Are All Item Response Functions Monotonically Increasing?

Wenhao Wang and Neal Kingston
Center for Educational Testing and Evaluation University of Kansas

Abstract

Common parametric logistic IRT models assume logistic form, which is monotonically increasing. If true item response functions of some real items are nonmonotonic, examinees with lower proficiency levels might receive higher scores.

This study checked whether the nonparametric smooth regression method could detect nonmonotonic IRFs accurately using simulated data. In addition, this method was used to identify items with nonmonotonic IRFs on real assessments.

Results show that the nonparametric smooth regression method can detect nonmonotonic IRFs and that one assessment has some items with nonmonotonic IRFs.

Investigation of possible reasons for and consequences of nonmonotonicity are presented for one item 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.

Purposes

- Check whether nonparametric smooth regression method can detect the nonmonotonic IRF accurately using simulated data.
- Check the existence of the real items with nonmonotonic IRFs.

Methods

Simulation Data. The responses of 10,000 examinees on a simulated 60 item test were generated. Six of 60 items were simulated to have non-monotonic form and the other 54 items had IRFs that were three -parameter logistic (3PL). A total of 500 replications are generated. A null distribution of the nonmonotonic area was used to evaluate whether the estimated IRFs were monotonic. Type I error and power rates were calculated.

Real Data. A random sample of 10,000 high school examinees (grades 9 to 11) who responded to all 84 items on a mathematics assessment were selected. The IRFs of these items were estimated using the nonparametric smooth regression method. The area of the nonmonotonic IRF was calculated and is compared to the corresponding posterior distribution to determine the extent of monotonicity which is measured by the posterior predictive p value (PPP-value). The PPP-value is equal to the probability that the nonmonotonic area of the real item is bigger than the values in the posterior distribution. In this study if the PPP-value of one item whose IRF estimated by the nonparametric smooth regression method is less than 0.05, then this item was identified as the item with nonmonotonic IRF.

Practical Consequences. Distractor analysis and item content analysis were conducted together to investigate the reasons for and the consequences of the nonmonotonicity.

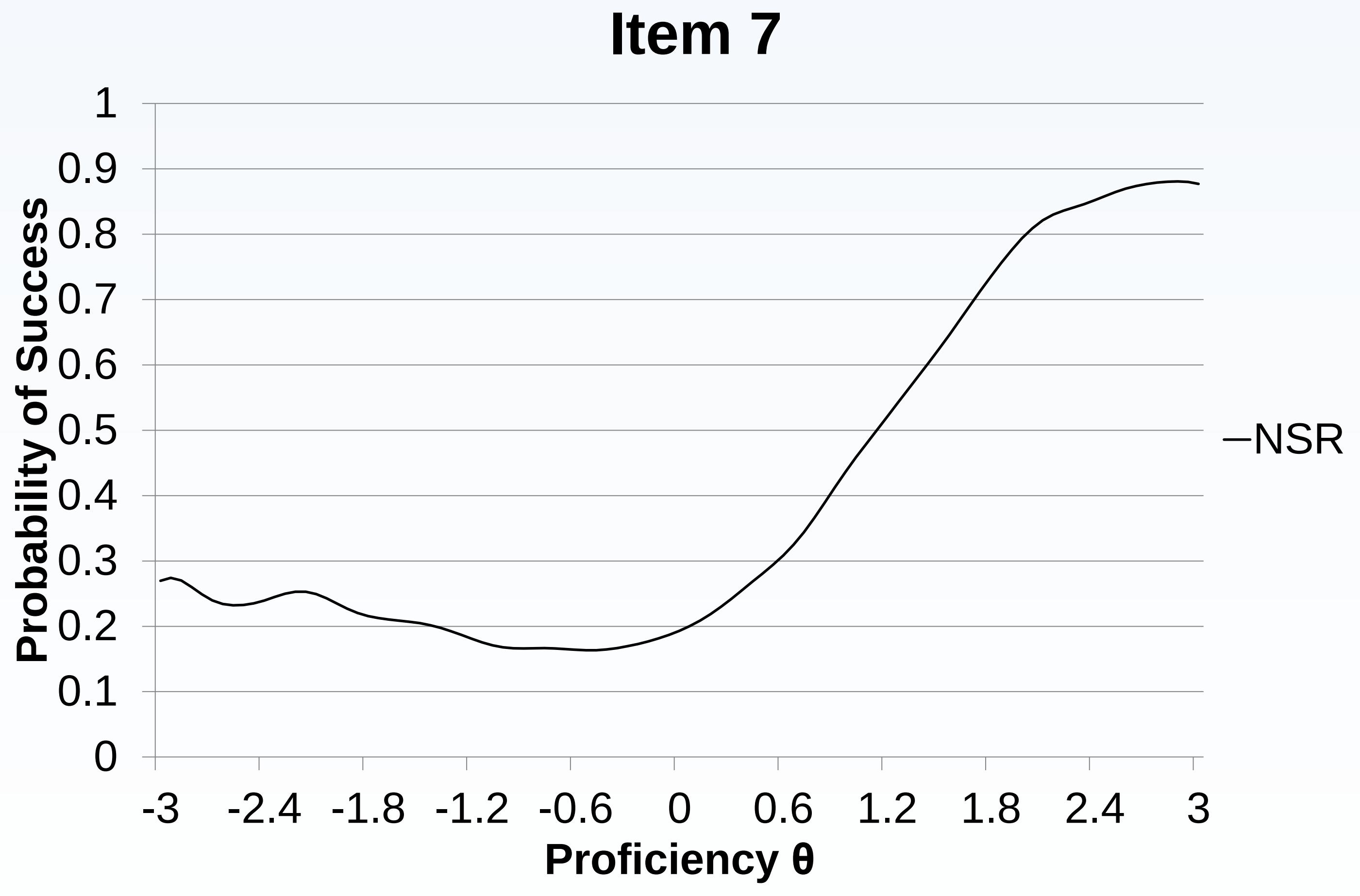
Results

The type I error rate is 0.051. With 10,000 examinees the average power rate is 0.502.

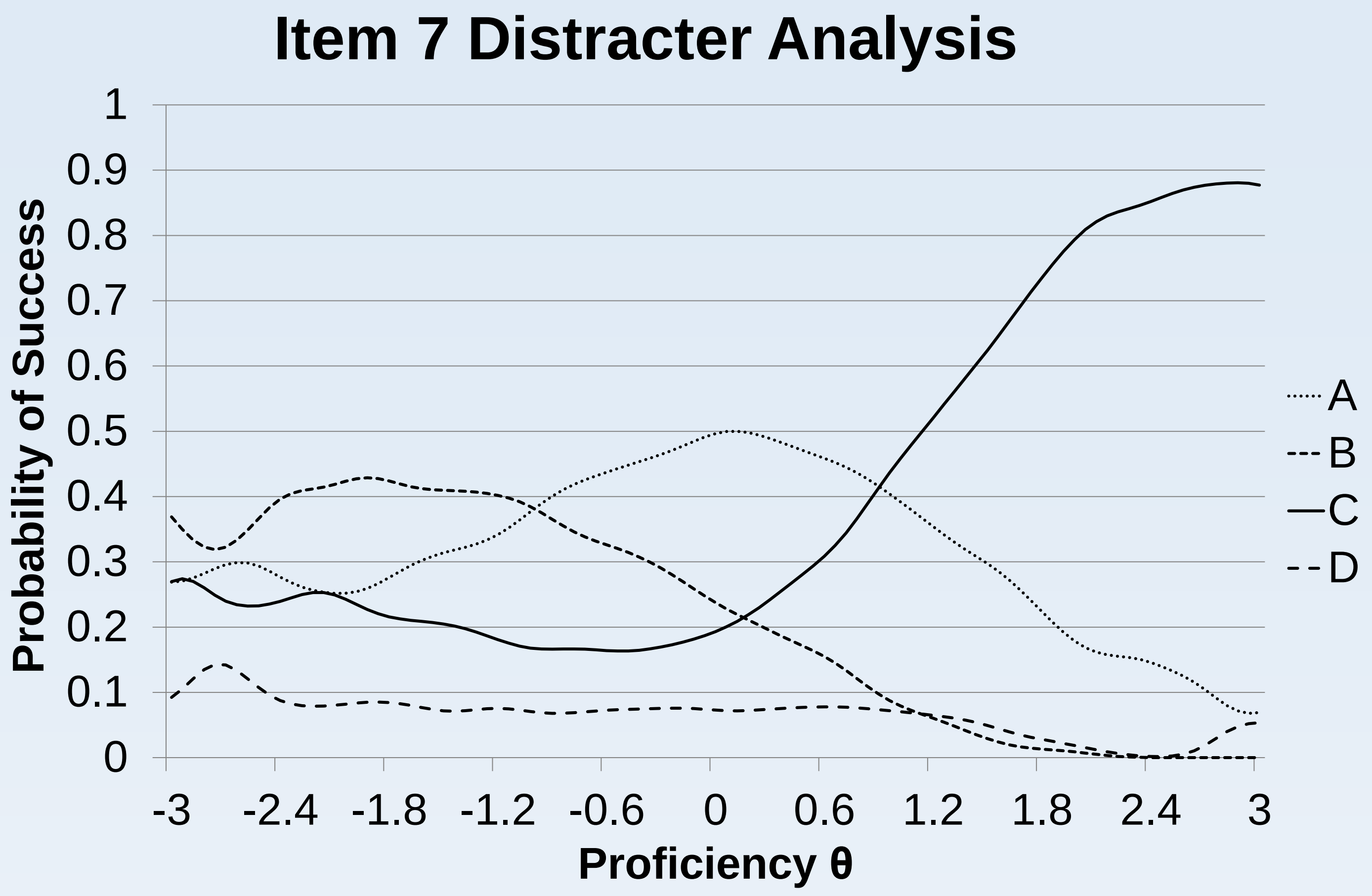
True Item Parameters and Power Rates for Nonmonotonic IRFs

Item	a	b	c	Power
10	0.855	1.654	0.139	0.184
20	0.695	-2.416	0.336	0.600
30	0.369	0.863	0.163	0.372
40	1.800	0.123	0.295	0.254
50	0.833	-1.509	0.089	0.806
60	0.784	-0.841	0.522	0.794

Fourteen items out of 84 math items are identified as having nonmonotonic IRFs using the nonparametric smooth regression method.



The PPP-value of item 7 is 0.044. The nonmonotonic parts of the items 7 started in the low theta range and ended in the middle theta range.



Conclusion

The result of this study provided evidence that the nonparametric smooth regression could be used as a tool to assess IRF non-monotonicity.

Looking at one real test, nonmonotonic IRFs are not rare. Items with distractors that function well in the middle of the score range may cause non-monotonicity. Some items identified with nonmonotonic IRFs can lead to students with higher proficiency levels receiving lower scores.

Since use of parametric IRT models for estimating theta might increase estimation error or add bias, nonmonotonicity should be checked before applying parametric logistic models. Also, Distracter analysis to investigate which distracter is selected most at the ability range where the non-monotonic IRF happens might be useful to avoid this issue.

KU THE CENTER FOR EDUCATIONAL TESTING & EVALUATION
The University of Kansas

Are All the Item Response Functions Monotonically Increasing?

Wenhao Wang

Neal Kingston

Center for Educational Testing and Evaluation

University of Kansas

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 checked whether the nonparametric smooth regression method could detect the nonmonotonic IRF accurately using simulated data. In addition, this method was used to identify items with nonmonotonic IRFs on real assessments. Results present that the nonparametric smooth regression method can detect the nonmonotonic IRF and the math assessment has some items with nonmonotonic IRFs. Investigations on the reasons for and the consequences of the nonmonotonicity were conducted for one item 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.

Are All the Item Response Functions Monotonically Increasing?

Purpose

Most item response models assume the item response function (IRF) have a logistic form with certain parameters. These models are called the parametric logistic IRT models and the IRF with a logistic curve is monotonically increasing. However, when Lord (1970) estimated item response functions with a nonparametric approach some items had local nonmonotonicity. Similarly, in their work on item-ability regressions, Kingston and Dorans (1985) discovered an item type that led to nonmonotonic results. Since if not properly modeled nonmonotonicity could lead to examinees with lower proficiency receiving higher scores, this issue could lead to reduced score validity. Therefore, this study focused on the assumption that the IRF with a logistic curve is monotonically increasing using both simulated and real data.

The statistical tests to check the model fit, for example the Pearson chi-square test (Yen, 1981), can only answer the global yes/no question about whether the model fits the data or not and cannot assess whether certain assumption is met or not. The model fit checking methods focusing on the monotonicity assumption should be able to estimate the true IRF from the observed data without assuming monotonicity of the IRF. If an item has an obvious nonmonotonic IRF estimated from these methods, it presents that this item violates assumption that the logistic curve is monotonically increasing. The nonparametric smooth regression method (Douglas & Cohen, 2001; Ramsay, 1991) is a "relatively easy-to-implement nonparametric regression method" and an "exploratory tool for assessing IRF monotonicity" (Junker & Sijtsma, 2001, p231). The first purpose of this study was to check whether nonparametric smooth regression method can detect the nonmonotonic IRF accurately using simulated data.

Once the IRF of a real item is estimated by the nonparametric smooth regression method, one additional step should be conducted to judge whether the area of the nonmonotonic IRF part is large enough to conclude that the parametric logistic IRT models are misfit. The Posterior Predictive Model Checking (PPMC) method (Sinharay, Johnson, & Stern, 2006) will be used to serve this purpose. Thus, in order to identify items with nonmonotonic IRFs, this study used nonparametric smooth regression method to estimate the IRF and PPMC method to judge the extent of monotonicity.

Hambleton, Swaminathan and Rogers (1991) stated that it is always helpful to evaluate the practical consequences of model misfit. Once there are some real items with nonmonotonic IRFs, the insightful analysis will be conducted to investigate the practical consequences of the nonmonotonic IRFs.

Method

Subjects. The participants in this study are a random selected sample of 10,000 high school students (grade 9 to 11) who took an 84-item four-choice summative mathematic assessment in spring 2010.

Simulation Data. The responses of 10,000 examinees on a simulated 60 item test were generated. Six of 60 items were simulated to have a non-monotonic IRF with the mathematic form and the other 54 items have IRFs followed the three parameter logistic (3PL) model. The 10,000 ability levels were randomly drawn from a standard normal distribution. The item parameters for the nonmonotonic model were selected purposely and the item parameters for the monotonic model were randomly selected from the real item parameter estimates of the high school summative mathematic assessment. A total of 500 replications are generated. For each replication, the IRF of every item is estimated by the nonparametric smooth regression method and the nonmonotonic area was calculated. This area is compared with the null distribution of the nonmonotonic area 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. This null distribution was generated from 500 replicated data simulated from the 3PL model using the true item parameters. Since the true situation was known, the Type I error and power rates of the nonparametric smooth regression method to detect the nonmonotonic IRF were calculated from 500 replications.

Real Data. The IRFs of 84 real items were estimated by the nonparametric smooth regression method. The area of the nonmonotonic IRF was calculated and is compared to the corresponding posterior distribution to determine the extent of monotonicity which is measured by the posterior predictive p value (PPP-value). The posterior distribution was generated from 500 replicated data. These 500 replicated data were simulated from the 3PL model using the draws from the posterior distributions of item and proficiency parameters. These posterior distributions were simulated through the MCMC algorithm using WinBUGS. The prior distributions used for MCMC algorithm were: $\log(a_i) \sim N(0,2)$, $b_i \sim N(0,2)$, $c_i \sim Beta(5,17), \theta_j \sim N(0,1)$. The PPP-value is equal to the probability that the nonmonotonic area of the real item is bigger than the values in the posterior distribution. 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). In this study if the PPP-value of one item whose IRF estimated by the nonparametric smooth regression method is less than 0.05, then this item was identified as the item with nonmonotonic IRF.

Practical Consequence. The procedures of investigating the practical consequences of the item with non-monotonic IRF might include:

1. Asking content experts to look at the question and answer choices to investigate the item writing.
2. Conducting distracter analysis to investigate which distracter is selected most at the ability range where the non-monotonic IRF happens.

Results

The type I error rate of the nonparametric smooth regression method is 0.051. The power rate of the nonparametric smooth regression method is 0.502. The type I error rate is very low but the power rate is not very high for this method. Table 1 indicated the true item parameters and the power rates for six items with true nonmonotonic IRFs. 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 model and the extreme item parameters being used for generating data.

Table 1

True Item Parameters and Power Rates for Nonmonotonic IRFs

Item	a	b	c	Power
10	0.855	1.654	0.139	0.184
20	0.695	-2.416	0.336	0.600
30	0.369	0.863	0.163	0.372
40	1.8	0.123	0.295	0.254
50	0.833	-1.509	0.089	0.806
60	0.784	-0.841	0.522	0.794

The power rates are low for items 50 and 60. These two either have low or high item guessing parameters but middle range item discriminate and difficulty parameters. On the other hand, the power rates are relatively high for other four items especially items10 and 40. These items either have extreme item discriminate parameters or extreme item difficulty parameters. This result indicates the nonparametric smooth regression method 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 this method significantly for this nonmonotonic data.

Fourteen items out of 84 math items are identified with nonmonotonic IRFs by the nonparametric smooth regression method. Table 2 includes the PPP-values for these 14 items on this assessment.

Table 2

PPP-Values for 14 Items on Math Assessment

No.	Item	Nonparametric Smooth Regression
1	2	0.012
2	7	0.044
3	9	0.034
4	20	0.022
5	24	0.046
6	25	0.038
7	29	0.018
8	31	0.022
9	34	0.000
10	49	0.020
11	53	0.002
12	56	0.038
13	80	0.010
14	81	0.006

Figure 1-4 present the estimated IRFs of the item 2, 7, 25, and 80 on the summative mathematic assessment. These items have nonmonotonic IRFs estimated by the nonparametric smooth regression method.

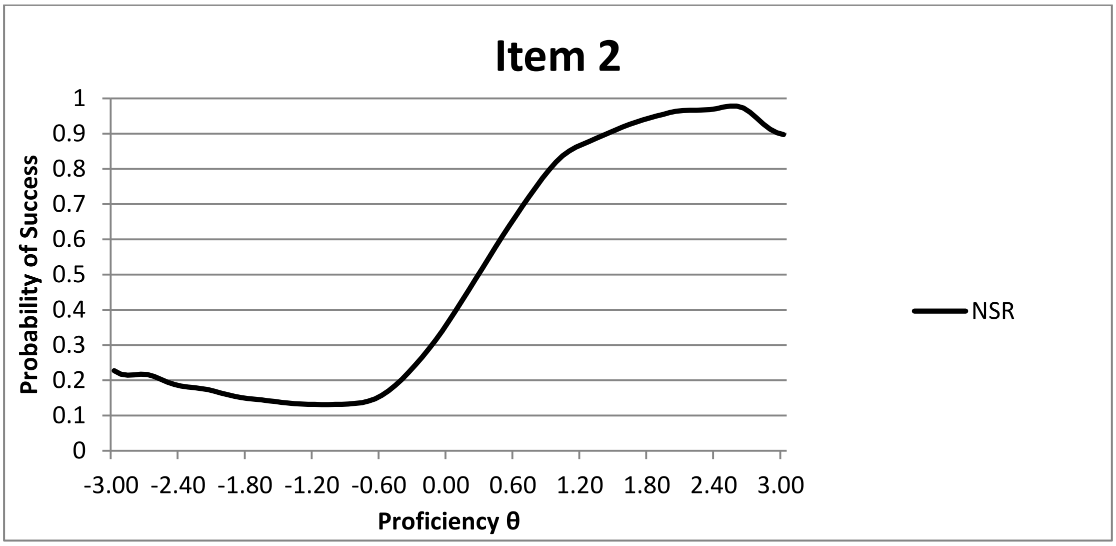


Figure 1. IRFs estimated by nonparametric smooth regression method for item 2. The PPP-value of item 2 is 0.012.

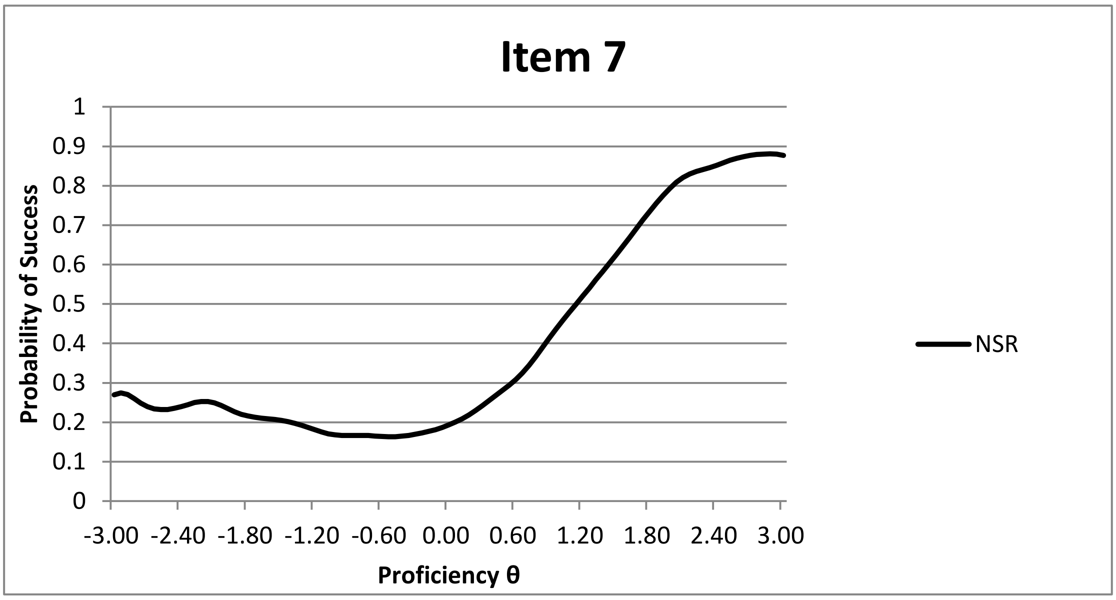


*Figure 2.*IRFs estimated by nonparametric smooth regression method for item 7. The PPP-value of item 7 is 0.044.

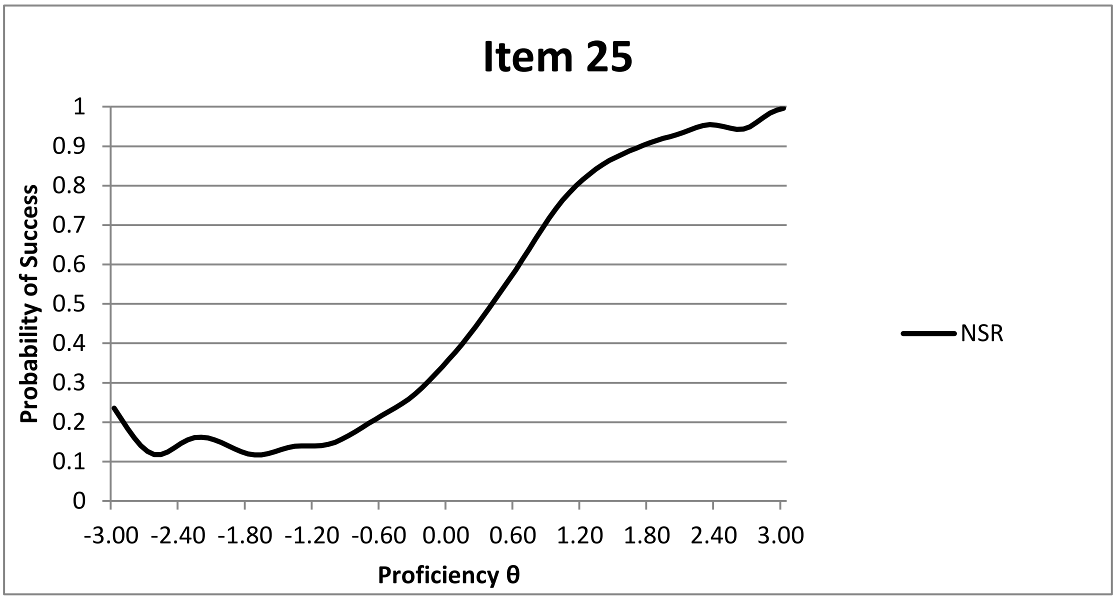


Figure 3. IRFs estimated by nonparametric smooth regression method for item 25. The PPP-value of item 25 is 0.038.

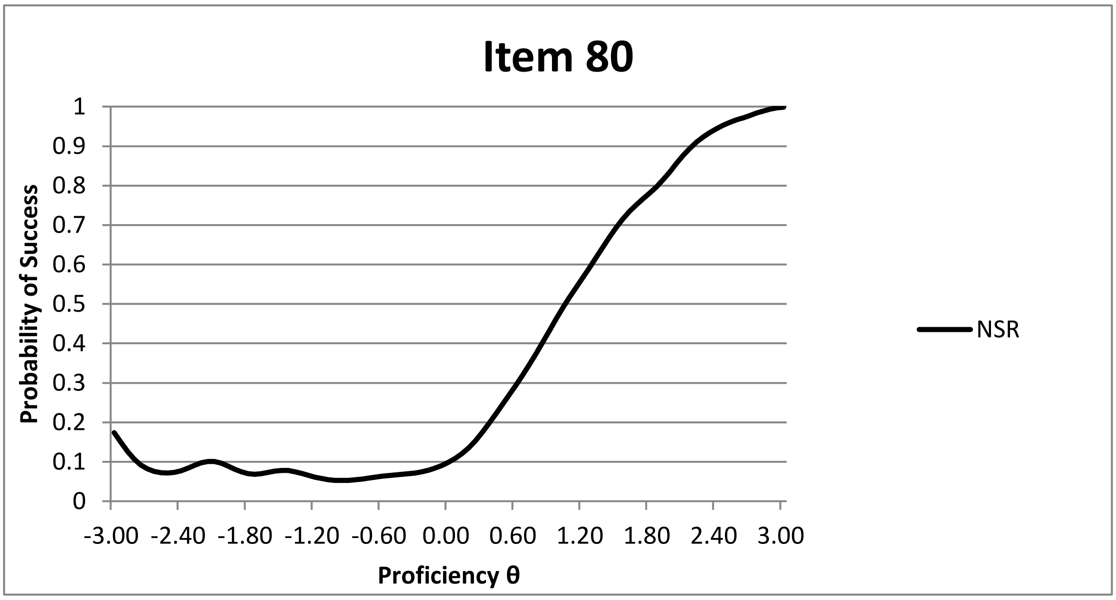


Figure 4. IRFs estimated by nonparametric smooth regression method for item 80. The PPP-value of item 80 is 0.010.

For these four items, the nonmonotonic parts all started in the low theta range. The nonmonotonic parts of the items 2 and 25 ended in the low theta range. The nonmonotonic parts of the items 7 and 80 ended in the middle theta range. There are a large number of students at the middle theta range. Thus, the items identified with nonmonotonic IRFs at the middle theta range might lead to many students with lower proficiency levels receiving a higher score. Moreover, the nonmonotonic part of item 7 is very big and this item was studied on the reasons for and the 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 5 presents the distracter analysis results of item 7 using the nonparametric smooth regression method.

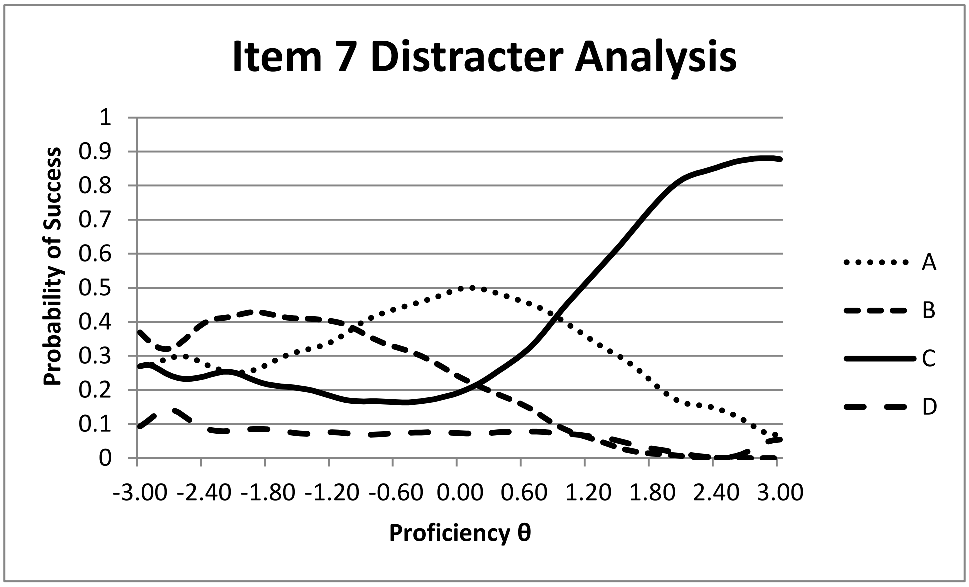


*Figure 5.*Distractor analysis results of item 7 using the nonparametric smooth regression method. Each line is the IRF of each option estimated by the nonparametric smooth regression method.

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.

Conclusion

The misfit of parametric logistic IRT models caused by the violation of the assumption that the IRF is monotonically increasing has not been investigated yet. The result of this study provided evidence that the nonparametric smooth regression could be used as a tool to assess this problem. Looking the sample of real items used in this study, nonmonotonic IRFs are not rare. 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 based on the investigation of the reasons for and consequences of the nonmonotonicity. To summarize, the nonmonotonicity should be checked before applying parametric logistic models.

References

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

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

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

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.

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.

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

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

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