Paper 957-2017

Does factor indeterminacy matter in multi-dimensional item response theory?

Chong Ho Yu, Ph.D., Azusa Pacific University

ABSTRACT

This paper aims to illustrate proper applications of multi-dimensional item response theory (MIRT), which is available in SAS's PROC IRT. MIRT combines item response theory (IRT) modeling and factor analysis when the instrument carries two or more latent traits. While it may seem convenient to accomplish two tasks simultaneously by employing one procedure, users should be cautious of mis-interpretations. This illustration utilizes the 2012 Programme for International Student Assessment (PISA) data set collected by Organization for Economic Cooperation and Development. Because there are two known sub-domains in the PISA test (reading and math), PROC IRT was programmed to adopt a two-factor solution. Additionally, the loading plot, dual plot, item difficulty/discrimination plot, and test information function plot in JMP were utilized to examine the psychometric properties of the PISA test. When reading and math items were analyzed in SAS's MIRT, 7-13 latent factors are suggested. At first glance these results are puzzling because ideally all items should be loaded into two factors. However, when the psychometric attributes yielded from a 2–parameter IRT analysis are examined, it is evident that both the reading and math test items are well-written. It is concluded that even if factor indeterminacy is present, it is advisable to evaluate its psychometric soundness based on IRT because content validity can supersede construct validity..

INTRODUCTION

Item response theory (IRT) is a psychometric tool that can amend shortcomings of the classical test theory (CTT). Specifically, in CTT, estimations of item difficulty and person trait are sample-dependent; in IRT, however, estimations of item parameters and person theta are more precise (Embretson & Reise, 2000). IRT and its close cousin, Rasch modeling, assume uni-dimensionality (Yu, 2013). In other words, a test or a survey used with these approaches should examine only a single latent trait of the participants. In reality, many tests or surveys are multi-dimensional; to address this issue multi-dimensional IRT (MIRT) was introduced (Hartig & Hoher, 2009).

To a certain extent, MIRT is a fusion of factor analysis and IRT. Although factor analysis and IRT share common ground in that some parameterizations of model parameters in factor analysis can be transformed into parameters in item response theory (Kamata & Bauer, 2008), the underlying philosophy of IRT is vastly different from that of factor analysis, which belongs to the realm of CTT. For example, the psychometric attributes of an instrument yielded from factor analysis attach to the entire scale. If one selects items from a validated scale, then the original psychometric properties will be damaged. In contrast, each item developed by IRT has its own characteristic (item difficulty parameter, discrimination parameter, guessing parameter, item information function, etc.). Hence, it is legitimate to generate an adaptive test by selecting items from an item bank. In addition, when responses are dichotomous (e.g. “right” or “wrong”) instead of ordinal (e.g. Likert scale ratings), conventional factor analysis utilizing Pearson’s correlation matrix is invalid. Nevertheless, throughout the last decade, numerous algorithms have been developed to make this fusion successful (Han & Paek, 2014). For example, in MIRT software packages the tetrachoric correlation matrix has replaced Pearson’s correlation matrix. At first glance it is efficient to accomplish two tasks concurrently (identify the factor structure and the item characteristics). However, it is important to point out that sometimes IRT and factor analysis results might not concur with each other. Specifically, it is common that while IRT yields excellent psychometric properties of the test items, factor analysis failed to yield a sound solution.

METHODOLOGY

This illustration utilizes the 2012 Programme for International Student Assessment (PISA) data set collected by Organization for Economic Cooperation and Development (OECD, 2013). Every four years OECD delivers assessments in three subject matters (reading, math, and science) to its members and

partner countries. In 2012, 51,000 students from 65 countries participated in PISA. In order to rule out cultural difference as an extraneous variable, in this analysis only USA students are selected (*n* = 4978). However, not all students answered all test items; rather, different booklets were assigned to different students and the missing values were imputed to obtain the plausible value for each student. To simplify the illustration, a subset of US students and a subset of items are selected so that imputation is not necessary. As a result, only 411 students are retained.

The original PISA exam contained 13 booklets and 206 items across the three domains. One may argue that both math and science require similar reasoning approaches and thus the two factors are indistinguishable in the data is expected. To illustrate factor indeterminacy in this large-scale assessment, the authors retain reading and math items only. According to Gardner (2006), linguistic skill and math skill belong to different dimensions of intelligence.

Because there are two known sub-domains in the test (reading and math), PROC IRT in SAS was forced to adopt a two-factor solution. The loading plot, dual plot, item difficulty/discrimination plot, and test information function plot in JMP were utilized to examine the psychometric properties of the PISA test. The dual plot is a graphic display that places item parameters and student ability (theta) on the same scale, whereas the Test Information Function (TIF) is the sum of all item information functions in the test.

RESULT

Figure 1 depicts the scree plot and variance explained. Based on this information, as well as on the eigenvalues of the polychoric matrix of math and reading items, it seems that there are 13 underlying factors in the PISA test. Figure 2 displays the loading plot of all item vectors in a 2-factor solution. Obviously, all vectors are jammed together and no clustering pattern can be detected.

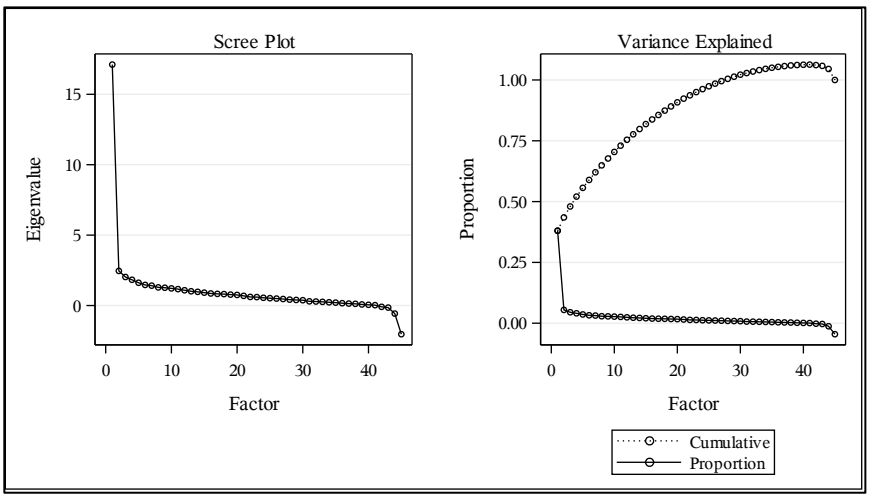

Figure 1. Scree plot from PROC IRT for PISA math and reading items.

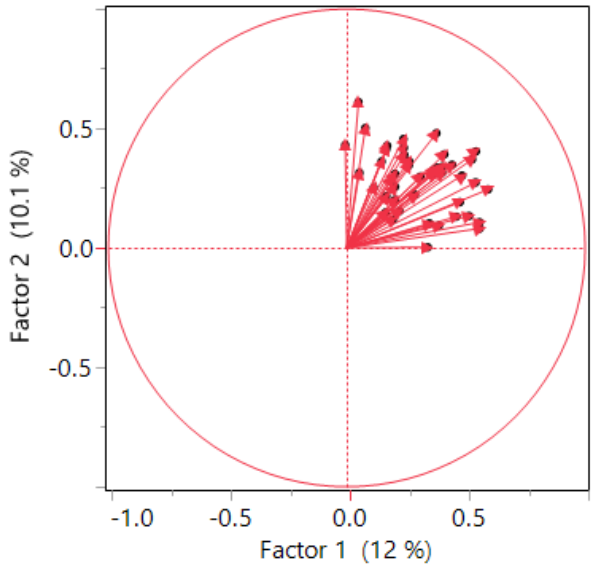

Figure 2. Loading plot of PISA items.

As mentioned before, one may argue that there might be two latent traits (reading and math skills) that contribute to the test performance of a math test. When math items are analyzed in SAS's MIRT, seven latent factors are suggested (see Figure 3).

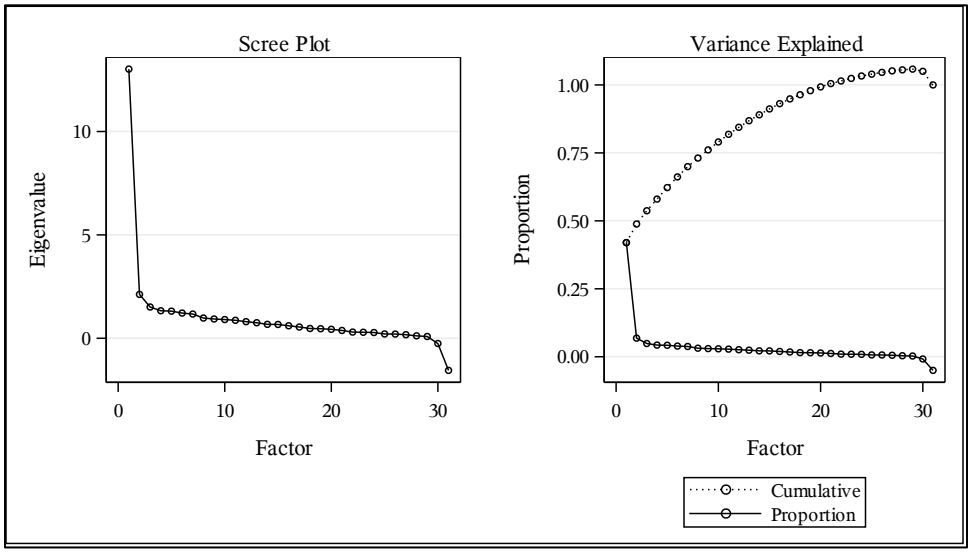

Figure 3. Scree plot from PROC IRT for PISA math and reading items.

At first glance these results are puzzling or even discouraging. However, when the dual plot (Figure 4) yielded from a 2–parameter IRT analysis is examined, it is evident that the math test items are well-written. First, the item difficulty level is evenly distributed; there is no extremely difficult or super-easy item. The student ability distribution also forms a fairly normal curve. More importantly, when the item and student attributes compared, it is obvious that the test difficulty matches the student ability. If the best students can outsmart the test then there will be no corresponding items on the left side of the dual plot. Conversely, if the test is too challenging there will be no corresponding student on the right side of the plot. But this dual plot shows that every item has a matching student and every student has a matching item.

The item difficulty/discrimination plot (Figure 5) provides additional support for the psychometric soundness of the test by showing that all items have positive discrimination parameters. Further, the test

information function plot (Figure 6) shows that much information about users whose ability estimates concentrate around the center (zero) can be learned from the exam. The peak of the TIF is above zero, indicating that more information about the students with slightly above-average ability estimates can be obtained. Math is considered a challenging subject and this psychometric information presents an important contribution to the field of mathematics education.

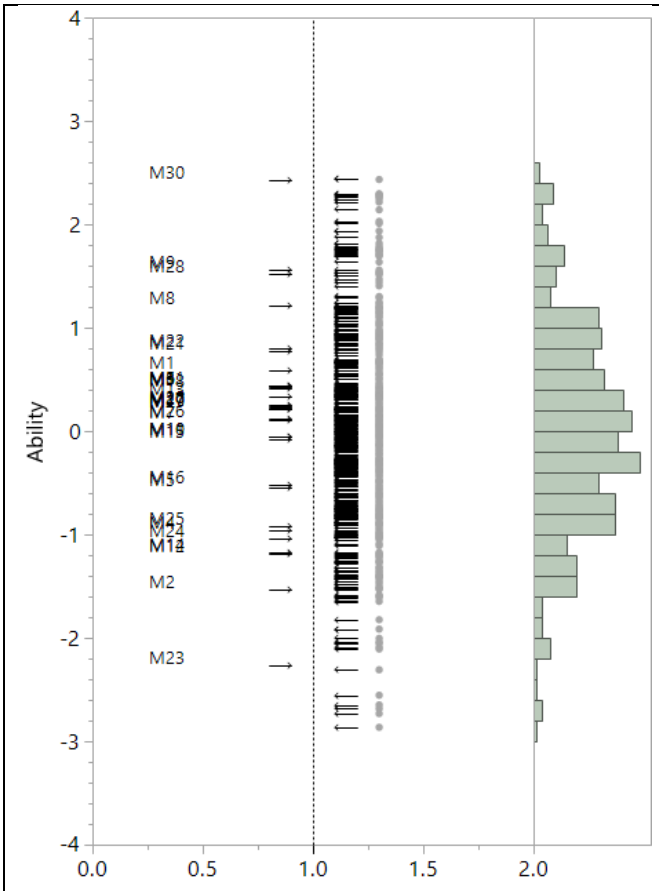

Figure 4. Dual plot of PISA math items.

Item	Difficulty	Discrimination
M1	0.5907	1.4874
M2	-1.5340	0.7029
M3	0.4321	1.5175
M4	-0.9551	1.5605
M5	-0.5497	0.8897
M6	0.4432	1.6813
M7	0.1058	0.6833
M8	1.2082	1.7824
M9	1.5620	2.6163
M10	-0.0576	1.9344
M11	-0.0542	1.2755
M12	-1.1925	1.0609
M13	0.3326	2.4837
M14	-1.1764	0.7011
M15	-0.0782	1.4960
M16	-0.5148	1.2578
M17	0.2201	0.9056
M18	0.4081	1.1894
M19	0.2093	1.5697
M20	0.2474	0.8379
M21	0.7750	2.3186
M22	0.7953	2.1087
M23	-2.2685	1.3003
M24	-1.0449	1.2897
M25	-0.9222	1.4980
M26	0.1243	1.3750
M27	0.2359	0.8302
M28	1.5180	2.2114
M29	0.2076	2.6948
M30	2.4211	2.7404
M31	0.4420	1.6899

Figure 5. Item difficulty/discrimination plot of PISA math items and USA student ability estimates.

Similar findings are observed in the IRT modeling of PISA reading items. Figures 7 to 9 show that the reading items also have fine psychometric properties. Specifically, both the item difficulty level and the student ability are normally distributed. The items are well balanced in that students cannot outsmart the test, but the test is also not overly challenging. Further, all discrimination parameters are positive and the test information function concentrates around zero. The peak of the TIF is slightly below zero, meaning that more information about the students whose reading skill is mildly below average can be obtained. Given all these IRT psychometric attributes, it is absurd to deny the validity of the PISA exam just because there is no clear factor solution. Further, when math and reading tests are separately evaluated in two IRT models it returns a single ability estimate for each student. But if MIRT model is used, then it will yield individual ability profiles as test results rather than single scores (Hartig & Hoher, 2009). The question is: could we obtain more useful information by adding this extra layer of complexity?

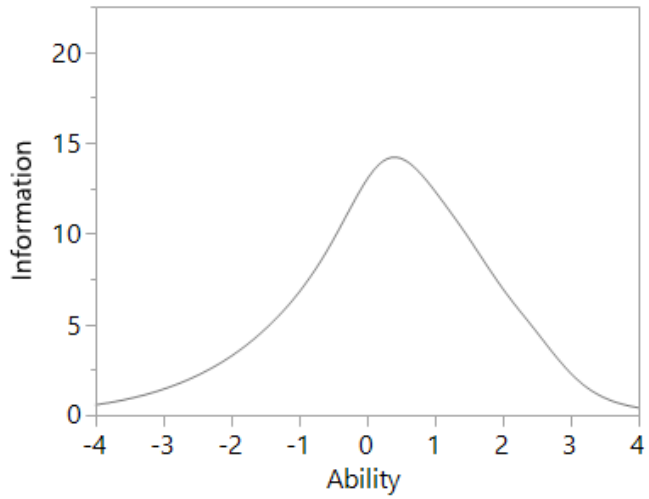

Figure 6. Test Information Function Plot of PISA math items.

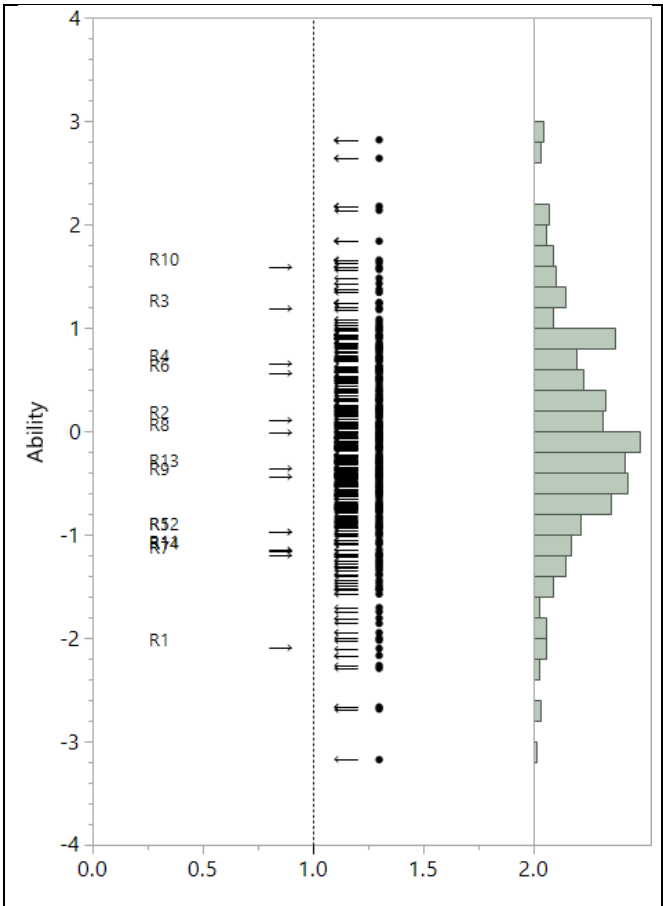

Figure 7. Dual plot of PISA reading items.

Item	Difficulty	Discrimination
R1	-2.0926	1.2044
R2	0.1043	1.7469
R3	1.1894	0.6491
R4	0.6547	1.5222
R5	-0.9772	1.5257
R6	0.5538	0.7577
R7	-1.1935	1.6693
R8	-0.0088	1.4276
R9	-0.4370	1.1741
R10	1.5862	1.1349
R11	-1.1526	2.2098
R12	-0.9674	1.4423
R13	-0.3624	1.8990
R14	-1.1660	1.9855

Figure 8. Item difficulty/discrimination plot of PISA reading items and USA student ability estimates.

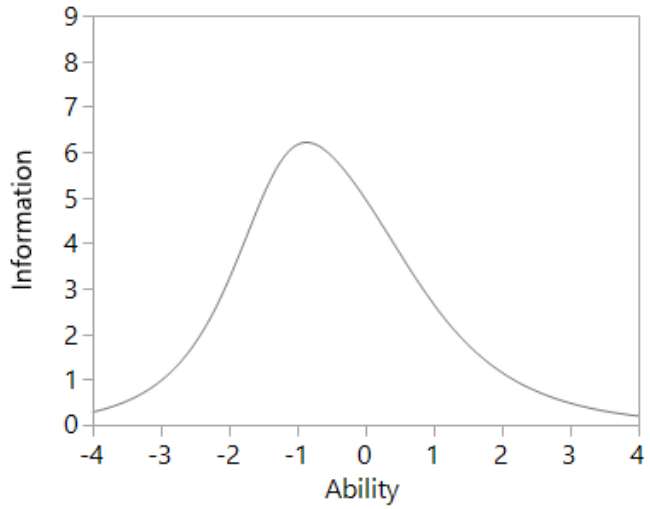

Figure 9. Test Information Function Plot of PISA reading items.

CONCLUSION

In the past, CTT and IRT were believed to be incompatible. As such, merging these methodologies seemed to be a remote dream. Nevertheless, with the advance of MIRT the dream became a reality. Indeed, combining these methodologies can remediate limitations in both camps. While traditional IRT allows for measurement of a single construct, infusing factor modeling enhances IRT with multidimensionality. In CTT the psychometric property is tied to the entire instrument; consequently, when a researcher wants to customize a scale to a specific population, he or she needs to repeat the tedious process of EFA and CFA. IRT provides users with the flexibility of generating an ad hoc, on-the-fly test. Despite the versatility of MIRT, it is not uncommon that PROC IRT could not yield a sound IRT model and a factor model at the same time. When factor indeterminacy is present, it is advisable to evaluate its psychometric soundness based on IRT alone because content validity can supersede construct validity.

REFERENCES

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. New York, NY: Psychology Press.

Gardner, H. (2006). *Multiple intelligences: New horizons in theory and practice*. New York, NY: Basic Book.

Han, & Paek, I. (2014). A review of commercial software packages for multidimensional IRT modeling. *Applied Psychological Measurement, 38*, 486-498.

Hartig, J., & Hohler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation, 35*, 57-63.

Kamata, A. & Bauer, D. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15*, 136–153.

Organization for Economic Co-operation and Development [OECD] (2013). *The PISA international database*. Retrieved from http://pisa2012.acer.edu.au/

Yu, C. H. (2013). *A simple guide to the item response theory (IRT) and Rasch modeling*. Retrieved from http://www.creative-wisdom.com/computer/sas/IRT.pdf

ACKNOWLEDGMENTS

Special thanks to Ms. Anna Lee and Ms. Samantha Douglas for their valuable input to this article.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Chong Ho Yu, Ph.D., D. Phil.
Azusa Pacific University
cyu@apu.edu
http://www.creative-wisdom.com/pub/pub.html