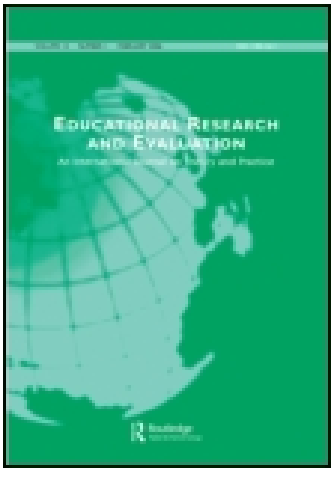

Educational Research and Evaluation: An International Journal on Theory and Practice

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/nere20

An empirical examination of IRT information for school climate surveys

Lun Mo [a] , Fang Yang [b] & Xiangen Hu [c]

[a] Memphis City Schools , TN, USA

[b] University of Detroit Mercy , Detroit, MI, USA

[c] The University of Memphis , Memphis, TN, USA

Published online: 21 Jun 2011.

To cite this article: Lun Mo , Fang Yang & Xiangen Hu (2011) An empirical examination of IRT information for school climate surveys, Educational Research and Evaluation: An International Journal on Theory and Practice, 17:1, 33-45, DOI: 10.1080/13803611.2011.583033

To link to this article: http://dx.doi.org/10.1080/13803611.2011.583033

Educational Research and Evaluation
Vol. 17, No. 1, February 2011, 33–45

Routledge
Taylor & Francis Group

An empirical examination of IRT information for school climate surveys

Lun Mo[a]*, Fang Yang[b] and Xiangen Hu[c]

[a]*Memphis City Schools, TN, USA;* [b]*University of Detroit Mercy, Detroit, MI, USA;* [c]*The University of Memphis, Memphis, TN, USA*

(*Received 17 November 2010; final version received 23 March 2011*)

School climate surveys are widely applied in school districts across the nation to collect information about teacher efficacy, principal leadership, school safety, students' activities, and so forth. They enable school administrators to understand and address many issues on campus when used in conjunction with other student and staff data. However, these days each district develops the questionnaire according to its own needs and rarely provides supporting evidence for the reliability of items in the scale, that is, whether an individual item contributes significant information to the questionnaire. The *Item Response Theory* (IRT) is a useful tool that helps examine how much information each item and the whole scale can provide. Our study applied IRT to examine individual items in a school climate survey and assessed the efficiency of the survey after the removal of items that contributed little to the scale. The purpose of this study is to show how IRT can be applied to empirically validate school climate surveys.

Keywords: school climate survey; item information; IRT; ICC; school district

Introduction

School climate surveys assess the perception of teacher efficacy, principal leadership, school safety, students' activities, and so forth. When the information collected in a climate survey is used with educational outcomes such as achievement test scores, attendance rates, and behavioral problems, many interesting questions can be addressed and reported to district administrators and schools as evidence to evaluate their policies or to assist their future policy-making. Several studies (Freiberg, 1998; Johnson & Johnson, 1993; Johnson, Johnson, Gott, & Zimmerman, 1997; Kuperminc, Leadbeater, & Blatt, 2001; Kuperminc, Leadbeater, Emmons, & Blatt, 1997; Manning & Saddlemire, 1996) found that a positive school climate is strongly associated with positive educational outcomes of students and school personnel. Similarly, a negative climate can interfere with optimal learning and development. However, a school district usually develops its own survey instrument to explore specific questions, and this instrument is created without systematic examination of its reliability. To confidently use the results from the survey, it is necessary to

*Corresponding author. Email: lun222@gmail.com

ISSN 1380-3611 print/ISSN 1744-4187 online

DOI: 10.1080/13803611.2011.583033
http://www.informaworld.com

examine the instrument first. Otherwise, one may draw misleading conclusions from the results.

One way to address reliability of the instrument is to analyze item and scale information by using *item response theory* (IRT; Bock, 1972; Cronbach, 1960; Hambleton, 1991). The concept of *information* reflects how precisely the parameters are estimated for each item in the survey or for the whole scale of the survey. Basically, the richer the amount of information is, the more precisely the parameters are estimated. If these estimates are precise, the conclusion about the effectiveness of the instrument could be reliable. The most popular IRT models are two- and three-parameter models. The two-parameter model includes two parameters estimating item characteristics: difficulty b and discrimination a. The three-parameter model resembles the two-parameter model but with an additional element of guessing. For the purpose of illustration, we use the two-parameter model to describe the underlying rationale and ignore the factor of guessing.

The formula for item information in the two-parameter IRT model is depicted as follows:

$$I_j(\theta) = a_j^2\{1 + \exp[-a_j(\theta - b_j)]\}^{-1}\{1 - [1 + \exp(-a_j(\theta - b_j))]^{-1}\} \quad (1)$$

where $I_j(\theta)$ indicates the information that item j provides given a competency θ. Parameter b indicates the difficulty of item j, which reflects where the item functions along the competency continuum. The more difficult the item is, the larger the location value is. Parameter a indicates discrimination of item j, which describes how well item j can differentiate between a low-competency group below the item location (difficulty value) and a high-competency group above the item location. From Equation (1), we can see that a, b, and θ decide the magnitude of the information for j. An *item information curve* (IIC) can be plotted by the amount of individual item information against the competency (see Figure 1). We can see that the shape of the information function is defined by discrimination a, and the maximum item

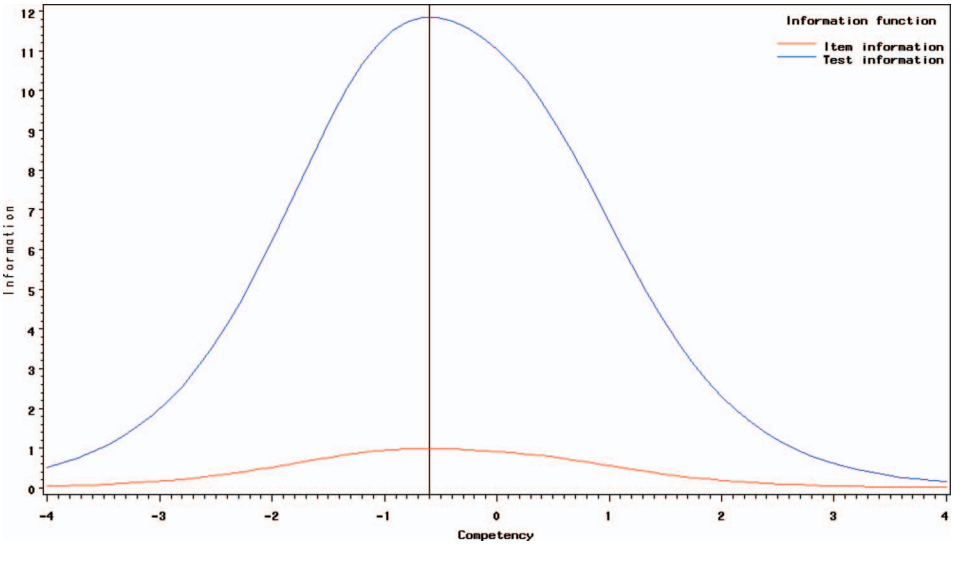

Figure 1. The information function curve.

information is at the point along the competency continuum corresponding to the item's difficulty parameter.

The test information function is the sum of the item information function at θ (Equation (2)). The maximum point of test information function is equal to the θ at which the set of items provides the most accurate measurement.

$$I(\theta) = \sum_{j=1}^{n} I_j(\theta) \tag{2}$$

where $I(\theta)$ is the amount of the scale information at a specific competency level of θ, and n is the number of items in the scale. Generally speaking, the more items there are in a survey, the richer information that survey has, and the more precise the scale would be. However, this is not always the case. The amount of individual item's information can be changed when other items are deleted or added, and furthermore, the item changes may affect the total amount of information in the scale. Plotting of the amount of the scale information against competency yields a graph of the *test information curve* (TIC) in Figure 1.

The maximum test information function is around the center of the competency continuum, and when competency levels approach the extreme, the amount of test information decreases significantly. By analyzing the IIC and TIC, we may find that some items are not informative: That is, the information provided by the whole questionnaire would not change much if these items were removed from the scale. The relative efficiency of the modified questionnaire can be computed as follows:

$$RE(\theta) = I_A(\theta) / I_B(\theta) \tag{3}$$

where $RE(\theta)$ denotes relative efficiency, $I_A(\theta)$ and $I_B(\theta)$ are the information functions for modified scale A and original scale B, respectively, defined over a common competency scale θ.

In addition to considering the amount of information provided by individual items and the whole survey, we also expect school climate surveys to be understandable for most participants with average competency level. Thus, it is necessary to remove items with low discrimination value or extreme difficulty values. The *item characteristic curve* (ICC) helps identify items with these features. For instance, as can be seen in Figure 2, the ICC shows the relationships among item performance, competency, and item characteristic. The performance of three items increases with the increase of competency level. Items 1 and 2 have similar discrimination levels but different difficulty levels, with Item 1 being easier than Item 2. Item 3 has a different discrimination level from Items 1 and 2, but has a similar difficulty level as Item 2.

The combination of ICC, IIC, and TIC curves enables us to evaluate how well an item performs in terms of its relevance or contribution to measuring the underlying construct of competency, and enables us to evaluate the relative efficiency by comparing one scale to another. The results of analyses at both the item and scale levels will finally assist us to select items that match the purposes of the survey. Based on this rationale, the IRT model should be suitable for the purpose of creating a reliable school climate survey. Consequently, the present study was conducted to apply the IRT model to improve school climate surveys. More specifically, to better develop an instrument to investigate campus climates, we applied the IRT model to select a variety of items that could fairly measure the average-level competency of

participants, demonstrated reasonable discrimination value and middle level of difficulty. Furthermore, we expected that those selected items would create an ideal test information curve, which should be flat around mean competency. The purpose of this study is to use Memphis City Schools' (MCS) climate survey as a vehicle to demonstrate that the IRT model provides a powerful tool for the development of climate surveys that are concise, reliable, and targeted toward their study population.

Method

Participants

In the 2007–2008 academic year, 62,043 students in Memphis City Schools (MCS) took the school climate survey. However, only those students who answered 90% of the survey questions were selected for this study. Among these 43,824 selected participants, 48.88% were male and 51.12% were female. In terms of ethnicity, 86.25% were African American, 6.89% were Caucasian, and 6.86% were other.

Instrument

The MCS school climate survey consisted of 43 items, of which 17 were intended to measure teacher efficacy, 24 to measure principal leadership, 3 to measure school safety, and 7 to measure student engagement. The assumption of unidimensionality (a single latent competency variable that fully explains performance) was separately examined for the whole scale and for the items in several different categories (see Table 1) by using statistical software SPSS. The software *Parscale* was used to run the IRT analysis.

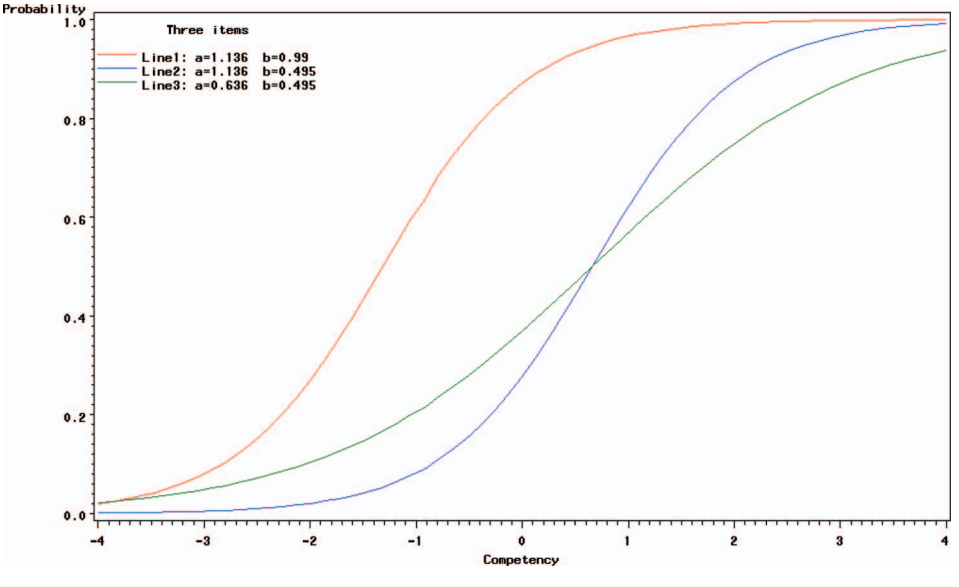

Figure 2. The item characteristic curve (ICC).

Results

SPSS examination on unidimensionality assumption

For the whole scale, the first-factor loading explained 23.4% of the variance and the ratio of first-factor loading to second-factor loading was more than three, indicating that the data met the requirement of unidimensionality. For questions investigating teacher efficacy and principle leadership, the ratios of first-factor loading to second-factor loading were 4.28 and 4.58, respectively, suggesting that they also met the criteria of unidimensionality. For questions measuring school safety and student activities, the ratios were 1.92 and 2.73, respectively, suggesting that they did not meet the requirement of unidimensionality.

Two-parameter grade response model analysis

Most school climate surveys were designed on Likert scales, and the optional response categories were polytomous and ordered. If there are m response

Table 1. Examination on unidimensionality assumption.

Question Category	First-Factor Loading	Second-Factor Loading	First-Factor Loading/Second-Factor Loading
Teacher efficacy	4.24	0.99	4.28
Principal leadership	2.75	0.60	4.58
Safety	1.61	0.84	1.92
Student activities	6.22	2.28	2.73
Total	10.07	3.20	3.15

Table 2. Parameter values estimated from IRT.

Item	Discrimination *a*	Difficulty *b*	Item	Discrimination *a*	Difficulty *b*
01	1.059	–0.954	23	0.721	–3.736
02	0.792	0.053	24	0.620	–1.121
03	0.038	126.581	25	0.567	–3.440
04	1.112	0.017	26	0.644	–3.054
05	0.511	–7.319	27	0.704	–2.635
06	1.141	–0.439	28	1.0	3.907
07	0.739	0.124	29	1.0	5.674
08	0.742	0.259	30	1.0	3.456
09	0.996	–0.980	31	0.935	–0.972
10	1.201	–0.764	32	0.857	0.979
11	1.088	–0.599	33	0.724	–1.319
12	0.116	21.332	34	0.806	–4.883
13	0.972	–1.977	35	0.864	1.503
14	0.990	–0.256	36	1.216	1.166
15	0.752	1.315	37	0.352	–4.1
16	0.665	–4.189	38	0.772	0.037
17	0.199	–4.620	39	1.0	5.132
18	0.590	3.901	40	0.050	70.049
19	0.938	1.840	41	0.329	–3.977
20	0.440	–1.617	42	0.571	–0.411
21	0.383	1.223	43	0.429	0.222
22	0.426	0.184			

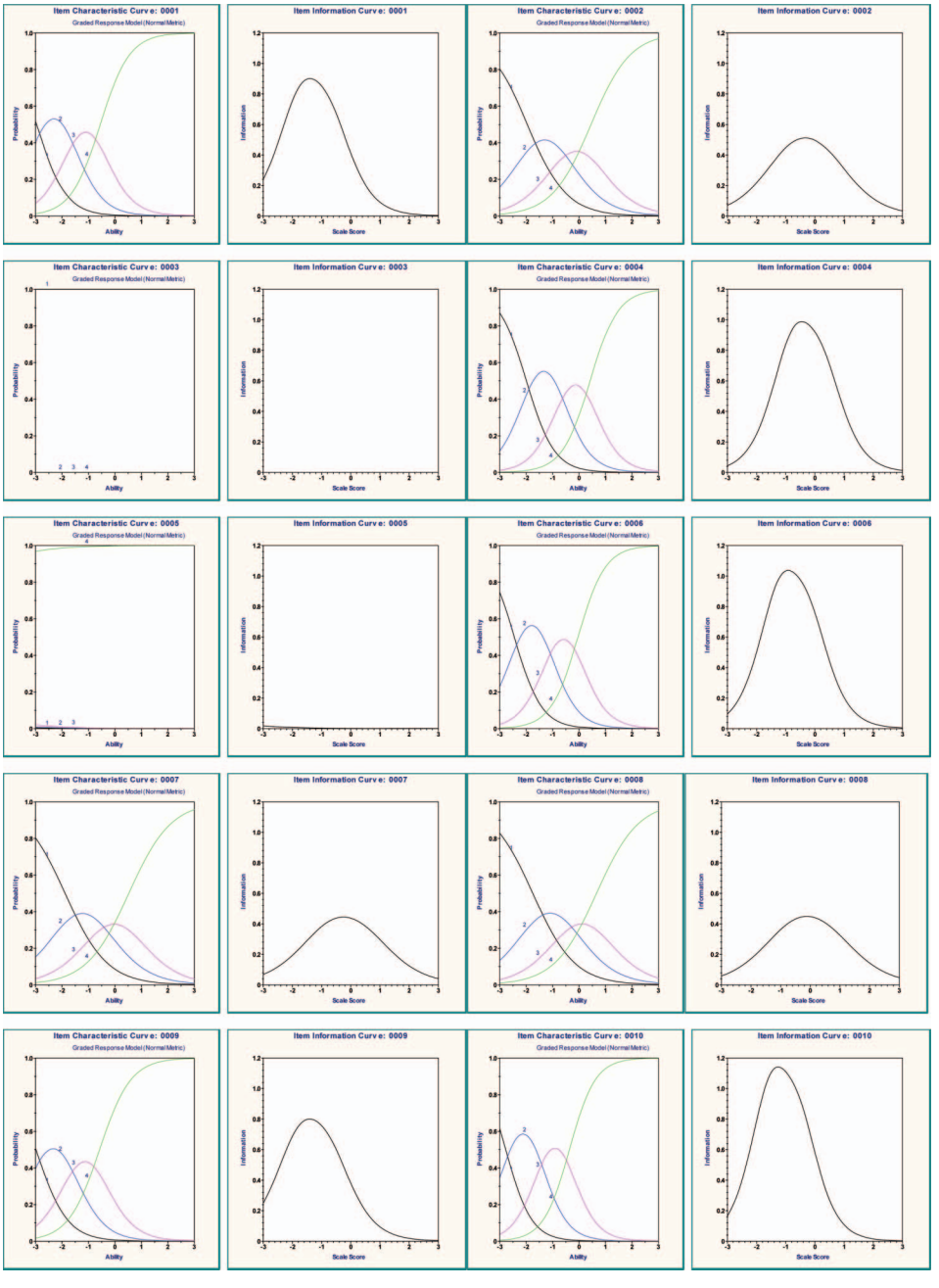

Figure 3. The ICC and IIC for each item in the questionnaire.

categories, then there are $m - 1$ threshold values to distinguish response categories. Accordingly, a graded response model (Samejima, 1969) was applied to analyze the data, which dichotomized the response categories into two overall categories: (1) greater than or equal to score category k; (2) less than score category k. The probability of an item response greater than or equal to score

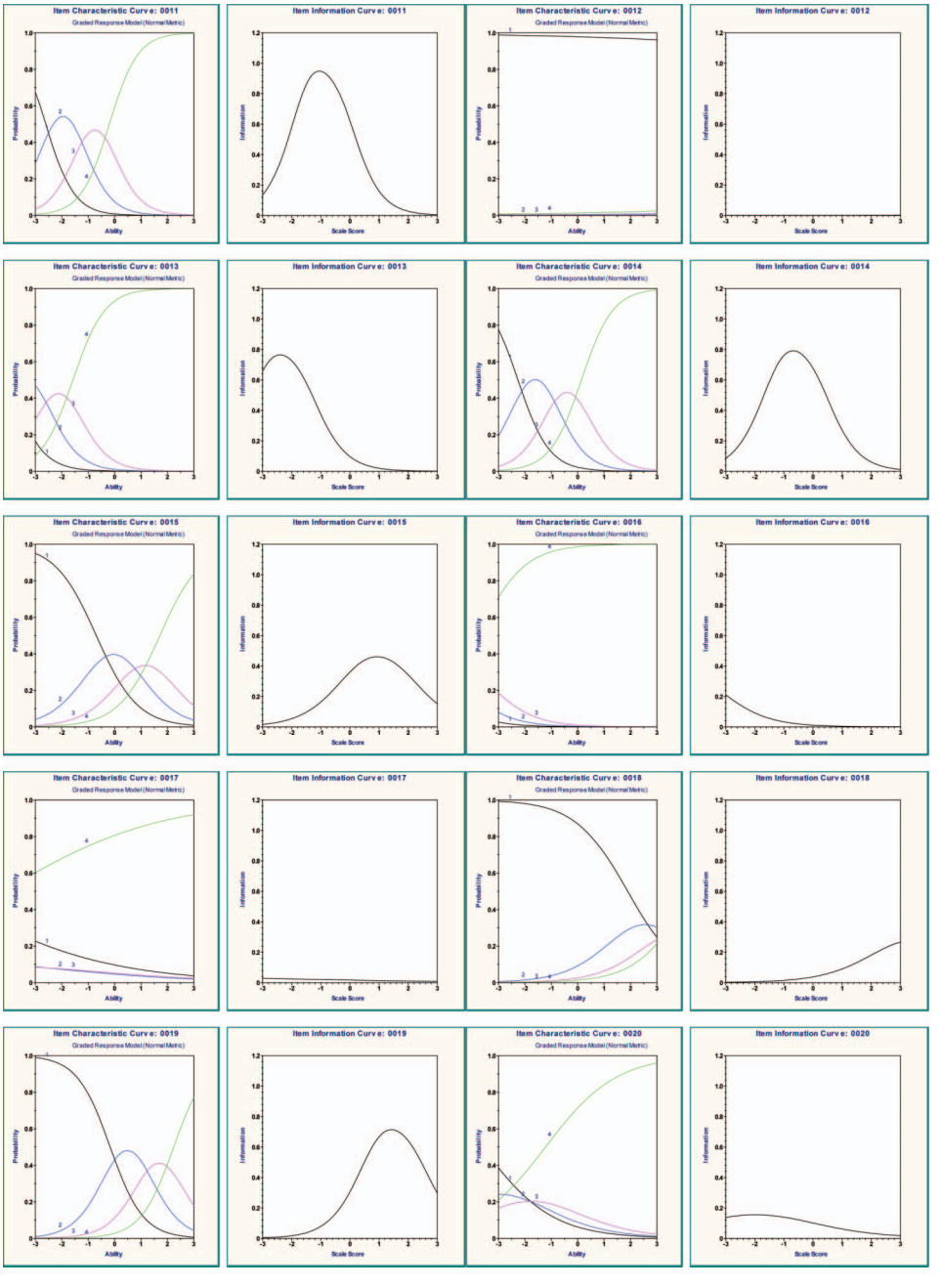

Figure 3. (*Continued*).

category k can then be described by a dichotomous ITR model. On the basis of this assumption, an examinee's probability of choosing a score category k is described by the difference in probabilities for the person having scored greater than or equal to k and having scored greater than or equal to $k + 1$. Under a

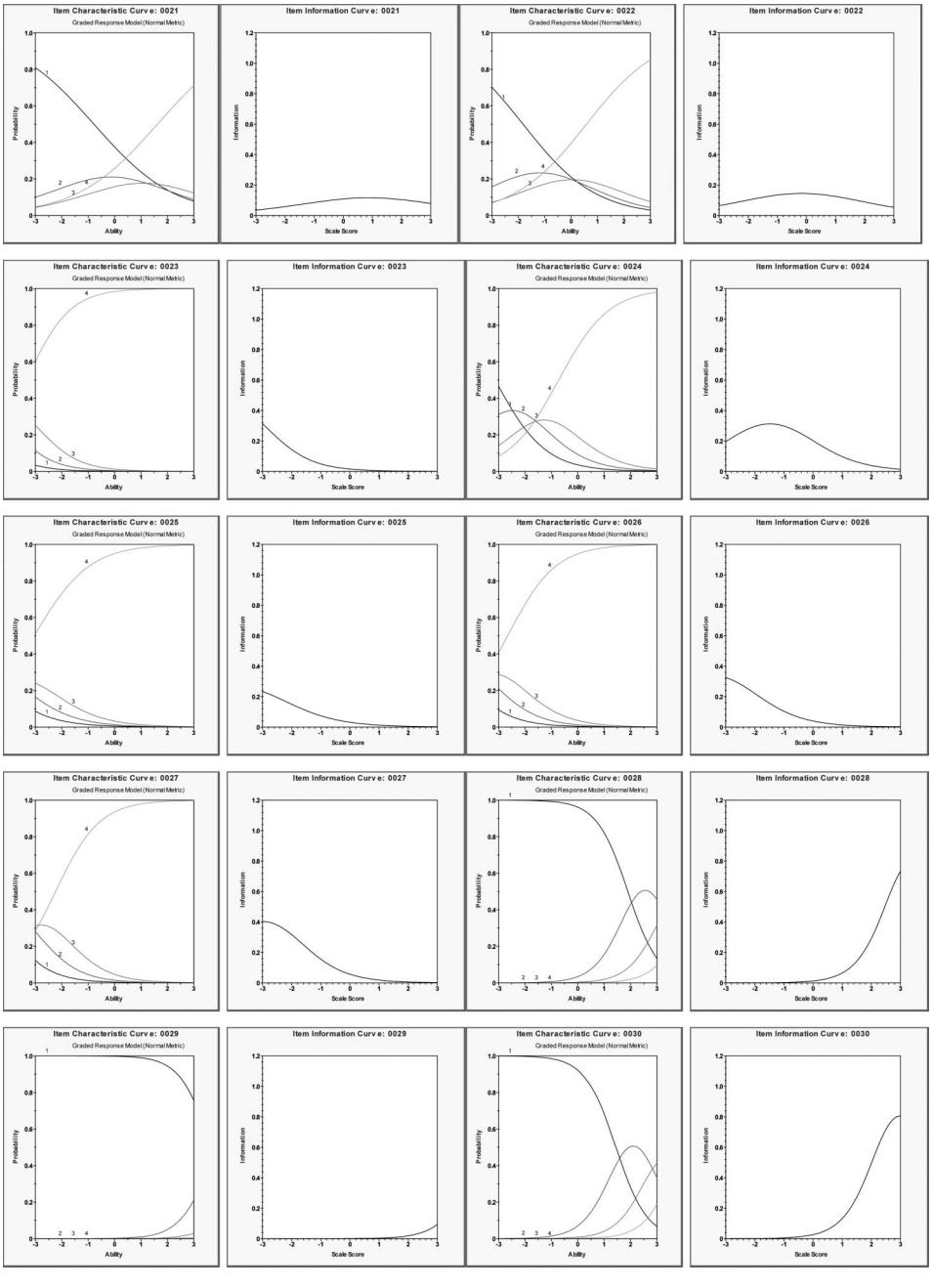

Figure 3. (*Continued*).

graded item response model, each item has a discrimination parameter, a difficulty parameter, and the same $m - 1$ threshold parameters across items. Meanwhile, the ICC and IIC would be depicted for each item; TIC would be derived for the whole scale.

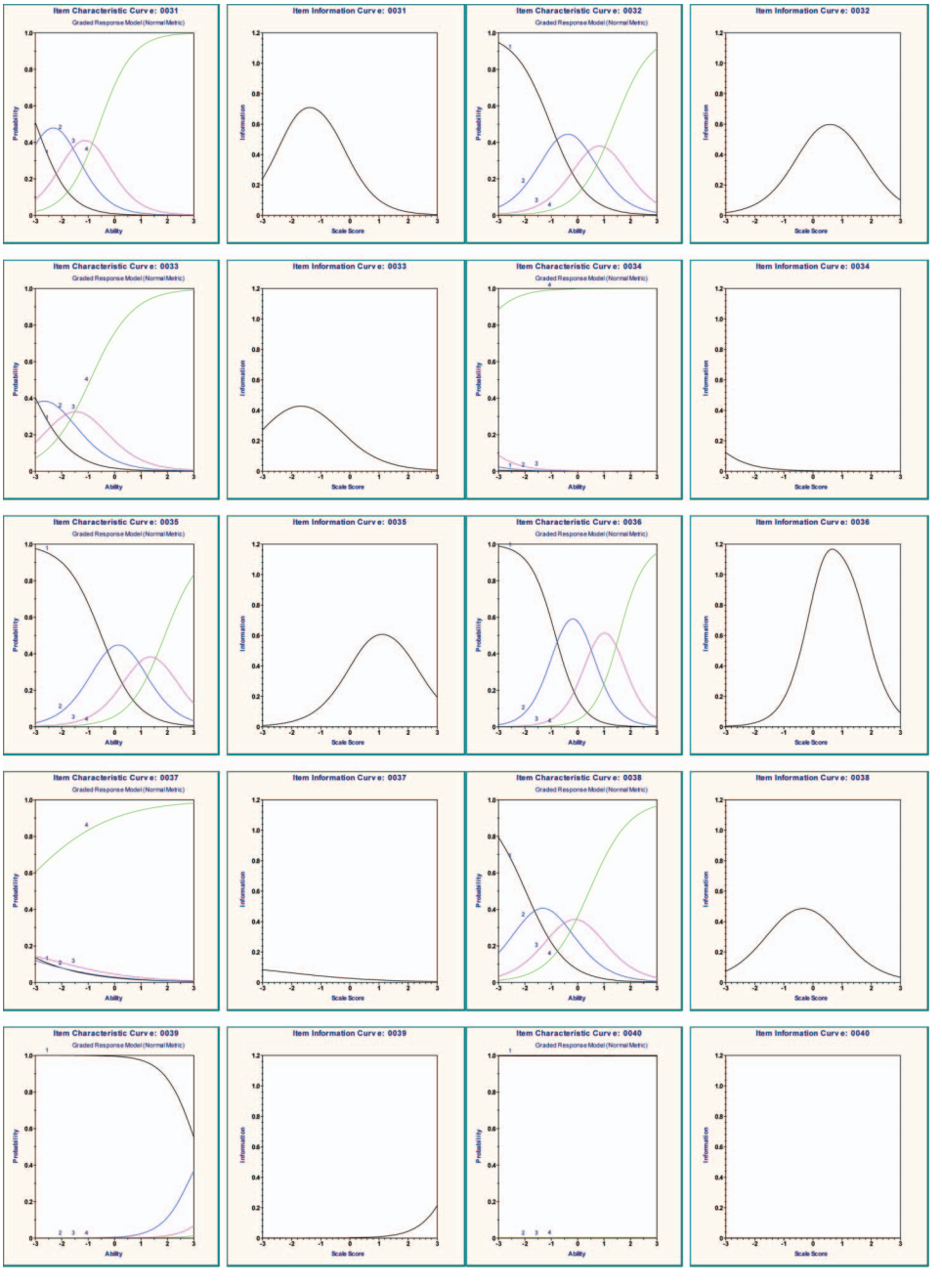

Figure 3. (*Continued*).

Estimated parameter values in IRT

The item discrimination parameter a substantially influences the amount of information available to assess competency. If the value of parameter a is too small, then a cannot provide enough information to estimate the competency. If the value of a is too large, it causes biased estimation. From Table 2, we found that

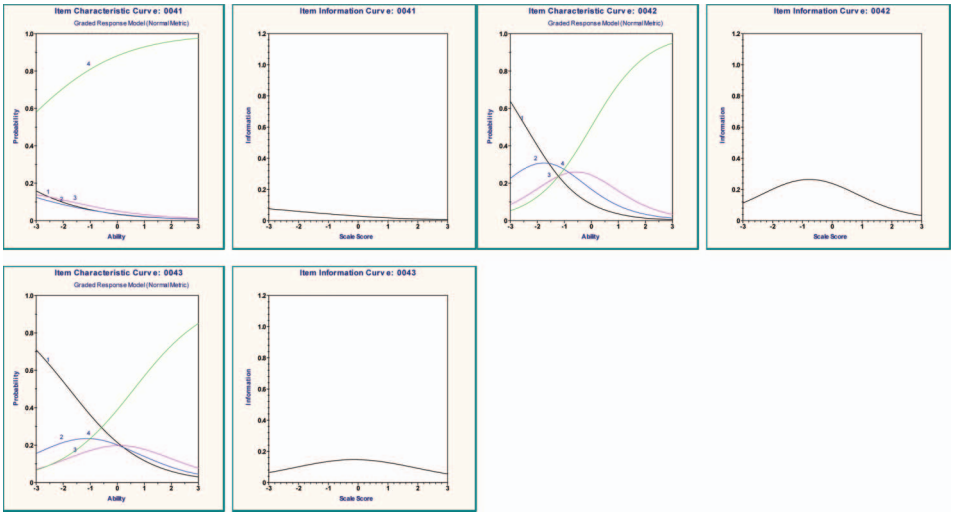

Figure 3. (*Continued*).

Items #3, #7, #12, and #40 had very low discrimination value, indicating that the performance was almost the same at the low-competency and high-competency levels. For this reason, we suggest checking these items.

The item difficulty parameter b is a point along the competency continuum where an item provides its maximum information to estimate competency. Theoretically, the value of b is defined on the scale of $(-\infty, \infty)$; however, under rare circumstances the value of difficulty is outside the range of $(-4, 4)$. Items with negative difficulties are considered easier and are more likely to be endorsed by respondents with low competency. Items with positive difficulties are considered more difficult and are more likely to be endorsed by respondents with high competency. Parameter b will contain a greater amount of information when its value is closer to θ than when it is further from θ. Although there is no rule to follow, we want items to be more useful for most students with average-level competency. We listed items whose b values were less than -4 or larger than 4. These items included #3, #5, #6, #12, #16, #17, #29, #34, #37, #39, #40, and #42. Since it is possible to obtain different estimation values of difficulty by modifying items, we suggest making changes on items whose difficulty values were close to the absolute value 4, such as #6, #16, #17, #18, #28, #34, #37, #41, and #42.

ICC and information curve

The pairs for ICC and IIC of each item were plotted in Figure 3.

By examining IIC curves, we found that Items #3, #5, #12, #17, and #40 contributed little information to the survey. Items #16, #20, #23, #25, #26, #27, #34, #37, and #41 dominated in Response category 4 and provided limited information. Items #18, #29, and #39 dominated in Response category 1 and provided limited information.

Combining all of the above information, Items #3, #7, #12, and #40 failed all selection criteria. Since they had extreme difficulty values, had low discrimination values, and contributed little information, they were removed from the scale. Items #3, #5, #12, #17, #29, #34, #39, and #40 failed two criteria. They fell out of the reasonable range of

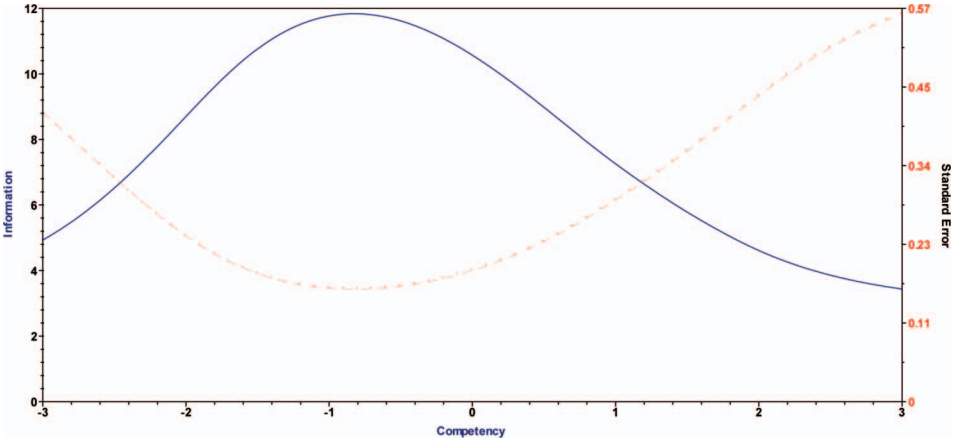

Figure 4. Test information before deleting items.

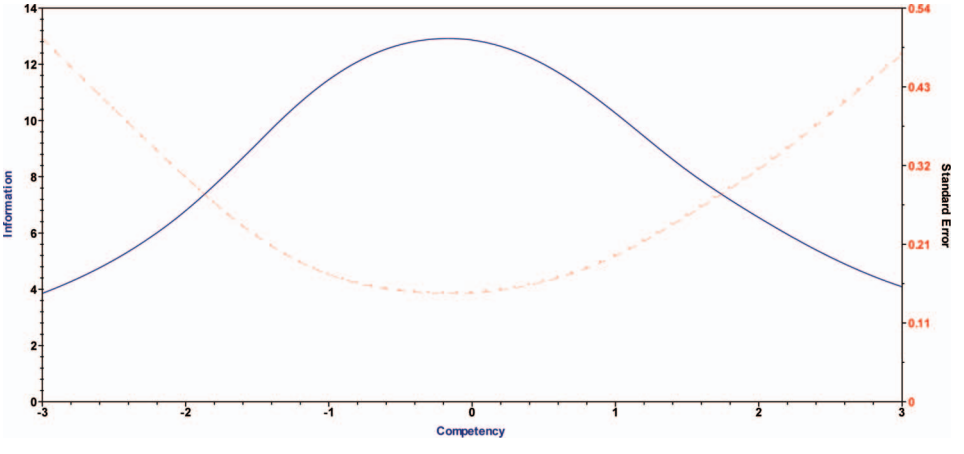

Figure 5. Test information after deleting items.

difficulty values and contributed little information. Thus, we deleted these items. Meanwhile, we suggested editing those items that either had extreme difficulty values or contributed little information, especially those items whose difficulty values were close to the cut-off point of absolute value 4, such as Items #16, #18, #34, #37, and #42.

Figures 4 and 5 present the whole scale information before and after deleting some items as described above. Before removing these items, the maximum amount of information with a competency value of -1.2 was 11.90. After removing those items, the maximum amount of information with a competency value of 0 was 13.

Discussion and conclusion

Compared to their wide application in the assessment of achievement tests, few empirical studies have been found to apply IRT in survey assessments. In the present study, we followed the principle of the maximum of item information function and the maximum of scale information function in the IRT model, used empirical data

from Memphis City Schools, and explored the possibility of improving the school climate survey as a whole.

According to the IRT model, the ICC and IIC can identify the performance of individual items on a scale. An individual item that contributes low information is usually attributed to three reasons: (1) The item is poorly worded, (2) the item contains a low discrimination value and low correlation with other items, and (3) the item is either too difficult or too easy. As a solution to this issue, survey developers may first consider rewriting the item with poor performance. If the information function of this item still cannot be improved, a simple solution would be to remove it from the scale. In the present study, we deleted some items and gave suggestions as to which items may be rewritten.

It is surprising that after deleting some items that introduced biased estimation, more information was obtained for the scale. Specifically, we found that the information curve shifted to the right, and the maximum amount of information was observed when the competency value was around the mean point. The improved competency value indicated that the modified questionnaire was appropriate for respondents with average-level competency.

A school climate survey provides an in-depth profile about the particular strengths and needs of a school. Since different school districts have different concerns, it is necessary for them to develop questionnaires that address issues according to their interests. On the other hand, it is also necessary for the schools to empirically examine the survey tools through various psychometric tools, including IRT. The examination of item and scale's reliability can decrease the impact of a biased item rather than the question itself and increase the accuracy of the survey measurement. The ultimate goal is to accurately assess students', parents', and school personnel's perception of school climate and collect the maximum information to make informed decisions on educational improvement.

IRT is a useful tool to measure how much information each item in the survey and the survey as a whole can provide. Based on the information provided by the items, survey developers can modify the items by rewriting or removing them. The modification can make the instrument more concise, more reliable, and better designed for the target participants. One limitation of the study is that only the two-parameter IRT model was adopted as an effort to improve the school climate survey. The two-parameter IRT model implemented in the present study may have different estimates from the three-parameter IRT model or nonparametric IRT model. To overcome this limitation, future studies may apply different IRT models, compare the results from different models, and determine which one is more appropriate to use in the context of questionnaire studies.

Acknowledgements

This study was sponsored in part by the grant (R305A090528) from the Institute of Education Science within the U.S. Department of Education. The first author would like to express thanks to his former colleagues: Dr. Marie Sell, Mr. Floyd Deal, Dr. Brant Riedel, Dr. Jeff Shive, Dr. John Barke, Dr. Florence Calaway and Mr. John Nickey.

Notes on contributors

Lun Mo is an Educational Cognitive Psychologist and Statistician. During the preparation of the present study, the author was employed as an Evaluator at Memphis City Schools. Dr. Mo's research interests include Educational Measurement, Cognitive Psychology, and

Educational Practice (Respect, AP and ACT). To date, he has publications on measurement of source memory, examination of AP students' academic performance, and application of cognitive learning theory in educational practices.

Fang Yang is Assistant Professor at the University of Detroit Mercy. Dr. Yang received her PhD at the University of Memphis and a Postdoctoral training at the University of Florida. Her research interests focus on the combined topics in Financial Accounting and Industrial/Organizational Psychology.

Xiangen Hu is Professor in the Department of Psychology at the University of Memphis. Dr. Hu's primary research areas include Mathematical Psychology, Research Design and Statistics, and Cognitive Psychology. More specific research interests include General Processing Tree (GPT) models, categorical data analysis, and knowledge representation, computerized tutoring, and advanced distributed learning.

References

Bock, R.D. (1972). Estimating item parameters and latent competency when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Cronbach, L.J. (1960). *Essentials of psychological testing*. New York, NY: Harper.

Freiberg, H.J. (1998). Measuring school climate: Let me count the ways. *Educational Leadership*, *56*(1), 22–26.

Hambleton, R.K. (1991). *Fundamentals of item response theory: Measurement methods for the social science*. Newbury Park, CA: Sage.

Johnson, W.L., & Johnson, A.M. (1993). Validity of the quality of school life scale: A primary and second-order factor analysis. *Educational and Psychological Measurement*, *53*, 145–153.

Johnson, W.L., Johnson, A.M., Gott, R., & Zimmerman, K. (1997). Assessing the validity of scores on the Charles F. Kettering Scale for the junior high school. *Educational and Psychological Measurement*, *57*, 858–869.

Kuperminc, G.P., Leadbeater, B.J., & Blatt, S.J. (2001). School social climate and individual differences in vulnerability to psychopathology among middle school students. *Journal of School Psychology*, *39*, 141–159.

Kuperminc, G.P., Leadbeater, B.J., Emmons, C., & Blatt, S.J. (1997). Perceived school climate and difficulties in the social adjustment of middle school students. *Applied Developmental Science*, *1*, 76–88.

Manning, M.L., & Saddlemire, R. (1996). Developing a sense of community in secondary schools. *National Association of Secondary School Principals (NASSP) Bulletin*, *80*(584), 41–48.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No.17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf