the nonparametric smooth regression method and the B-splines nonparametric IRT method is highest. Each real assessment has several items identified with nonmonotonic IRFs. The math assessments have more items with nonmonotonic IRFs than the reading assessments.

Chapter 5: Discussion

The present study aimed to answer the following research questions:

1. Can nonparametric smooth regression, item-ability regression, and B-spline nonparametric IRT detect the nonmonotonic IRF accurately in a simulated data study?
2. Do real items with nonmonotonic IRFs exist, and if so, how common are they?

These two research questions are discussed based on the results from the simulation study and the real data study. The discussion of the simulation study focuses on comparing three nonparametric methods, which corresponds to the first research question. The discussion of the real data study focuses on real items identified with nonmonotonic IRFs on five assessments, which corresponds to the second research question. Limitations of this study and suggestions for further research follow these discussions. Finally, the conclusion provides the implication of this study for educational research and practice.

Nonparametric Methods Comparison in the Simulation Study

Three nonparametric methods were compared from three aspects. First, three methods were compared on how well they identify six items with true nonmonotonic IRFs. Second, the average type I and type II error rate of three methods were compared. Third, similarities among the three methods on estimating the nonmonotonic area were compared.

Method comparison on the estimation of items with true nonmonotonic IRFs. Among 60 simulated items, six of them were generated from the nonmonotonic model (equation 36). The type II error rate of one method for one item indicates how well this method identifies one item with true nonmonotonic IRF. The type II error rates for the last two items are much lower than the type II error rates for the first four items of every method. The last two items have extreme item guessing parameters but their item discriminate and difficulty parameters are in the middle range. The other four items either have extreme item difficulty parameter or extreme item discriminate parameter. Figures 84, 85, and 86 present the relationship between the type II error rate and each true item parameters.

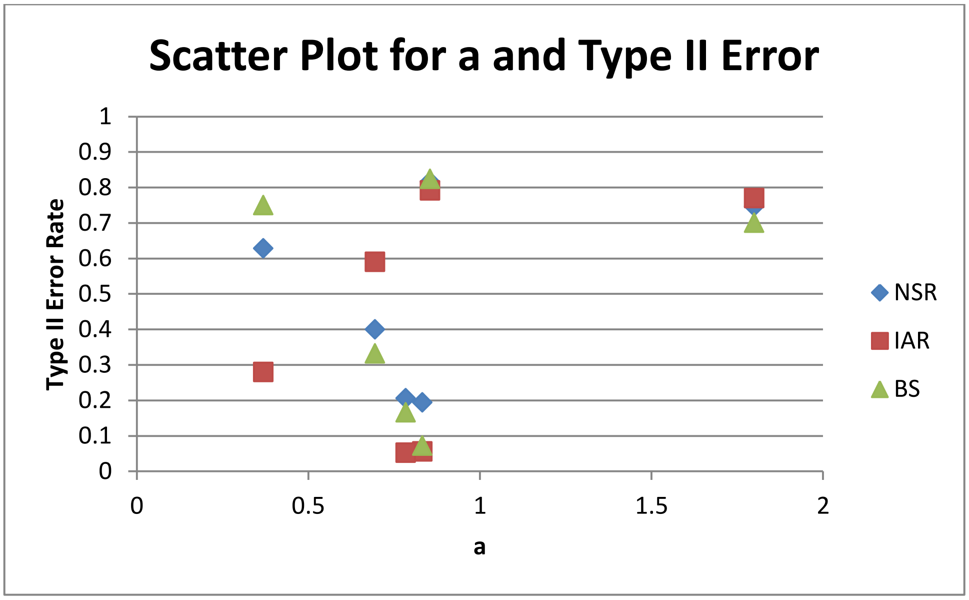

Figure 84. Scatter plot for a and type II error rate for each nonparametric method. Each method has six type II error rates correspond to six items with true nonmonotonic IRF.

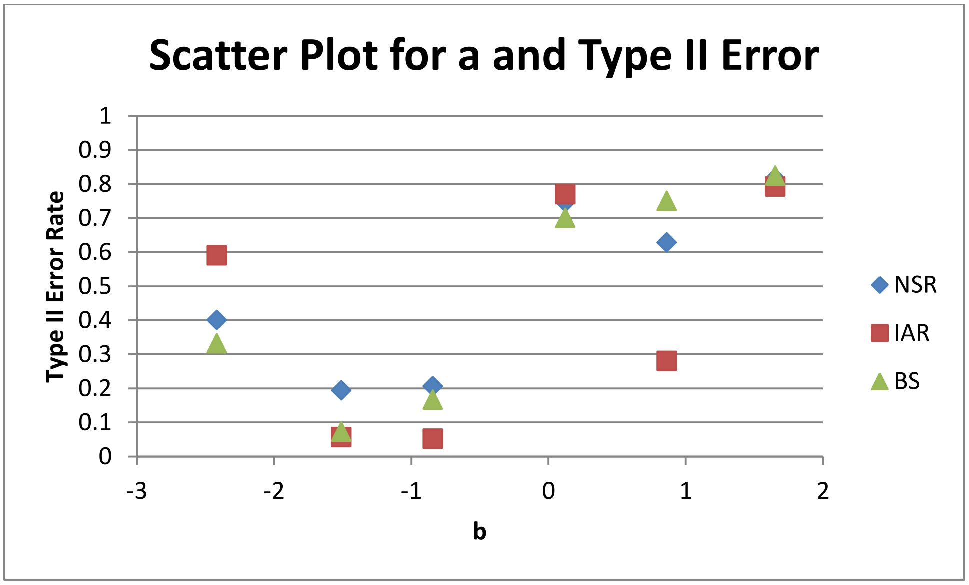

Figure 85. Scatter plot for b and type II error rate for each nonparametric method. Each method has six type II error rates correspond to six items with true nonmonotonic IRF.

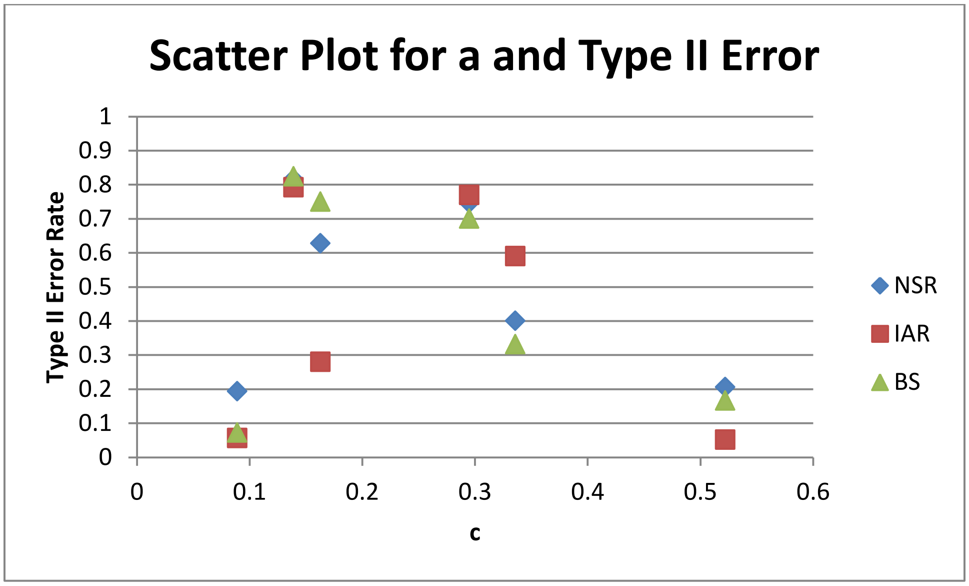

Figure 86. Scatter plot for c and type II error rate for each nonparametric method. Each method has six type II error rates correspond to six items with true nonmonotonic IRF.

This result indicates three nonparametric methods can estimate accurately the nonmonotonic IRFs of items with middle range item discriminate and difficulty parameters but extreme item guessing parameter. One reason is that the guessing parameter did not affect the nonmonotonicity estimation of these three methods significantly, when equation 36 was used to generate the nonmonotonic data.

Since the B-splines nonparametric IRT method used the estimation from the nonparametric smooth regression method as priors, these two methods have a similar IRF estimation and the type II error rate of these two methods for each item are very similar. Moreover, the type II error rate of the item-ability regression method for each item is close to the type II error rate of the other two methods. Also the mean p-values for six items with true nonmonotonic IRFs are small. Thus, these three nonparametric methods can identify the nonmonotonic IRF equally well.

Method comparison on the average type I and type II error rate. The differences of the average type I and type II error rate among three methods are not large. This result confirms the first hypothesis that three nonparametric methods can detect the nonmonotonic IRF equally. However, the type II error rate is much higher than the type I error rate of each method. The high type II error is equivalent to the low power to identify the nonmonotonicity. This result means that the probability that each method identifies a monotonic item as nonmonotonic is very low but they sometimes cannot detect nonmonotonicity very well. The reasons might be the nonmonotonic models and the extreme item parameters being used for generating data.

Method comparison on nonmonotonic area correlation. The nonmonotonic area correlation is an index of the similarity between two nonparametric methods on estimating the nonmonotonic part of IRF. The result of nonmonotonic area correlation indicates nonmonotonic area estimated by the nonparametric smooth regression method and the B-splines nonparametric IRT method have a similar pattern. The item-ability regression method has fewer similarities with these two methods. The reason for the similar estimation pattern between the nonparametric smooth regression method and the B-splines nonparametric IRT method is that the B-splines nonparametric IRT method used the estimation from the nonparametric smooth regression method as priors.

Real Items Identified with Nonmonotonic IRFs

IRFs of items on five assessments were estimated by three nonparametric methods. Then the PPMC method was used to judge the extent of the nonmonotonicity of IRFs. Each assessment has items identified with nonmonotonic IRFs by at least one method. Hambleton and Han (2004) have discussed that little attention has been given to investigating the consequences of model misfit. Moreover, Hambleton and Swaminathan (1985, p.168) stated that it is always helpful to evaluate the consequences of model misfit whenever possible and that the number of investigations should not be limited. Investigation of the consequences of model misfit means that the assessment of model fit does not end. A rejection of a model cannot be made by an extreme statistical value without assessing the consequences of model misfit. Since the nonmonotonic IRF leads to the model misfit, the reasons for and the consequences of the nonmonotonic IRF should be investigated. Thus, this study also investigated whether the consequences of nonmonotonic IRFs affect the fairness and comparability of the test score. This investigation from different aspects was conducted on several items. Since all the real items are security items, they are not presented during investigation and explanation.

In this section, the classification of the nonmonotonic IRFs is discussed first. The items identified with nonmonotonic IRFs are grouped by the subjects and discussed from the content perspective next. The reasons for and consequences of several items identified with nonmonotonic IRFs were carefully examined and presents in both classification part and subject discussion part.

Categories of the real items identified with nonmonotonic IRFs. Two ways of classification for items identified with nonmonotonic IRFs are discussed. One way of classification is based on the similarity of IRFs from three estimation methods at the nonmonotonic part. And the other way of classification is based on the theta range where nonmonotonicity appears.

Separate items identified with nonmonotonic IRFs based on the similarity of IRFs from three estimation methods at the nonmonotonic part. Items identified with nonmonotonic IRFs can be separated into four categories based on the similarity of IRFs from three estimation methods at the nonmonotonic part: (1) IRFs from three estimation methods are very similar at the nonmonotonic part, (2) IRFs from three estimation methods have the similar theta range where the nonmonotonic part appears but are different on this range, (3) IRFs from two estimation methods are nonmonotonic and IRF from one estimation method is monotonic, and (4) IRF from one estimation method is nonmonotonic and IRFs from two estimation methods are monotonic.

Three nonparametric methods estimated IRF using different models. The IRFs which are estimated by the nonparametric smooth regression method and the B-spline nonparametric IRT method have more curvatures compared with the IRFs estimated by the item-ability regression method. This is the main reason leading to the differences among IRFs from three estimation methods. When there is one monotonic IRF, it is usually estimated by the item-ability regression method.

When IRFs from three estimation methods are very similar at the nonmonotonic part, it provides convincing evidence that the IRF of this item is nonmonotonic. When there is monotonic IRF from one or two estimation methods and the other nonmonotonic IRFs from two or one estimation method have a very small nonmonotonic part, these two or one nonmonotonic estimated IRFs might be different from the true IRF which might be monotonic. In this situation, it might be incorrect to state that the IRF of this item is nonmonotonic because the nonmonotonic IRF might be caused by inaccurate estimation.

Separate items identified with nonmonotonic IRFs based on the theta range where nonmonotonicity appears. Items identified with nonmonotonic IRFs can be separated into three categories based on the theta range where the largest nonmonotonicity appears: the theta range where the largest nonmonotonicity appears is at the low theta range (-3 to 1), the theta range where the largest nonmonotonicity appears is at the middle theta range (-1 to 1), and the theta range where the largest nonmonotonicity appears is at the high theta range (1 to 3).

Table 17 summarizes the number of items identified with nonmonotonic IRFs by at least one method on five assessments at each category.

Table 17
Number of Items Identified with Nonmonotonic IRFs by at Least One Method at Three Categories on Five Assessments

Theta Range	Math A	Math B	Math C	Reading A	Reading B	Sum	Percent
Low	10	15	22	6	6	59	88.06%
Middle	1	1	3	0	0	5	7.46%
High	0	0	3	0	0	3	4.48%
Sum	11	16	28	6	6	67	100%

Most items (88.06%) are at the low theta range category. Only five out of 67 items are at the middle theta range category and only three out of 67 items are at the high theta range. The nonmonotonicity of some math items identified with nonmonotonic IRFs is at the middle or high theta range. But the nonmonotonicity of all reading items identified with nonmonotonic IRFs is at the low theta range.

The items identified with nonmonotonic IRFs at the low theta range category affect the test score of students at the low theta range. However, the number of students at the very low theta range is not large. The largest nonmonotonic parts of some IRFs of items at this category appear at the very low theta range. Moreover, only some IRFs from the three estimation methods of these three items are nonmonotonic and one or two IRFs of these items are monotonic. This kind of nonmonotonicity might be caused by inaccurate estimation and is different from the true IRFs which might be monotonic. Since three nonparametric methods are easily affected by the sample size, the small sample size at the very low theta range leads to the inaccurate estimation. The nonmonotonicity of these items do not affect the majority of students and the consequence of the nonmonotonicity on test fairness and comparability is not significant. On the other hand, there are some items at this category with three IRFs that are very similar at the largest nonmonotonic part and the theta range where the largest nonmonotonic part appears covers most part of the low theta range. The reasons for and the consequences of this kind of nonmonotonicity should be investigated because items identified with this kind of nonmonotonic IRFs can affect a large number of students. Item 71 on the reading assessment A is used as an example to investigate the reasons for and consequences of this kind of nonmonotonicity. This item asks student the meaning of a word appearing in the paragraph. Three options of item 71 are very similar in meaning. C is the correct answer but a number of students with low proficiency levels chose A and B. This indicates that students understood the context but they did not really understand the differences of meanings among these three options. Thus, their probability of success is lower than the probability of guessing. This might be the reason for the nonmonotonicity and how it affects the proficiency level estimation. The nonmonotonicity of this item can lead to the estimated proficiency levels of students with higher proficiency levels lower than their true proficiency levels. Figure 87 presents the distracter analysis results of item 71 on the reading assessment A using the nonparametric smooth regression method.

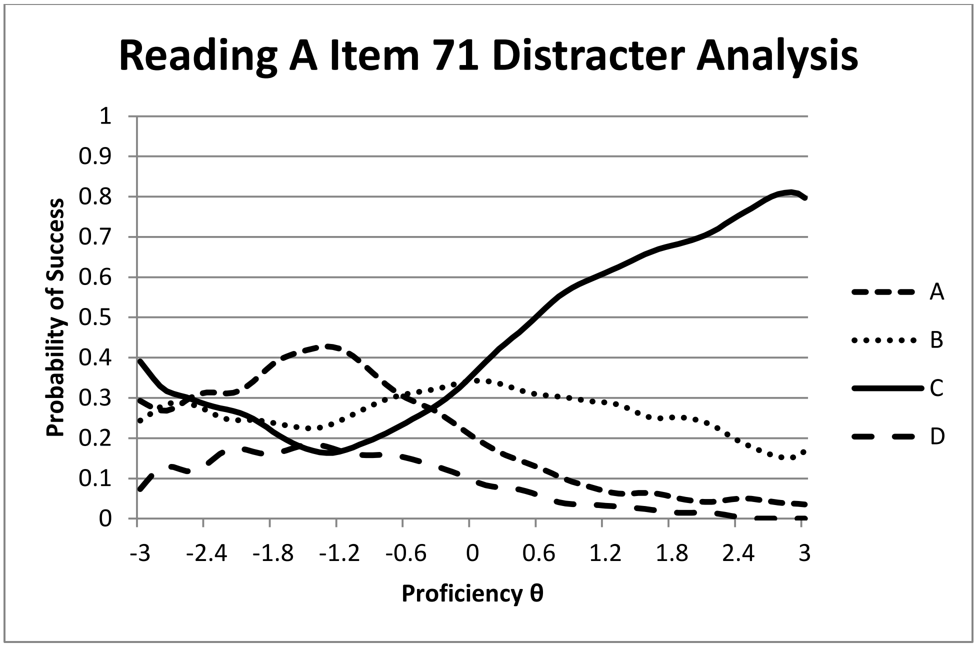

Figure 87. Distracter analysis results of item 71 on the reading assessment A using the nonparametric smooth regression method. Each line is the IRF of each option estimated by the nonparametric smooth regression method.

C is the correct options. But the probabilities that students with low proficiency levels chose option A and B are higher than the probability of guessing.

There are a large number of students at the middle and high theta range. Thus, the items identified with nonmonotonic IRFs at the middle and high theta range category might lead to many students with lower proficiency levels receiving a higher score. Item 7 on the math assessment C provides an example of these two categories to investigate the reasons for and consequences of the nonmonotonicity. Item 7 on the math assessment C is a difficult item because it requires students to find a maximum number in an applied problem. There are two numbers (for example 60 and 0.08) both in the questions and options. Most students with middle and low proficiency levels just chose two incorrect options, A and B. A is "≤ 60.8" and B is "≥60.8". This pattern indicates that they did not know how to solve the problem. Thus, they used another approach to solve this problem. Students just chose the options which are the combination of two numbers in the questions (60.8). Many students with middle proficiency levels chose option A (≤ 60.8) because they understood that they had to find a less than inequality sign because the problem provides an upper bound and asks for the maximum number smaller than the bound. A number of students with lower proficiency levels chose option B (≥60.8) because they did not understand the questions and the term “more than” appears in the question. This might be the reason for the nonmonotonicity of this item at the low and middle theta range, which leads to the estimated proficiency levels of many students lower than their true values. The IRF of this item also indicates that students do not guess at random when they do not know how to solve the problem. Figure 88 presents the distracter analysis results of item 7 on the math C assessment using the B-spline nonparametric IRT method.

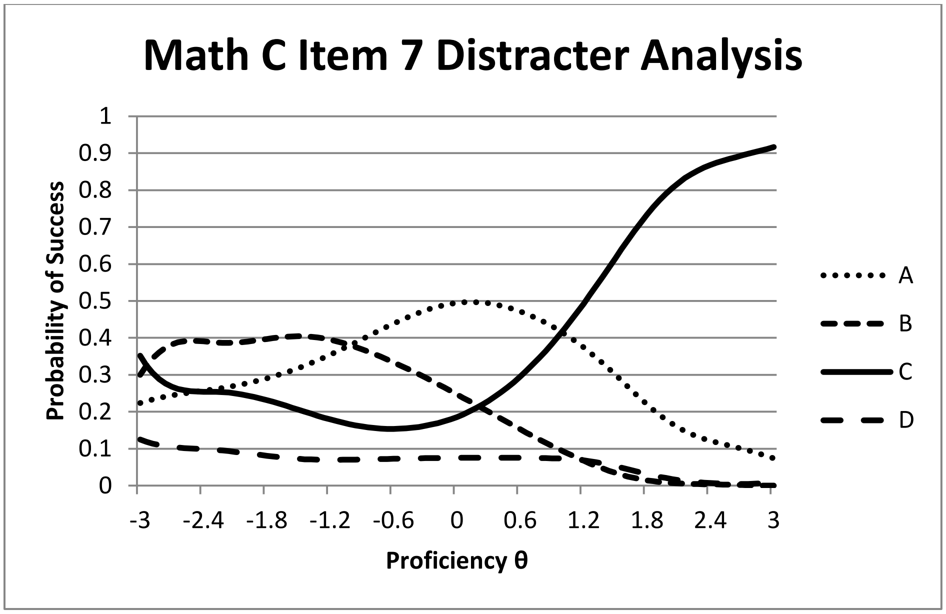

*Figure 88.*Distractor analysis results of item 7 on the math assessment C using the B-spline nonparametric IRT method. Each line is the IRF of each option estimated by the B-spline nonparametric IRT method.

Option C is the correct choice. But option A has a very high probability of success at the middle theta range and option B has a very high probability of success at the low theta range.

Real items identified with nonmonotonic IRFs on the math and reading assessments. Most items identified with nonmonotonic IRFs have one large nonmonotonic part. There are some nonmonotonic IRFs have one large and other very small nonmonotonic parts. But these nonmonotonic parts might be caused by unsmooth estimation methods. Both math and reading assessments have items identified with nonmonotonic IRFs by at least one method. Table 18 summarizes the number of items and the number of items identified with nonmonotonic IRFs by at least one method on each assessment.

Table 18
Number of Items and Number of Items Identified with Nonmonotonic IRFs by at Least One Method on Five Assessments

	Math A	Math B	Math C	Reading A	Reading B
No. of Items Identified with Nonmonotonic IRFs	11	16	28	6	6
No. of Items	83	84	84	80	81
Percent	13.25%	19.05%	33.33%	7.50%	7.40%

Table 18 shows that every math assessment has more items identified with nonmonotonic IRFs than every reading assessment.

Table 19 includes the number of items and the number of items identified with nonmonotonic IRFs by at least one method for each indicator on math assessments.

Table 19

Number of Items and Number of Items Identified with Nonmonotonic IRFs by at Least One Method for each Indicator on Math Assessments

Indicator	No. of Items Identified with Nonmonotonic IRFs	No. of Items	Percent
M.10.4.2.K5	10	20	50.00%
M.10.2.3.A2	7	18	38.89%
M.10.3.3.A1	6	18	33.33%
M.10.3.4.K6	7	21	33.33%
M.10.1.4.A1	4	15	26.67%
M.10.4.2.A1	4	15	26.67%
M.10.1.3.A1	3	15	20.00%
M.10.2.2.A2	3	18	16.67%
M.10.4.1.K3	2	12	16.67%
M.10.4.2.K4	2	12	16.67%
M.10.2.3.K6	2	15	13.33%
M.10.2.2.K3	2	18	11.11%
M.10.3.4.K4	2	18	11.11%
M.10.1.2.K3	1	24	4.17%
M.10.3.1.A1	0	12	0.00%

Indicator M.10.4.2.K5 has the highest percentage (50.00%) of number of items identified with nonmonotonic IRFs. Items aligned to this indicator require students to analyze a scatter plot and make predictions based on the line of best fit. Ten items aligned to this indicator and identified with nonmonotonic IRFs are examined carefully to investigate the reasons for the nonmonotonicity. The nonmonotonicity of three items appears at the very low theta range. Moreover, only some IRFs from the three estimation methods of these three items are nonmonotonic and one or two IRFs of these items are monotonic. This kind of nonmonotonicity might be caused by inaccurate estimation and the consequence of the nonmonotonicity on test fairness and comparability is not significant. The other seven items can be separate into three types based on the reason for nonmonotonicity. Item 27 on math B, item 24 on math C, and item 50 on math C belong to the first type and ask students estimate points that lie on the line of best fit which is not shown but can be modeled based on the scatter plot. Regardless of whether the question of these items includes the term "line of best fit," students with low or low and middle proficiency levels chose the option consistent with the estimation based on pattern of the data on the scatter plot rather than the line of best fit which should be modeled first. These three items ask student to estimate the y-axis value of a given value on x-axis. Students with low or low and middle proficiency levels just chose the option that is the smallest number larger than all the plotted y-values whose correspond x-values are smaller than the given x-value based on the data on the scatter plot. Beside this reason, another reason might also lead to the nonmonotonicity of item 50 on math C is that the y-value of the correct option was not marked according to the scale on the y-axis. Yet one of the distracters contained a value marked on the y-axis and larger than all the plotted y-values whose correspond x-values are smaller than the given x-value. Thus, many students in the low and middle theta range chose this option. Consequently, the probability of success of students who chose these distractors is lower than the guessing probability and affects their proficiency level estimation. Figures 89, 90 and 91 present the distracter analysis results of item 27 on the math assessment B, item 24 on the math assessment C and item 50 on the math assessment C using the item-ability regression method.

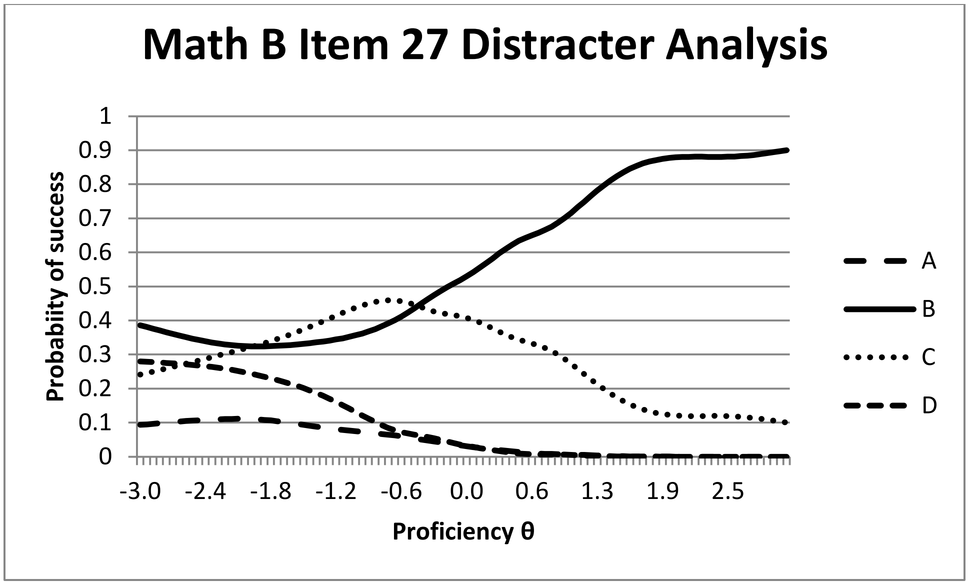

*Figure 89.*Distractor analysis results of item 27 on the math assessment B using the item-ability regression method. Each line is the IRF of each option estimated by the item-ability regression method.

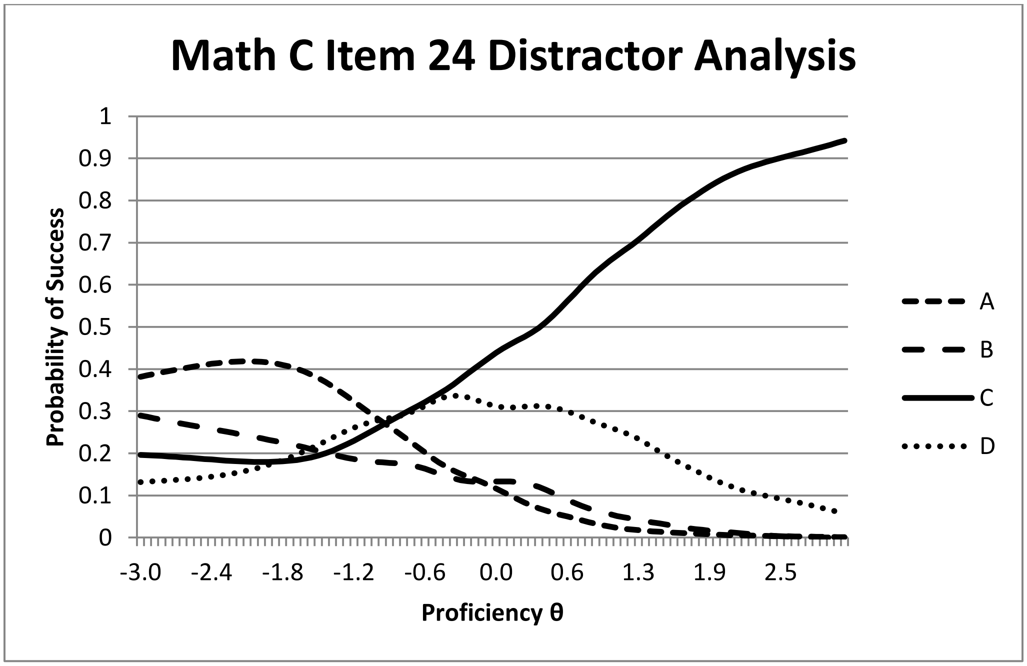

*Figure 90.*Distractor analysis results of item 24 on the math assessment C using the item-ability regression method. Each line is the IRF of each option estimated by the item-ability regression method.

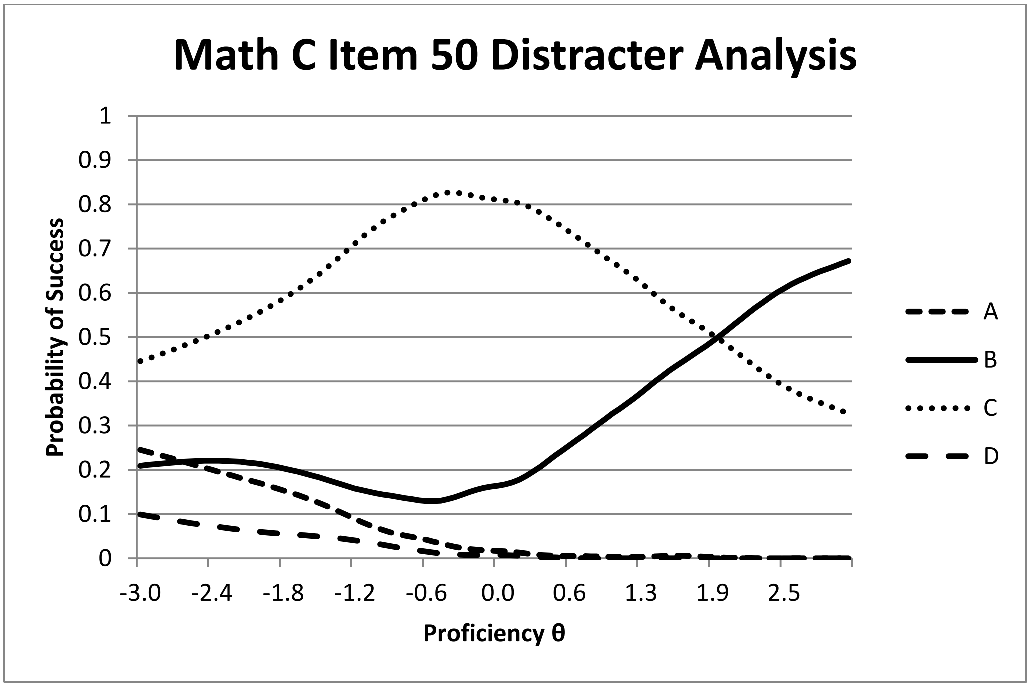

*Figure 91.*Distracter analysis results of item 50 on the math assessment C using the item-ability regression method. Each line is the IRF of each option estimated by the item-ability regression method.

For item 27 on the math assessment B, option B is the correct choice. But option C has a very high probability of success at the low theta range. For item 24 on the math assessment C, option C is the correct choice. But the probability of success of option D is increasing as the proficiency level increases at the low and middle theta range. For item 50 on the math assessment C, option B is the correct choice. But option C has a very high probability of success at the low and middle theta range.

The second type of items identified with nonmonotonic IRFs and aligned to indicator M.10.4.2.K5 asks student to choose the equation of the line of best fit. Among 20 items aligned to this indicator, these two items (item 69 on math A and item 53 on math C) are the only two items asking the equation of the line of best fit. The nonmonotonicity of these two items appears at the low theta range and at this range the probabilities of success of other three distractors are higher than the probability of success of the correct option. However, the probability of success of one distractor is increasing as the proficiency level increases at the low or low and middle theta range as Figure 92 and 93 present. The equation of this distractor has the same slope direction as the equation of the correct option but either different slope value for item 69 on math A or different slope value and interception value for item 53 on math C. The distractor analysis might indicate that student with low proficiency levels understood the concept of the direction of the slope but not the value of slope. For item 69 on math A, the slope value of the equation of the distracter with increasing probability of success at the low theta range is equal to the unit of the scale. Student who chose this distracter might be distracted by the unit of scale and their probability of success is lower than the probability of guessing. For item 53 on math C, the y-value of the intercept of the equation of the correct option is not marked according to the scale on the y-axis but the y-value of the intercept of the equation of the distractor with increasing probability of success at the low and middle theta range is marked according to the scale on the y-axis. Student who chose this distracter might be distracted by the value of the interception and their probability of success is lower than the probability of guessing. Figures 82 presents the distracter analysis results of item 69 on the math assessment A using the B-spline nonparametric IRT method and figure 93 presents item 53 on the math assessment C using the item-ability regression method.

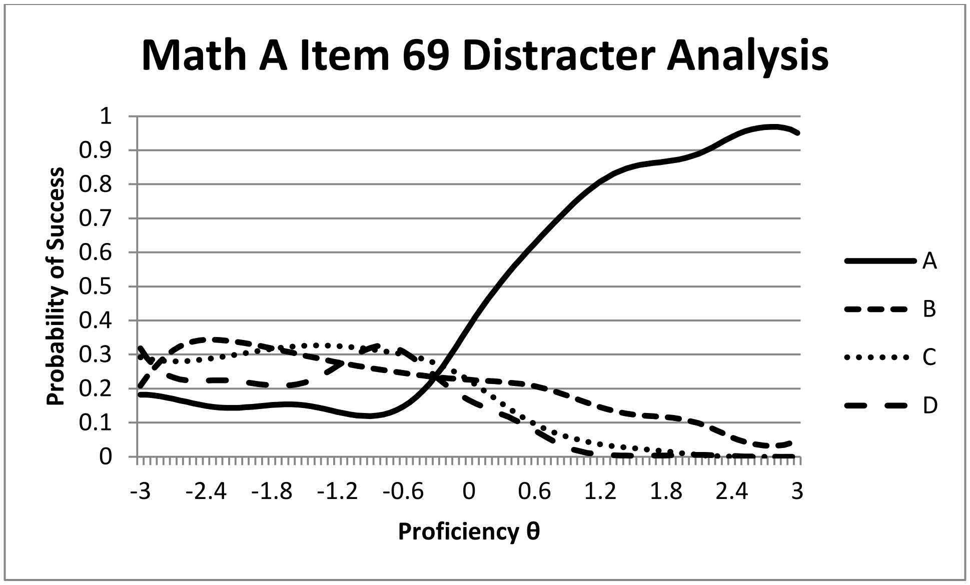

*Figure 92.*Distracter analysis results of item 69 on the math assessment A using the B-spline nonparametric IRT method. Each line is the IRF of each option estimated by the B-spline nonparametric IRT method.

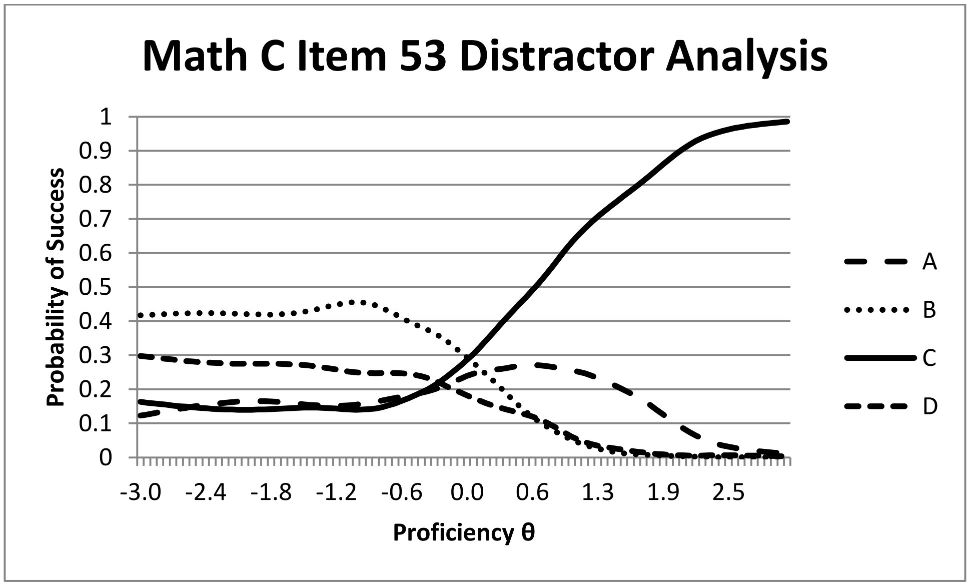

*Figure 93.*Distracter analysis results of item 53 on the math assessment C using the item-ability regression method. Each line is the IRF of each option estimated by the item-ability regression method.

For item 69 on the math assessment A, option A is the correct choice. But the probability of success of option D is increasing as the proficiency level increases at the low theta range. For item 53 on the math assessment C, option C is the correct choice. But the probability of success of option A is increasing as the proficiency level increases at the low and middle theta range.

The nonmonotonicity of the third type of items identified with nonmonotonic IRFs and aligned to indicator M.10.4.2.K5 appears at the high theta range. The nonmonotonic parts of the two items of this type (item 9 and item 57 on math C) are very small as Figure 59 and 60 present. The math content experts, who is familiar with math curriculum and involves in math item development, commended these two items are good items from the content perspective and the reasons for the nonmonotonicity might be the careless of the students with high proficiency levels. Since the nonmonotonic parts are small, they are not able to affect the test result significantly.

Indicator M.10.3.3.A1 also has a high percentage (33.33%) of number of items identified with nonmonotonic IRFs. This indicator requires students to analyze the impact on the perimeter, area, or volume of a geometric figure when one of these measurements is increased or decreased by a given factor. In order to answer this kind of items correctly, students need to understand the computation of and relationships among distance, area, and volume. Students with proficiency levels from about -3 to -1 only chose the options that are the multiples or divisors of the factor, indicating they were able to compute the product or quotient of rational numbers. They chose options indicating that they had multiplied or divided by a linear factor, but they did not apparently understand how to correctly use the linear factor to describe corresponding changes to area or volume. This might be the reason for the nonmonotonicity of this kind of items. The nonmonotonicity caused by this reason affects the proficiency level estimation of students. For students with higher proficiency levels, their estimated proficiency levels might be lower than their true proficiency levels. For example, item 57 on math assessment A, item 41 on math assessment B, and item 80 on math assessment C, require students to calculate effects on volume given a linear factor of change. Students often selected the option corresponding to the square of the factor or the factor itself instead of the cube of the factor. For these three items, the estimated IRFs have a similar shape.

Table 20 includes the number of items and the number of items identified with nonmonotonic IRFs by at least one method for each indicator on reading assessments.

Table 20
Number of Items and Number of Items Identified with Nonmonotonic IRFs by at Least One Method for each Indicator on Reading Assessments

Indicator	No. of Items Identified with Nonmonotonic IRFs	No. of Items	Percent
R.11.2.1.1	4	10	40.00%
R.11.2.1.2	3	10	30.00%
R.11.1.4.7	2	12	16.67%
R.11.1.3.1	1	8	12.50%
R.11.1.4.14	1	8	12.50%
R.11.1.4.8	1	11	9.09%
R.11.1.3.3	0	12	0.00%
R.11.1.3.4	0	12	0.00%
R.11.1.4.10	0	12	0.00%
R.11.1.4.11	0	8	0.00%
R.11.1.4.15	0	8	0.00%
R.11.1.4.2	0	8	0.00%
R.11.1.4.5	0	12	0.00%
R.11.1.4.6	0	12	0.00%
R.11.1.4.9	0	8	0.00%
R.11.2.1.3	0	10	0.00%

Indicator R. 11.2.1.1 has the percentage (40.00%) of number of items identified with nonmonotonic IRFs. The reading content expert, who is familiar with reading curriculum and involves in reading item development, commented that this kind of items is not straightforward and requires inference and interpretation of the entire passage which is a higher comprehensive skill. This might be the reason for the nonmonotonicity. More studies are needed for the reasons for the nonmonotonicity of the items for this indicator.

Limitations and Future Studies

There are several limitations and suggestions for future studies. These limitations and suggestions are discussed based on the order of studies: simulation study limitations and real data study limitations.

Simulation study limitations. One limitation of the simulation study is that the evaluation nodes of the item-ability regression method were affected by the software BILOG-MG. Although the pre-determined quadrature nodes, which are the same as the evaluation nodes of the other two methods, were input to BILOG-MG code, the quadrature nodes became different after running BILOG-MG because adjust command was used (which adjusts the scale of empirical ability distribution). These different nodes were used to estimate IRFs nonparametrically. The different nodes added a variable to the comparison among three nonparametric methods since IRFs were estimated on different nodes. Therefore, it is hard to determine if the differences between the item-ability regression method and the nonparametric smooth regression method or the B-spline nonparametric IRT method are caused by the method differences itself or the evaluation nodes differences. In order to solve this problem, “noadjust” command could be used to run BILOG-MG.

Additional limitation is related to the models and item parameters used for generating nonmonotonic data. The nonmonotonic model used in the simulation study generated IRFs with very high probabilities at the low theta range. But the real nonmonotonic IRFs have low probabilities at the low theta range. This nonmonotonic model might be different from the real situation. Moreover, the change of guessing parameters did not affect the nonmonotonic IRF estimation of three methods very much when the nonmonotonic model (equation 36) was used. It is hard to use this model to study the relationship between the guessing parameters and the nonmonotonic IRF estimation. In addition, in the simulation study extreme item parameters were used to generate nonmonotonic data. However, in the real situation there are few items with nonmonotonic IRFs and with extreme item parameters. Using extreme item parameters might lead to the comparison among three nonparametric methods lack of generalization. Thus, more generalized item parameters and other different nonmonotonic models should be used in future studies for method comparison.

Real data study limitations. First limitation of real data study is that the small sample size at the very low theta range leads to inaccurate estimation at this range. Since there are a small number of students at the very low theta range and the nonmonotonicity at this range does not affect the test result a lot, the evaluation nodes ranging from -2.5 to 3 can be chosen for future studies because the large population can lead to more accurate IRF estimation.

Another limitation is that there are only a few items being examined on the reasons for and consequences of the nonmonotonicity. The nonmonotonic parts of the estimated IRFs of items examined are large. But there are some items with estimated IRFs with one or several small nonmonotonic parts. The reasons for and consequences of these items should also be studied. If the consequences are not significant, these items should not be rejected because of the nonmonotonicity.

The third limitation is that the relationship between the indicator and the nonmonotonicity was not studied enough. The conclusion will be more convincing if all the items identified with nonmonotonic IRFs for one indicator are investigated on the reasons for and consequences of the nonmonotonicity as for indicator M.10.4.2.K5.

The fourth limitation is that the reasons for that there are more math items identified with nonmonotonic IRFs than reading items identified with nonmonotonic IRFs are not clear. The further studies should focus on the reasons for this difference based on the investigation of the reasons for and consequences of all the math and reading items identified with nonmonotonic IRFs. Moreover, the collaboration work of both math and reading content experts is needed because the content experts should compare the reasons for and consequences of the nonmonotonicity between the math and reading items together.

Conclusion

This study first compared the nonparametric smooth regression method, the item-ability regression method and the B-spline nonparametric IRT method using simulated data. The estimation results of three methods are very similar and indicate three methods can identify the nonmonotonic IRF equally well. The item-ability regression method and B-spline nonparametric method worked better compared with the nonparametric smooth regression method because their type I and II error rates are lower. Subsequently, these three methods were used to estimate the IRFs of real items on five assessments. Each assessment has several items identified with nonmonotonic IRFs. The math assessment has more items identified with nonmonotonic IRFs than the reading assessment. Most nonmonotonicity appears at the low theta range. Some items identified with nonmonotonic IRFs can lead to students with higher proficiency levels receiving lower scores and these items should be modified in future. In order to avoid nonmonotonicity, item writer could write items with more specific questions, less distractions in the graph, and mark correct response and distractors all on the y-axis for item with graph based on the investigation of the reasons for and consequences of the nonmonotonicity. To summarize, looking the sample of real items used in this study, nonmonotonic IRFs are not rare and can affect the fairness and comparability of the test score. The nonmonotonicity should be checked before applying parametric logistic models.

References

Azzalini, A., Bowman, A. W., & Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika, 76*(1), 1.

Bayarri, M., & Berger, J. O. (2000). P values for composite null models. *Journal of the American Statistical Association, 95*(452), 1127-1142.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (eds.), Statistical Theories of Mental Test Scores, (p. 397-472), Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 3*7, 29-51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26*(4), 381.

Copas, J. (1983). Plotting p against x. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 32*(1), 25-31.

Craven, P., &Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik, 31*(4), 377-403.

De Ayala, R. J. (2008). *The theory and practice of item response theory.* Location: Guilford Press.

De Boor, C. (2001). *A practical guide to splines*. Location: Springer.

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, 62(1), 7-28.

Douglas, J. (1999). *Asymptotic identifiability of nan-parametric item response models* (Technical Report No. 142). University of Wisconsin, Department of Biostatistics and Medical Informatics.

Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement, 25*(3), 234.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis.* New York, NY: Chapman & Hall.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733-759.

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement, 27*(3), 217.

Glöckner-Rist, A., & Hoijtink, H. (2003).The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, 10, 544-565.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society. Series B (Methodological), 29*(1), 83-100.

Hambleton, R., & Han, N. (2004, April). *Assessing the fit of IRT models: Some approaches and graphical displays*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Hambleton, R. K., & Swaminathan, H. (1984). *Item response theory: Principles and applications*. Boston: Kluwer Academic.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* CA: Sage.

Jacod, J., & Protter, P. E. (2003). *Probability essentials.*New York, NY: Springer Verlag.

Junker, B. W.,& Sijtsma, K. (2001). Nonparametric IRT in action: An overview of the special issue. *Applied Psychological Measurement*, 25, 211–220.

Kingston, N. M., &Dorans, N. J. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement, 9*(3), 281.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33*(7), 519.

Lord, F. (1952). *A theory of test scores*. (Psychometric Monograph No. 7). Iowa City, IA: Psychometric Society.

Lord, F. M. (1970). Item characteristic curves estimated without knowledge of their mathematical form—a confrontation of Birnbaum's logistic model. *Psychometrika, 35*(1), 43-50.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: L. Erlbaum Associates.

Lord, F. M., Novick, M. R., & Birnbaum, A. (Eds.). (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McKinley, R. L., & Mills, C. N. (1985).A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*(1), 49.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton.

Molenaar, I. W. (2001). Thirty years of nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 295.

Nadaraya, E. (1964). On estimating regression. *Theory of Probability and Appication., 9*, 141-142.

Nürnberger, G. (1989). Approximation by spline functions. New York, NY: Springer.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*(4), 342.

Ramsay, J. (1996). A geometrical approach to item response theory. *Behaviormetrika, 23*(1), 3-16.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611-630.

Ramsay, J. O., & Silverman, B. W. (Ed.).(1997). *Functional data analysis*. New York, NY: Springer.

Rossi, N., Wang, X., & Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of Educational and Behavioral Statistics, 27*(3), 291.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12*(4), 1151-1172.

Samejima, F. (1979). A new family of models for the multiple-choice item (Research Rep. No. 79-4). University of Tennessee, Department of Psychology.

Sijtsma, K., & Junker, B. W. (2006). Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika, 33*(1), 75-102.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, 42(4), 375-394.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*(4), 298.

Stout, W. (2001). Nonparametric Item Response Theory: A maturing and applicable measurement modeling approach. *Applied Psychological Measurement, 25*(3), 300.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*(4), 501-519.

Watson, G. S. (1964). Smooth regression analysis. *Sankhy : The Indian Journal of Statistics, Series A, 26*(4), 359-372.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika, 71*(2), 281-301.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*(2), 245.