Young Lives

An International Study of Childhood Poverty

YOUNG LIVES SCHOOL SURVEY

ITEM RESPONSE SCORING IN THE VIETNAM SCHOOL SURVEY ROUND 1

ABHIJEET SINGH

July 2013

Overview

This note summarizes the process of generation of comparable test scores for Math and Vietnamese across the test and retest components of the Vietnam school survey (2012). The two rounds of testing were carried out at the beginning and the end of the academic year and were designed specifically to allow for the analysis of the learning increments ('value-added') over the course of the academic year. For this purpose, it is very useful to generate scores on the same metric using anchor items in Item Response Theory (IRT).

In this note, I detail the specific methodology of generation of the IRT scores in the Vietnam data and provide diagnostic graphs illustrating item fit and the distribution of final test scores.

Methodology

All questions in the Vietnam school survey tests, both for math and reading comprehension, were multiple choice questions; accordingly a three-parameter (3PL) Item Response Theory Model was thought to be most appropriate. This is the approach also taken by international testing programs such as TIMSS and PISA[1].

IRT models posit a mathematical relationship between the latent ability of an individual and the probability that the individual will correctly answer a given item; this relationship differs from item-to-item but is assumed constant across individuals. In the 3PL model, this relationship is given by:

$$P_g(\theta_i) = c_g + \frac{1 - c_g}{1 + \exp[-1.7 \cdot a_g \cdot (\theta_i - b_g)]}$$

where θ_i is the individual's ability, c_g is the "pseudo-guessing parameter", b_g is the item difficulty parameter and a_g is the item discrimination parameter. Under three core assumptions of IRT – unidimensionality of the trait being measured, local independence of item responses conditional on ability and no differential item functioning, standard maximum likelihood techniques can be used for estimating item specific parameters (a ,b, c) and individual ability parameters (θ) which are referred to here as the test scores.

Estimation was carried out in Stata using the OpenIRT suite of commands written by Tristan Zajonc. Missing responses to particular questions on a given test booklet were marked as incorrect for the purpose of generating the test scores. Estimated scores were standardized to have a mean of 500 and a variance of 100 in the base period, allowing these to shift in the second round. Replicated items across the two rounds were used to anchor the ability distributions across the two test rounds.

Results

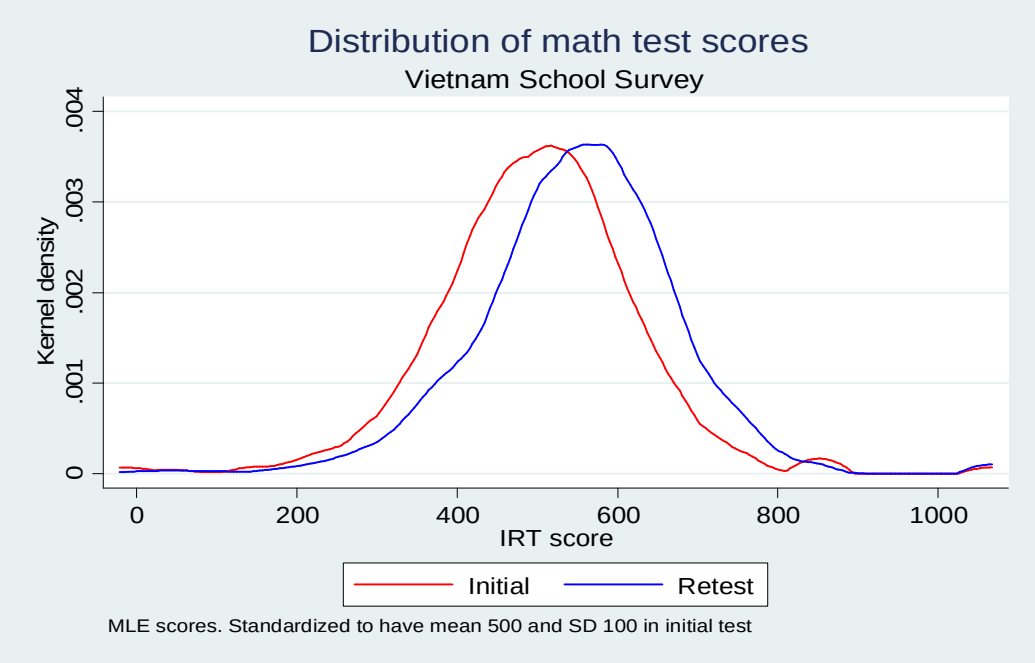
The OpenIRT software generates three sets of estimates of ability – maximum likelihood estimates, Bayesian expected posterior estimates and plausible values or multiple imputations estimates. Maximum likelihood estimates are unbiased estimates for an individual's ability (and for the sample mean) although it may be subject to significant noise especially at the ends of the distribution. Plausible values estimates draw plausible values from an individual's posterior distribution and then use these draws to estimate the true ability distribution (Das and Zajonc, 2010); these are better estimates of the higher moments (variance, skewness etc.) of the ability distribution[2].

The distribution of the math test score is given in Figure 1. As can be seen, the distribution has shifted rightwards i.e. indicating that additional skills have been acquired over the school year.

[1] See Van der Linden and Hambleton (1997) for an overview of IRT models.

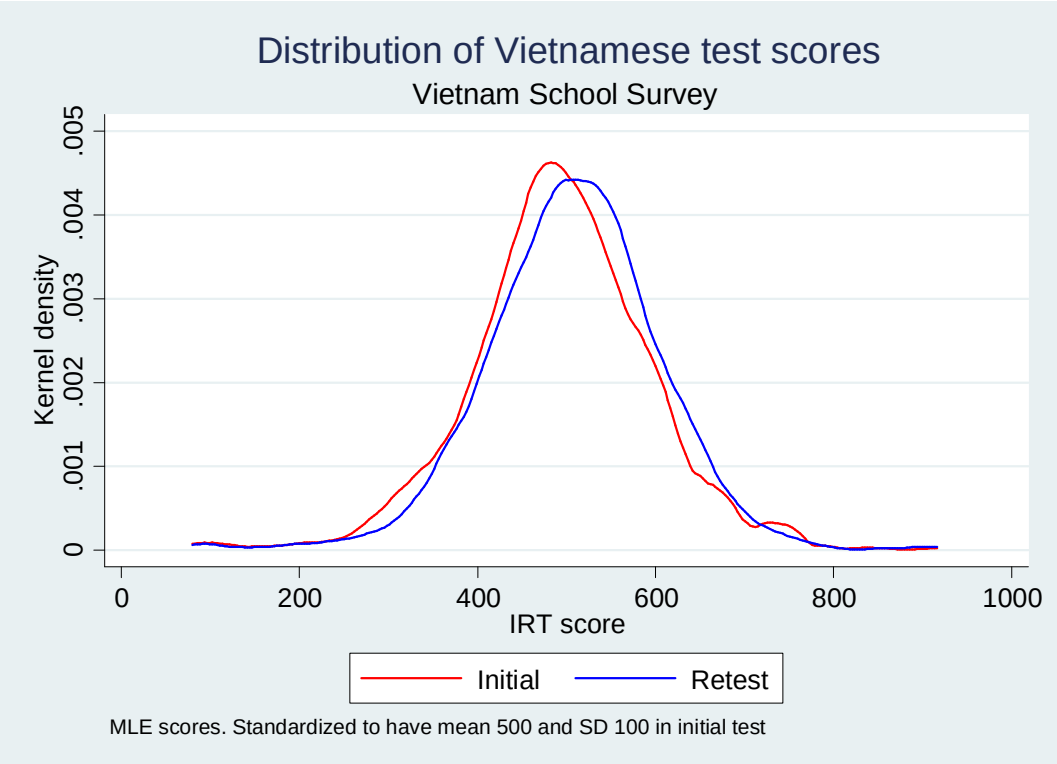
[2] See Das and Zajonc (2010) for a detailed discussion of the different estimates and their interpretation. EAP scores are generated without the inclusion of any manifest predictors.

Figure 1. Distribution of test scores in Mathematics



The distribution of the Vietnamese test scores is given in Figure 2. The near overlap between the two distributions indicates less evidence of increments in learning, although there are some signs that the distribution might have shifted to the right marginally and there is a statistically significant change in the mean score.

Figure 2. Distribution of test scores in Vietnamese



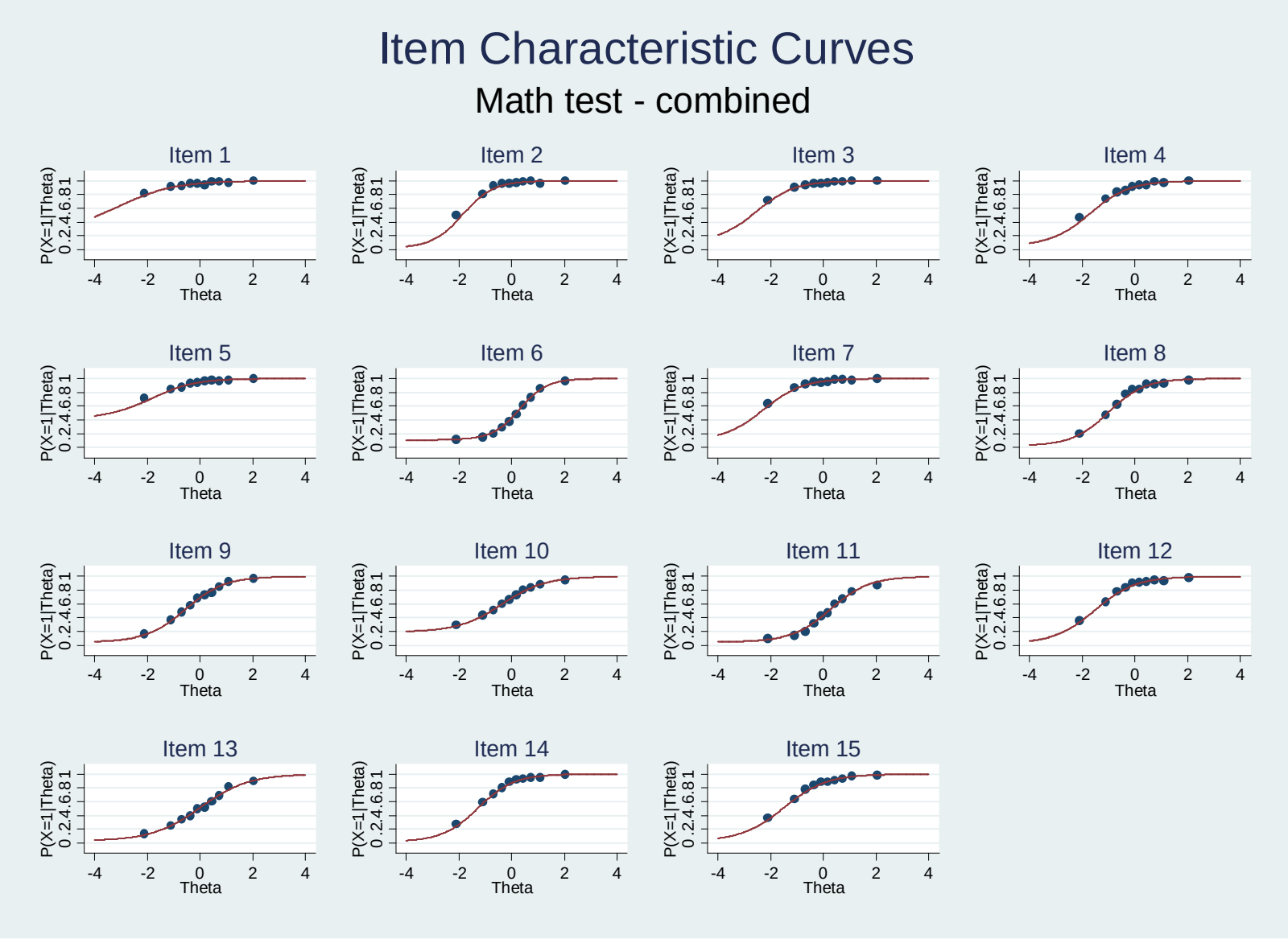
We also investigated the item fit of each of the test items by visually inspecting the Item Characteristic Curves (ICC) for each item. ICCs plot the predicted probability that an individual at a given level of ability answers the question correctly with the observed proportion correct in the data. Appendix 1 provides the ICCs for mathematics and Appendix 2 provides ICCs for Vietnamese.

References

Das, J. & Zajonc, T. (2010). India shining and Bharat drowning: Comparing two Indian states to the worldwide distribution in mathematics achievement. *Journal of Development Economics*, *92*(2), 175-187.

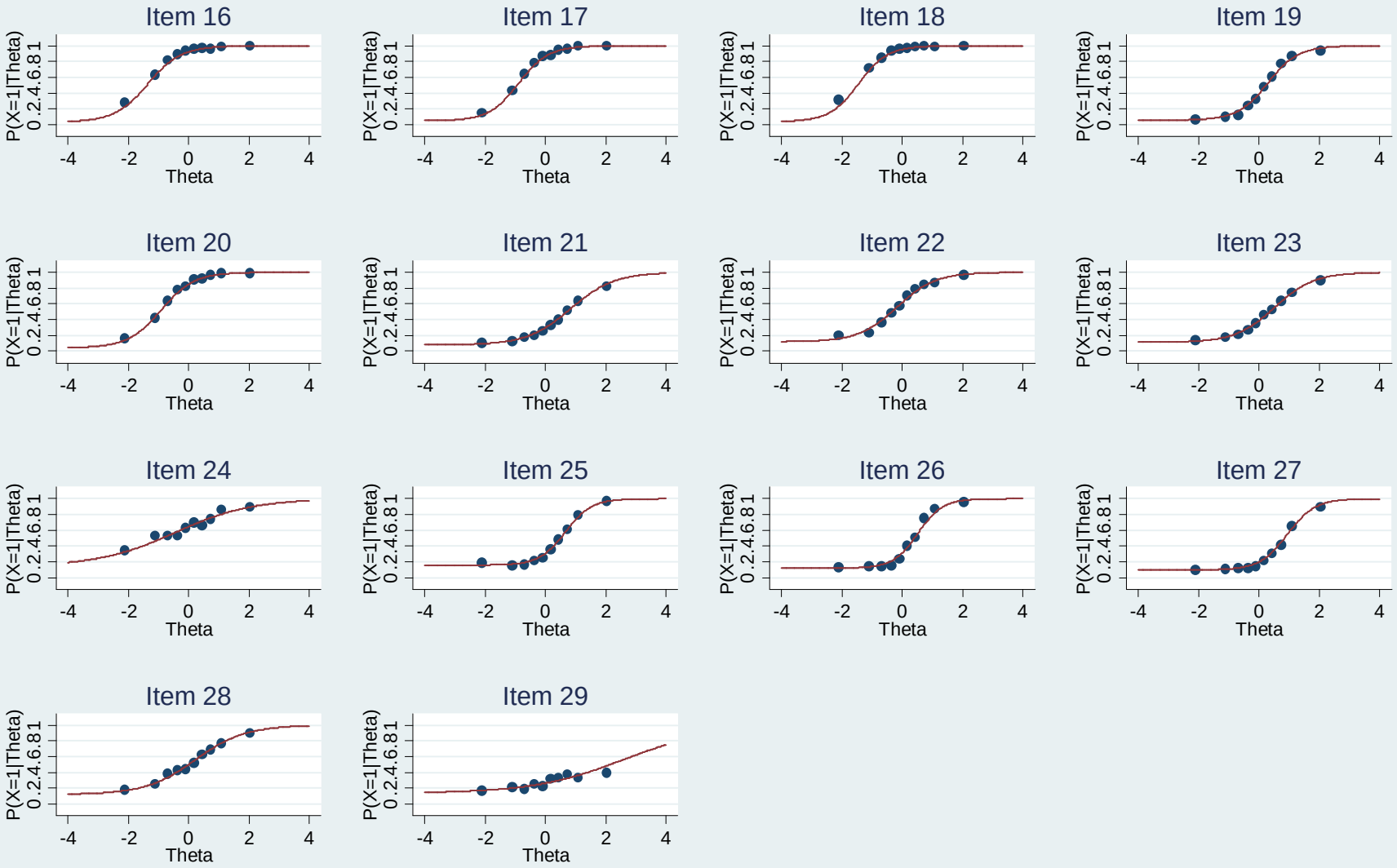
Van Der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. *Handbook of modern item response theory*, 1-28.

Appendix 1. Item Characteristic Curves for Mathematics



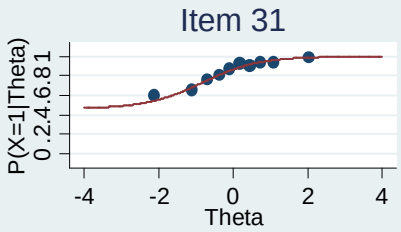
Item Characteristic Curves

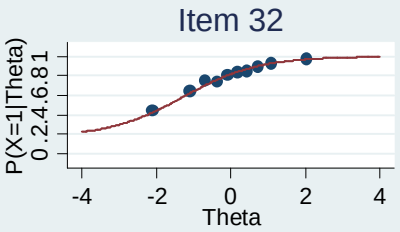
Math test - combined

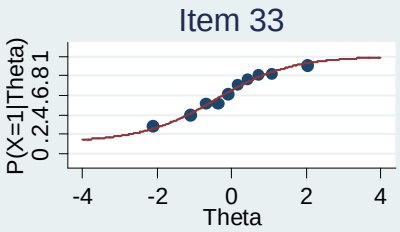


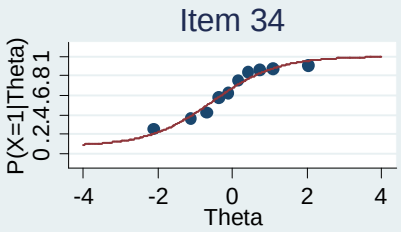
Item Characteristic Curves

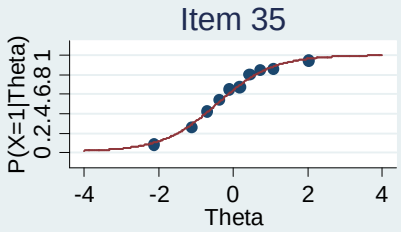
Math test - combined

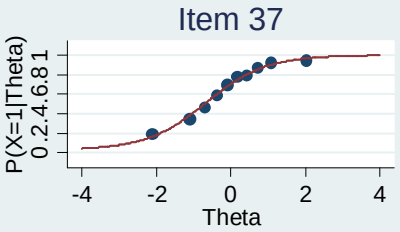


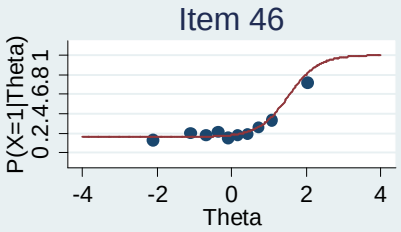


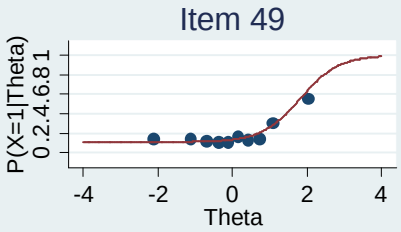


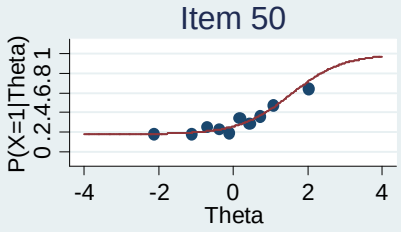


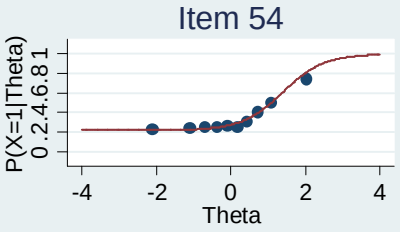


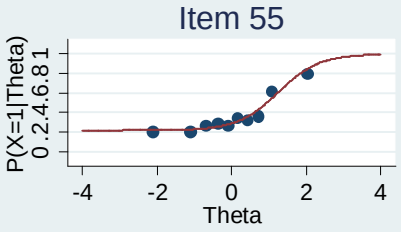


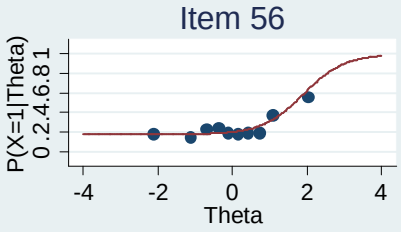


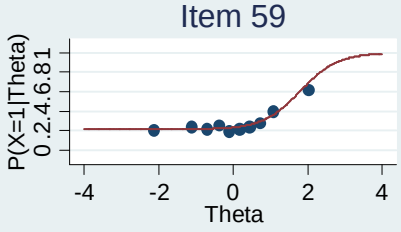




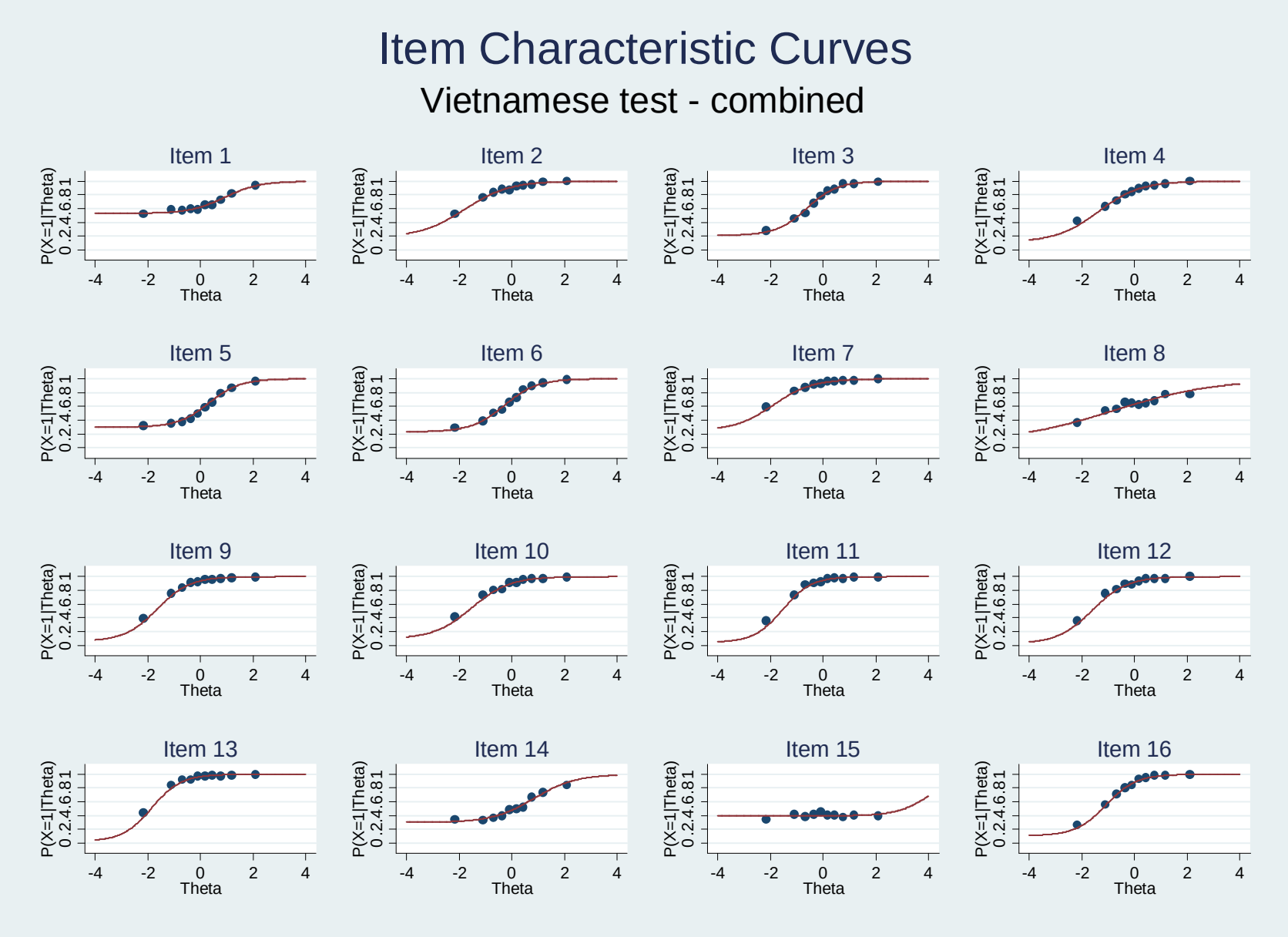






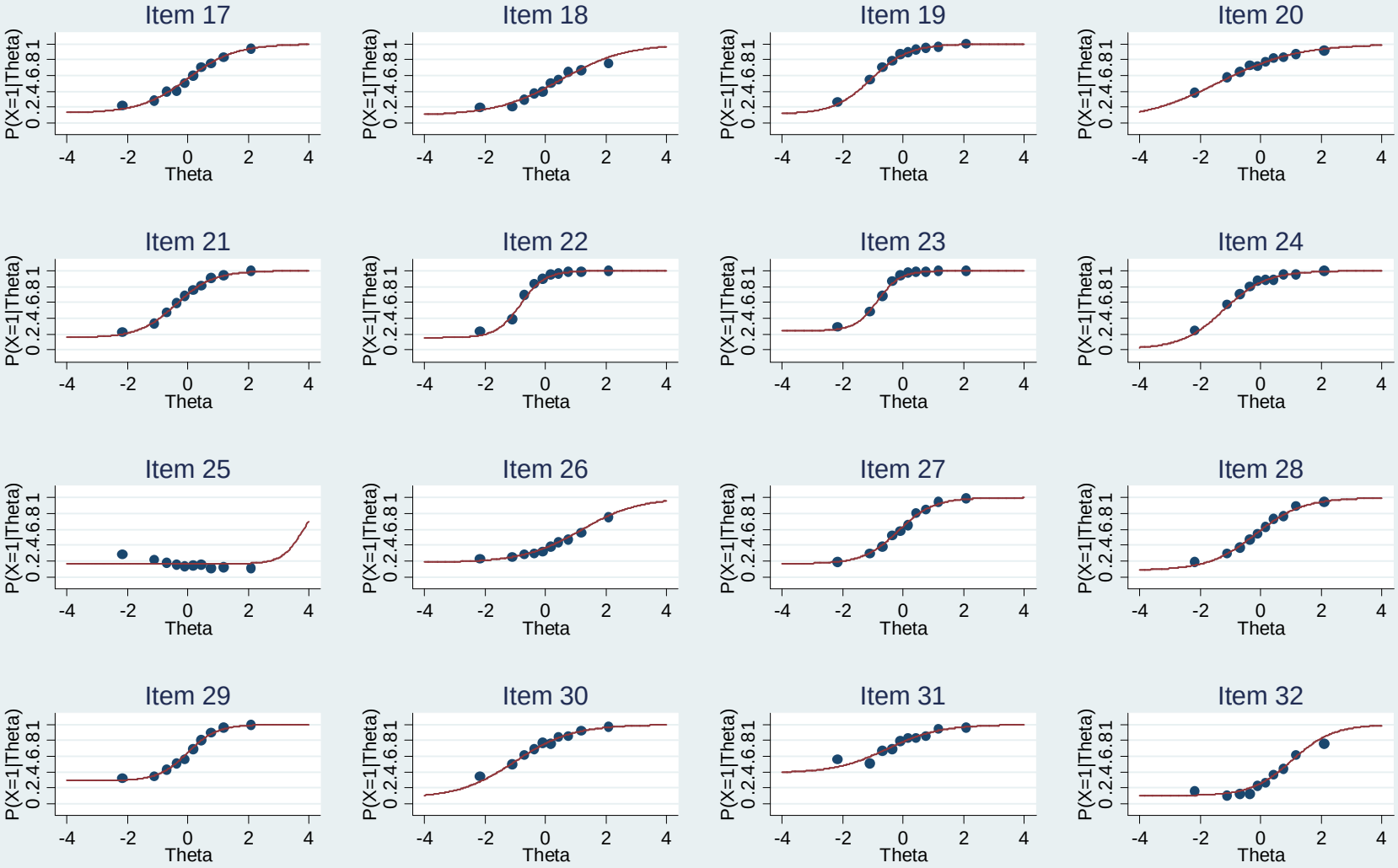


Appendix 2. Item Characteristic Curves for Vietnamese



Item Characteristic Curves

Vietnamese test - combined



Item Characteristic Curves

Vietnamese test - combined