International Education Studies; Vol. 9, No. 1; 2016
ISSN 1913-9020 E-ISSN 1913-9039
Published by Canadian Center of Science and Education

Developing a Numerical Ability Test for Students of Education in Jordan: An Application of Item Response Theory

Eman Rasmi Abed[1], Mohammad Mustafa Al-Absi[1] & Yousef Abdelqader Abu shindi[1]

[1] Faculty of Educational Sciences and Arts, UNRWA, Amman, Jordan

Correspondence: Eman Rasmi Abed, Faculty of Educational Sciences and Arts, UNRWA, Amman, Jordan. Tel: 9-626-534-8486, Mobile: 9-627-779-39277. E-mail: eabed67@yahoo.com

Received: June 18, 2015 Accepted: August 12, 2015 Online Published: December 29, 2015

doi:10.5539/ies.v9n1p161 URL: http://dx.doi.org/10.5539/ies.v9n1p161

Abstract

The purpose of the present study is developing a test to measure the numerical ability for students of education. The sample of the study consisted of (504) students from 8 universities in Jordan. The final draft of the test contains 45 items distributed among 5 dimensions.

The results revealed that acceptable psychometric properties of the test; items parameters (difficulty, discrimination) were estimated by item response theory IRT, the reliability of the test was assessed by: Cronbach's Alpha, average of inter-item correlation, and test information function (IRT), and the validity of the test was assessed by: arbitrator's views, factor analysis, RMSR, and Tanaka Index.

The numerical ability test can be used to measure the strength and weaknesses in numerical ability for educational faculty students, and the test can be used to classify students on levels of numerical ability.

Keywords: education, item response theory, numerical ability

1. Introduction

Assessment is one of the set of standards for mathematics curriculum that was suggested by the National Council of Teachers of Mathematics (NCTM, 1989). It is the process of determining how much of the objectives planned by curriculum were achieved, what is the level reached by the student, to check how the outcomes of learning and experience are gained. It is stated in the document of (NCTM, 1995) as the process of collecting evidence about the student's knowledge and ability to use mathematical knowledge and trends towards mathematics, and draw the provisions of this evidence for a variety of purposes.

So within a comprehensive system of assessment, students learn to assess their own progress and to set goals (Smith, 2013, p. 35). Schuwirth (2010) stated that "*Teachers are encouraged to make use of the results of assessment of learning to benefit the learners by reviewing their performance in the assessment activities with them and working out a plan for further improvement*" (p. 171).

Numerical ability and skills is one of the skills that focused on NCTM standards, which included some of the concepts and skills such as: Numerical sense and counting systems, the concepts of numbers operations, numbers and numerical relationships, the theory of numbers, calculation and estimation (NCTM, 1989). This standard in the field of the number and operations aims to enable students to understand the numbers and methods of representation and the relationships between numbers and numerical systems, understand the meanings of operations and how they relate to each other, and to ease of calculation and the work of reasonable estimates (NCTM, 2000).

In the numerical ability and skills test we can use a speed test, which may be consisted of basic arithmetic, such as: addition, subtraction multiplication and division, number sequences and simple mathematics, such as: percentages, powers, and fractions (Psychometric Success, 2013), obviously you will not be allowed to use a calculator.

Tests play an important role in the development of the learning process, and when the test were prepared and developed in a standardized method it will provide a quantitative data on the measured features in a high degree of validity and reliability. This type of tests is called a standardized test.

Popham (2005) the former president of the American Educational Research Association, defined the

standardized test as "any test that's administered, scored, and interpreted in a standard, predetermined manner". The tests often have multiple-choice questions that can be quickly graded by automated test scoring machines (Mitchell, 2006). Some tests also incorporate open-ended questions that require human grading, which is more expensive, though computer software is being developed to grade written work also (Jaschik, 2011).

Standardized testing in the United States of America ramped up in 2002 due to the adoption of the No Child Left behind Act (Evans, 2013). The act aimed to hold all public schools to a high standard of education, measured by their students' scores in statewide standardized tests. This type of assessment is important for students, teachers and schools; since the low scores can prevent a student from progressing to the next grade level or lead to teacher firings and school closures (Morin, 2011), while high scores ensure continued federal and local funding and are used to reward teachers and administrators with bonus payments.

Standardized tests have four basic functions in educational systems (Betz, Eickhoff, & Sullivan, 2013; Costas, 2014):

1) Selection: if you have many applicants with limited opportunities you can easily choose the best by depending on the test results.

2) Classification: the test results can be a good norm or standard for rating and classifying students to levels or disciplines.

3) Assessment: it is well known that students' assessment is an important feature for the development and improvement.

4) Diagnosis: the test results are important for the diagnosis of learning difficulties, by revealing weaknesses to find solutions. There are many commercial tests, but the factors that lead clinicians to select particular tests to use in clinical practice were unknown.

5) Re-application: standardized tests may be useful to other schools seeking to develop their own test.

The standardized tests have many technical characteristics, such as:

1) Validity: the valid test is the test that measures what it was placed accurately. 'Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations (AERA, APA, & NCME, 1999, p. 9).

2) Reliability: the reliable test is a test which gives almost the same results if it is applied to the same people in different time intervals. Reliability refers to the consistency of an instrument when the testing procedure is repeated on a population of individuals or groups (Cueto, Leon, Guerrero, & Munoz, 2009).

3) Objectivity: the objective test is a test that gives the same results no matter what assessors; it is not affected by values, or by persons (Higginson, 1992).

1.1 Literature Review

Galli, Chiesi, and Primi (2008) conducted a study to develop a scale to measure the math ability that psychology students need to enroll introductory statistics courses inside their degree program. The Rasch model was applied to construct the instrument. The principal component analysis (PCA) of the residual showed a one-dimensional construct; the fit statistics revealed a good fit of each item to the model. The item difficulty measures were examined and the area of ability accurately assessed by the items was identified. The validity of the scale was assessed: the measures obtained by the scale correlated with attitude toward statistics and statistics anxiety (concurrent validity), and a relationship with statistics achievement was found (predictive validity).

Adedoyin and Mokobi (2013) conducted a study aims at providing the psychometric analysis of 2010 Botswana mathematics Junior Certificate paper 1 in determining the quality of the Junior Certificate mathematics multiple choice examination test items. The mathematics paper 1 consisted of forty (40) multiple choice test items which was constructed using the three year Junior Certificate mathematics curriculum. The population for the study was all the 36,940 students who sat for the Junior Certificate mathematics examination in 2010, out of which a sample of 10,000 was selected randomly by the use of SPSS computer software. The students' responses were analyzed using IRT (3PL) model to examine the psychometric parameter estimates of the forty test items which were: item difficulty, item discrimination, and the guessing value. The item characteristics curves were also generated for each test item that fitted the IRT (3PL) model. Twenty three (23) items fitted the 3PLM out of the forty (40) items, and were used in examining the psychometric qualities of the Junior Certificate mathematics test paper 1.The findings from this study indicated that out of the twenty three (23) items that fitted the IRT model, twelve (12) items were classified as poor test items, ten (10) items were classified as fairly good test

items which could be revised or improved and one (1) item was considered to be good test item. It was therefore recommended that examination bodies should consider improving the quality of their test items by conducting IRT psychometric analysis for validation purposes.

Abed, Al-shayeb, and Abu Zeineh (2015) developed A Mathematical Thinking test for High Basic Stage Students in Jordan. They collected data from (1147) male and female 8th, 9th, and 10th during one scholastic year. Item analysis revealed an acceptable discrimination and difficulty indices. The final version of the test includes 27 items. The factor analysis identified three major factors: logical reasoning, induction and generalization, and model and symbolic. Acceptable values of internal consistency found. Results revealed significant differences on each sub-test due to the academic level variable, and no differences due to the student gender.

Jamhawy (2000) compared between classical test theory CTT and item response theory IRT in developing mathematics ability test for 1061 Jordanian students in grades 9. The final version of the test includes 39 items. Results showed that item statistics and examinees ability of the two procedures were comparable. 33 items were selected by CTT procedures, 20 items by rasch model, 35 by two-parameter model, and 38 by three- parameter model. And also results revealed that no statistical differences between thee reliability coefficient drives from CTT, two- parameter model, but there were between CTT and rasch model, and between CTT and three-parameter model. Also, results revealed that the two-parameter model was the most comparable model with CTT.

Galli, Chiesi, and Primi (2010) assessed mathematics competences in statistic course. The item response theory IRT was applied to construct a scale to measure the mathematical ability deemed necessary for psychology students to successfully complete introductory statistics courses. Factor analysis used to explore the test dimensionality. The second part of the research intended to establish whether the data fit the model chosen (RSL model). The dimensionality analysis indicates that the PMP refers a unidimensional construct. The item difficult measures showed that items contents are easy for the students, and the discriminative measures showed that items can discriminate students with different mathematical ability levels. The analysis of relationship between mathematical ability and achievement attests the Predictive validity of PMP.

In conclusion, having reviewed the above studies, we find that few researches related to construct general mathematics ability tests (Galli, Chiesi, & Primi, 2008; Adedoyin & Mokobi, 2013; Jamhawy, 2000; Galli, Chiesi, & Primi, 2010). Most of researches using Item response theory IRT in calibrating tests.

Reviewing the antecedent researches and studies revealed the importance of mathematical ability, and the importance of developing tests of mathematical ability.

This study used IRT to calibrate numerical ability test for students of educational sciences in Jordan.

1.2 Research Problem

Mathematics is an important and necessary science for every individual in any society and society itself, because there is a need to mathematical knowledge to solve problems that faced the individual in a lot of life situations, as that of mathematics play an important role in the progress of societies, which seeks to be a sophisticated societies scientifically and technically; so it is a key material for each grade in the different educational stages.

Teaching students mathematics requires that the teacher is eligible qualified academically and educationally, so that he can provide mathematical knowledge to students in a meaningful way, and for students of the faculties of educational sciences who are preparing to be teachers of the basic stage students, they are not specialists in the study of mathematics; so they may be fall in conceptual mistakes result from a mismatch between the student-teacher skills and the curriculum or instructional material (Daly, Witt, Martens, & Dool, 1997; Gravois & Gickling, 2002). For this reason, students of the faculties of educational sciences need to study academic courses, aim to equip them with the necessary mathematical knowledge to qualify them to teach mathematics.

The study aimed to develop a standardized test to measure the level in numerical ability for students of educational sciences in Jordan.

1.3 Research Questions

- Are items parameters of the numerical ability test (difficulty & discrimination) acceptable values?
- Are items of the numerical ability test effective to measure what dimensions and test measure?
- What are the indicators of the numerical ability test reliability?
- What are the indicators of the numerical ability test validity?

1.4 Limitations of the Study

- The interpretation and dissemination of the results of this study is limited by a sample chosen from students of the faculties of educational sciences in Jordanian universities for the first semester of academic year 2014/2015.
- The results generalizations depend on psychometric properties of the numerical ability test.

1.5 Significance of the Study

From the importance of mathematics, this study provide directly a practical guide that detect the level of numerical ability, which in turn helps to understand in-depth math, as the teachers are responsible for the quality of mathematical tasks that involved with the students, and they play an essential role in the mathematical understanding of students. In addition to this, it introduces a test to determine the level in mathematics among students in the faculties of educational sciences includes a number of numerical aspects in detail, components and dimensions. This leads to more effective programs in the process of teacher preparation, plan training, directing and in teacher training standards and building models of promote thinking. The results of this study may indirectly benefit in the development of curricula and textbooks of mathematics, and include the teachers' guides with appropriate activities for different levels of numerical ability. This study has opened up new avenues in the research for strategies that teachers should focus on teaching in the classroom for the development of numerical sense among students.

1.6 Procedural Definitions

Numerical ability: is the ability to deal with numbers, and do solving simple problems (Othman, 2006, p. 87); it is intended to make a major speed accuracy of numerical operations (+, - , ×, ÷). Numerical ability in this research includes: numerical justification, numerical logic, numerical computation, numerical estimation, and data representation.

Rationing test: is a preparation of the items that requires psychometric properties of: discrimination include in the interval [0.0, 0.2], and difficulty parameters include in the interval [-2.0, 2.0] in order to acquire the connotations of test validity and reliability to an acceptable degree.

Students of Education: students who are studying in multidisciplinary educational sciences in Jordanian universities.

2. Method

2.1 Participants

The study sample included (504) students of educational sciences at Jordanian universities, they were selected randomly from 8 universities (Jordanian, Yarmouk, Hashemite, Al-albeit, Technical Tafileh, Islamic science, Zarqa, Educational Science and Arts/UNRWA).

2.2 Pilot study

Five domains for numerical ability were detected from literature review (Falaye, 2006; Wei, Yuan, Chen, & Zhou, 2012; Hyde et al., 2011; Sasanguie, Global, Moll, Smert, & Rognvoet, 2013): numerical reasoning (14), numerical logic (9), numerical arithmetic (34), numerical estimation (15), and data interpretation (8).

The first draft of test contains 80 items applied on 109 students (3classes) that were randomly selected (by cluster sample) from Educational Science and Art Collage/UNRWA.

Depending on classical test theory CTT, Parameters of the test items were estimated (difficulty & discrimination). The items with difficulty parameter less than 0.2 or greater than 0.8 and discrimination parameter less than 0.39 were deleted.

2.3 Research Instrument

A test was prepared to examine the level of students in numerical ability, in the first phase of the development of the test; five appearances for the numerical ability were identified in the light of a review of studies on numerical ability and literature that dealt with its skills and manifestations.

It is consisted of (45) items, distributed into five domains: numerical reasoning (8) items, numerical logic (3) items, numerical arithmetic (21) items, numerical estimation (9) items, and data interpretation (4) items.

Assumptions of item response theory were evaluated: the dimensionality of the test was evaluated depend on residual analysis; NOHARM program estimated root mean square of residuals RMSR (closed to 0.0), and tanaka index (closed to 1.0), this results an indicators of appropriateness between the number of the test dimensions and

data

Furthermore, factor analysis for each dimension calculated; the ratio of variance greater than 20% (numerical reasoning (24.008%), numerical logic (50.524%), numerical arithmetic (20.268%), numerical estimation (22.690%), and data interpretation (34.082). this is an indicator of unidimensionality for each dimension.

Local independent for each dimension were estimated depending on index Q_3, the absolute value of Q_3 between all pairs of items in the same dimension less than 0.05, this is an indicator of availability of local independent in data.

The regression of examinee item performance (probability of success) and the set of traits (abilities) assumed to be influencing item performance were represented by item characteristic function (Appendix A), ICC' s illustrates the monotonically increasing function for all items.

3. Results

To answer the first research question: "Are items parameters of the numerical ability test (difficulty & discrimination) acceptable values?" Test items parameters were estimated by NOHARM program, and Table 1 reveals these results.

Table 1. Items parameters: difficulty, and discrimination

# item	difficulty	discrimination	# item	difficulty	Discrimination
1	0.65	0.63	24	-0.03	0.71
2	-0.8	0.77	25	-0.02	0.63
3	-0.19	0.49	26	-0.02	0.86
4	-0.69	0.65	27	-0.6	1.10
5	-0.75	0.40	28	-0.35	0.42
6	-0.42	0.47	29	0.11	0.65
7	-0.23	0.73	30	-0.39	0.71
8	-0.43	0.98	31	-0.22	0.61
9	-0.09	0.83	32	-0.52	0.26
10	-0.32	1.33	33	-0.81	0.70
11	0.03	0.71	34	-1.09	0.34
12	-0.91	0.33	35	-1.12	0.51
13	-0.25	0.66	36	-0.92	0.65
14	-0.08	0.59	37	-0.05	0.66
15	-0.04	0.96	38	-0.89	0.82
16	-0.82	0.60	39	-0.91	1.50
17	-0.16	0.70	40	0.03	0.50
18	0.26	1.06	41	-0.14	0.99
19	0.54	1.43	42	-0.16	0.40
20	0.25	0.78	43	0.03	0.56
21	-0.3	0.56	44	-0.79	0.56
22	0.65	0.58	45	-0.16	0.35
23	-0.8	0.83			

Hambleton, Swaminathan & Rogers (1991) mentioned that the value of difficulty parameter vary (typically) from about -2.0 to +2.0. Values near -2.0 correspond to items that are very easy, and values near +2.0 correspond to items that are very difficult. So the results in table 1 revealed acceptable values for the difficulty parameter [-1.12, 0.65].

Regard to discrimination parameter (Hambleton, Swaminathan, & Rogers, 1991) defined the usual rang is [0, 2], High value results in item characteristic curve that is very steep. So the results in table 2 revealed acceptable values for the discrimination parameter [0.26, 1.50].

To answer the second research question: "Are items of the numerical ability test effective to measure what dimensions and test measure?". The correlation coefficient between item marks and dimension marks and between items marks and total marks on test were estimated, and Cronbach's Alpha if item deleted were estimated. Table 2 reveals these results.

Table 2. Item-dimension, Total correlation and Cronbach's Alpha were item deleted

# item	Item-dimension Correlation	Item-Total correlation	**Cronbach's Alpha if Item Was Deleted**	# item	Item-dimension correlation	Item-Total Correlation	**Cronbach's Alpha if Item Was Deleted**
1	0.396**	.290**	0.815	24	0.466**	.422**	0.808
2	0.496**	.373**	0.812	25	0.422**	.398**	0.809
3	0.500**	.352**	0.810	26	0.422**	.379**	0.809
4	0.466**	.320**	0.810	27	0.525**	.491**	0.805
5	0.051**	.109*	0.822	28	0.270**	.268**	0.813
6	0.457**	.241**	0.813	29	0.368**	.348**	0.810
7	0.471**	.333**	0.810	30	0.434**	.403**	0.809
8	0.576**	.514**	0.805	31	0.326**	.315**	0.811
9	.136**	.370**	0.809	32	0.156**	.117*	0.816
10	.091*	.355**	0.810	33	.299**	.261**	0.812
11	.167**	.388**	0.809	34	.128**	.034*	0.819
12	0.051*	.026*	0.819	35	.434**	.229**	0.814
13	0.443**	.425**	0.808	36	.176**	.058**	0.817
14	0.415**	.391**	0.809	37	.337**	.189**	0.814
15	0.561**	.519**	0.805	38	.531**	.355**	0.810
16	0.148**	.180**	0.823	39	.603**	.444**	0.807
17	0.401**	.348**	0.810	40	.462**	.332**	0.811
18	0.488**	.428**	0.808	41	.578**	.432**	0.808
19	0.512**	.469**	0.806	42	.472**	.137**	0.816
20	0.448**	.417**	0.808	43	.608**	.325**	0.811
21	0.411**	.351**	0.810	44	.666**	.383**	0.809
22	0.445**	.405**	0.809	45	.548**	.196**	0.813
23	0.524**	.475**	0.807				

*Correlation is significant at the 0.05. **Correlation is significant at the 0.01.

Table 2 revealed that Item-dimension correlation is significant at the 0.01 level for all test items excepting items: 5, 12 a significant at the 0.05 level. This shows the effectiveness of the test items to measure what dimension measured.

Item-Total correlation is significant at the 0.01 level for all test items excepting items: 10, 12, 32, 34 a significant at the 0.05 level. This shows the effectiveness of the test items to measure what test measured.

Also Table 2 revealed the values of Cronbach's Alpha if item was deleted (0.805, 0.822), this convergent values provide more evidence for items effectiveness.

To answer the third research question: "What are the indicators of the numerical ability test reliability?"

The coefficient of internal consistency for the study instrument was calculated by using Cronbach Alpha formula for the test, and found to be equal to (0.828) which is accepted as a consistency value, with no difference between male (alpha = 0.88) and female (alpha = 0.86).

The average of inter-item correlation r equals 0.084, and this analysis offers an index of internal consistency reliability and investigates the existence of possible gender differences in the Manifestation of Numerical ability.

In item response theory, the concept of reliability is related to Item Information function Ii(θ) and test Information function IT(θ), and standard error of the estimate subjects abilities SEE, as explained by (Thissen, 2000). This is the best way to assess the reliability coefficient depending on the test information function. The relationship between reliability test information by the following equation:

$$R_{xx} = 1 - \frac{1}{\sum_{i=1}^{n} I(\Theta)} \quad (1)$$

Where R_{xx}: Test reliability, I(θ): Item Information function.

The estimated value of test reliability (by equation (1)) was 0.91(greater than Cronbach alpha). This establishes that the internal consistency of the test is good.

To answer the fourth research question: "What are the indicators of the numerical ability test validity?"

Content Test validity was verified by (8) arbitrators from supervisors and faculty members. They provide some comments and suggestions related to language expressions, and then the test was modified to fit these comments.

A factor analysis (analyzed by SPSS), using principal components and varimax rotation, was used to explore whether the test measures more than one construct. Results revealed five factors that accounted for 30.2% of the variance. (see Table 3).

Table 3. Total variance explained by the result of factor analysis

Component	% of Variance	Cumulative %
1	8.245	8.245
2	6.685	14.930
3	6.507	21.437
4	5.185	26.622
5	3.623	30.245

Hattie (1984) Confirmed that when the model corresponds to the data, the indicators through item response theory (IRT) models look more logical indicators in determining the dimensional data, these indicators are tested by the analysis of the residuals that it's estimated by NOHARM program. The software estimates: the Residuals matrix, Sum of Squares of Residuals (SSR), Root Mean Squares of Residuals (RMSR), and Tanaka Index.

In this study RMSR = 0.007 (closed to 0.0), and Tanaka Index = 0.97 (closed to 1.0), this is confirmation that the data dimensional is appropriate.

4. Discussion and Conclusion

A test to measure numerical ability for educational college students in Jordan was developed using Item Response Theory (IRT). A pool of 45 items has been developed. The dimensionality analysis indicates that the numerical ability test refers five domains (numerical reasoning (8) items, numerical logic (3) items, numerical arithmetic (21) items, numerical estimation (9) items, and data interpretation (4) items).

Items parameters (difficulty, discrimination) for 45 items remained (final draft of test) acceptable values. And the effectiveness of items to measure what test and dimensions measures quizzed by: Item-dimension correlation, Item-Total correlation, and Cronbach Alpha was item deleted.

The study investigating Psychometric properties of numerical ability test; test reliability was check out by Cronbach Alpha, average of inter-item correlation, and test information function (IRT). And test validity check out by: arbitrator's views, and factor analysis. Also test validity estimated by: RMSR, and Tanaka Index, this

result indicates that to the factor construct of the numerical ability test concert with numerical ability concept.

From reliability indicators, results revealed acceptable values for internal consistency of test, and for each dimension, this results establish the test confidence and used it as indicator of numerical ability level for educational faculty students.

In general, a test of numerical ability for students of educational sciences was developed with acceptable psychometric properties. This test can be used as exploring measure for strength and weaknesses in numerical ability to support the strength and evaluate the weaknesses by using appropriate strategies and activities for teaching numerical ability.

Numerical ability test for students of education can be used to categorize students depending on test dimensions.

Future research on the test reliability and validity is recommended, and relationship between the test and other measures (intelligence and achievement) can be quizzed.

Acknowledgments

At the end of this research, we appreciate the Faculty of Educational Sciences & Arts (FESA)/UNRWA for funded and unlimited support. We give thanks also any personal assistance given, such as in manuscript preparation.

References

Abed, E., Al-shayeb, A., & Abu Zeineh, F. (2015). Developing a mathematical thinking test for high basic stage students in Jordan. *Journal of Psychological Sciences*, *16*(3), 25-41.

Adedoyin, O., & Mokobi, T. (2013). Using IRT psychometric analysis in examining the quality certificate mathematics multiple choice examination test items. *International Journal of Asian Social Science*, *3*(4), 992-1011.

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME). (1999). *Standards for Educational and Psychological Testing*. Washington DC: American Educational Research Association

Betz, K., Eickhoff, R., & Sullivan, F. (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools*, *44*(2), 133-146. http://dx.doi.org/10.1044/0161-1461(2012/12-0093)

Costas, H. (2014). Commercial versus internally developed standardized tests: Lessons from a small regional school. *Journal of Education for Business*, *89*(1), 42-48. http://dx.doi.org/10.1080/08832323.2012.740519

Cueto, S., Leon, J., Guerrero, G., & Munoz, I. (2009). *Psychometric characteristics of cognitive development and achievement instruments in Round 2 of Young Lives*. Young Lives Technical Note, No. 15.

Daly, E., Witt, J., Martens, B., & Dool, E. (1997). A model for conducting a functional analysis of academic performance problems. *School Psychology Review*, *26*, 554-574.

Evans, J. (2013). *Problems with standardized testing*. Retrieved October 25, 2014, from http://www.education.com/reference/article/Ref_Test_Problems_Seven/

Falaye, F. V. (2006). Numerical ability course of study and gender differences in students' achievement in Practical Geography. *Research in Education*, *76*, 33-42. http://dx.doi.org/10.7227/RIE.76.3

Fraser, C., & McDonald, C. K. (1988). *NOHARM: A Computer Program for fitting both unidimensional and multidimensional normal Ogive models of latent trait theory*. New South Wales, Australia: Center for Behavioral Studies, The University of New England

Galli, S., Chiesi, F., & Primi, C. (2008). The construction of a scale to measure mathematical ability in psychology students: An application of the Rasch model. *Testing Psychometrics Methodology in Applied Psychology*, *15*(1), 3-18.

Galli, S., Chiesi, F., & Primi, C. (2010). Assessing mathematics competence in introductory statistic courses: An application of the Item Response Theory. *Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. Retrieved from www.stat.auckland.ac.nz/~iase/publications.php

Gravois, T., & Gickling, E. (2002). *Best practices in curriculum-based assessment*. In A. Thomas, & J. Grimes (Eds.), *Best practices in school psychology IV* (Vol. 2, pp. 885-898). Bethesda, MD: National Association of School Psychologists.

Hambleton, R. K., Swaminathan, H., & Rogers, H. (1985). *Fundamentals of Item Response Theory*. SAGE publication: Newbury Park, California. http://dx.doi.org/10.1207/s15327906mbr1901_3

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research*, *19*, 49-78. http://dx.doi.org/10.1207/s15327906mbr1901_3

Higginson, I. (1992). *The development, validity, reliability and practicality of a new measure of palliative care: the Support Team Assessment Schedule* (Doctoral thesis, University of London).

Hyde, D. C., Winkler-Rhoades, N., Lee, S., Izard, V., Shapiro, K., & Spelke, E. (2011). Spatial and numerical abilities without a complete natural language. *Neuro Psychology*, *49*(5), 924-936. http://dx.doi.org/10.1016/j.neuropsychologia.2010.12.017

Jamhawy, E. (2000). *Comparing Items characteristic between CTT and IRT in mathematical ability test* (Unpublished master dissertation). Al-Yarmouk University.

Jaschik, S. (2011). *Can you trust automated grading?* (Inside Higher Ed.) Retrieved October 25, 2014, from https://www.insidehighered.com

Mitchell, R. (2006). *A Guide to standardized testing: The nature of assessment*. Center for Public Education. Retrieved October 25, 2014, from http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance.

Morin, A. (2011). *What is high stakes testing?* Retrieved October 25, 2014, from http://www.http://childparenting.about.com/

NCTM. (1989). *Curriculum and evaluation standards for school mathematics*. The National Council of Teachers of Mathematics, Inc.

NCTM. (1995). *Assessment standards for school mathematics*. The National Council of Teachers of Mathematics, Inc.

NCTM. (2000). *Principles and standards of school mathematics*. The National Council of Teachers of Mathematics, Inc.

Othman, F. (2006). *The psychology of individual differences, and mental ability*. Al-Ameen press, Egypt, Cairo.

Popham, W. (2005). *Standardized testing fails the exam*. Retrieved October 25, 2014, from http://www.edutopia.org/f-for-assessment

Psychometric Success. (2013). *Numerical ability tests*. Retrieved October 25, 2014, from http://www.psychometric-success.com/aptitude-tests/numerical-aptitude-tests.htm

Sasanguie, D., Gobel, S., Moll, K., Smert, K., & Regnvoet, B. (2013). Approximate number sense, symbolic number processing or number-space mapping: what underlies mathematics achievement? *Journal of Experimental Child Psychology*, *14*(3), 418-431. http://dx.doi.org/10.1016/j.jecp.2012.10.012

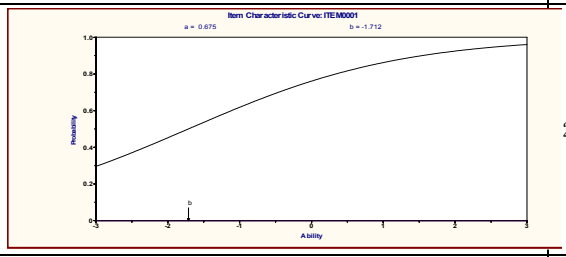
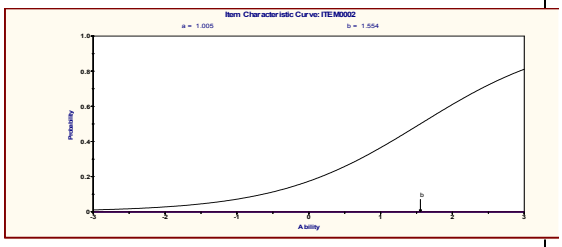
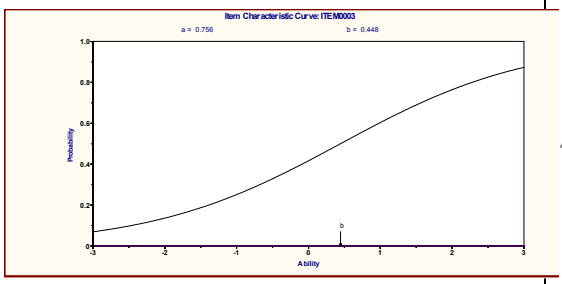
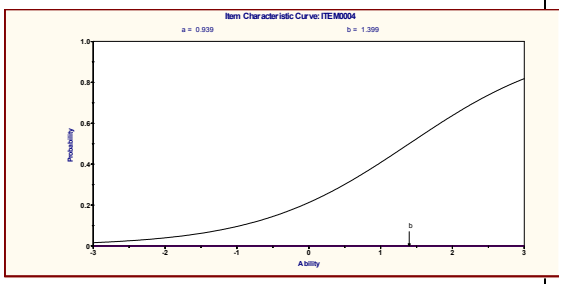
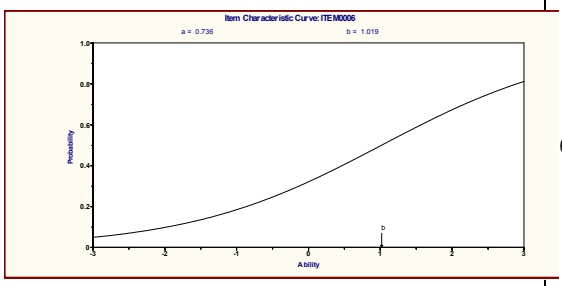
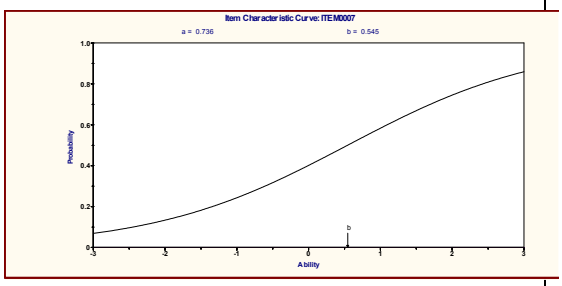
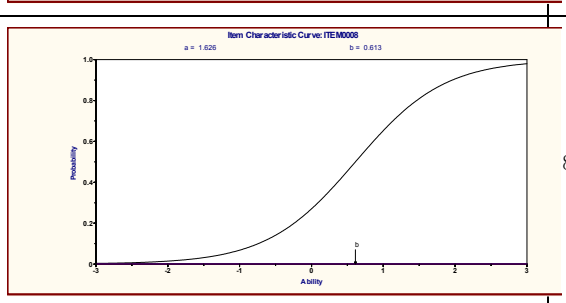
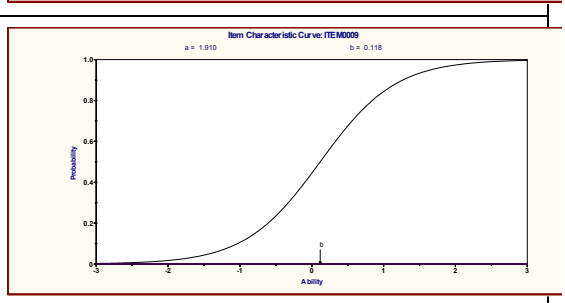
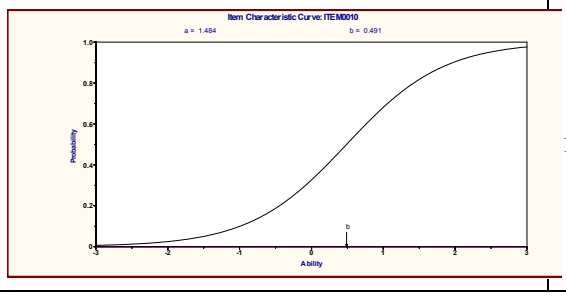
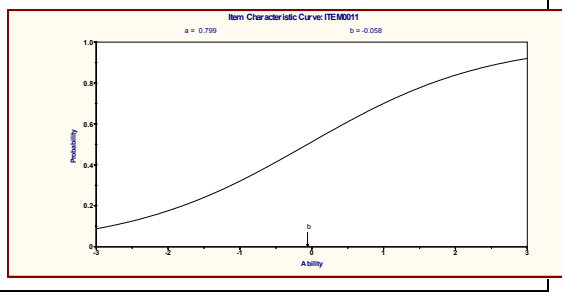
Schuwirth, L. (2010). From assessment of learning to assessment for learning. *JIAMSE*, *20*(2), 170-172.

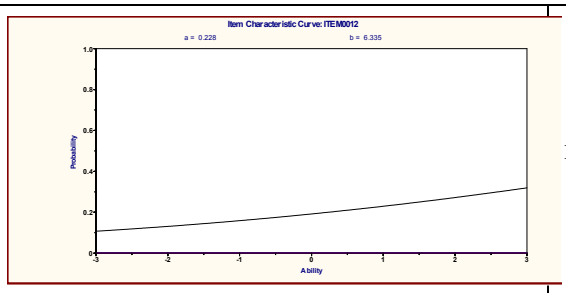
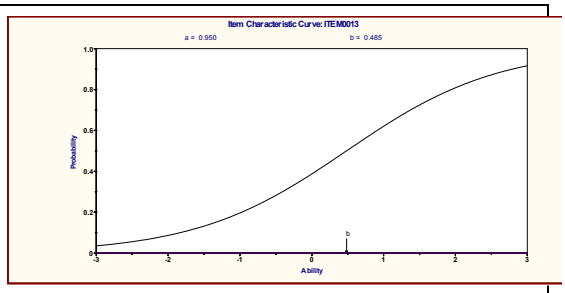
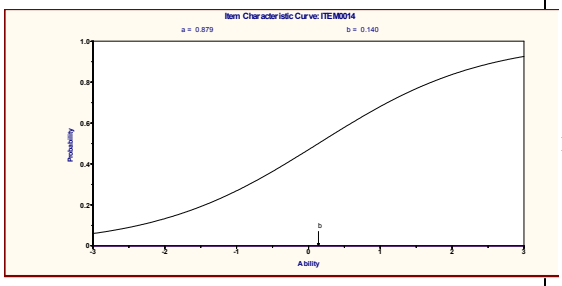
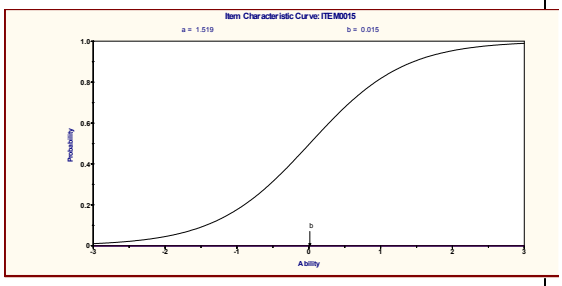
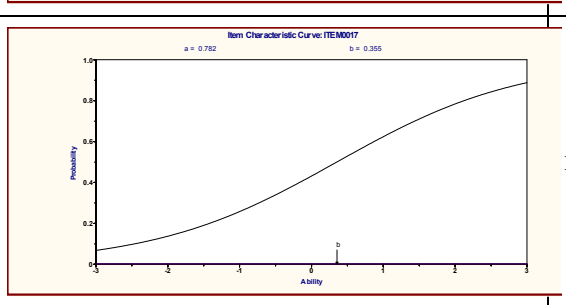
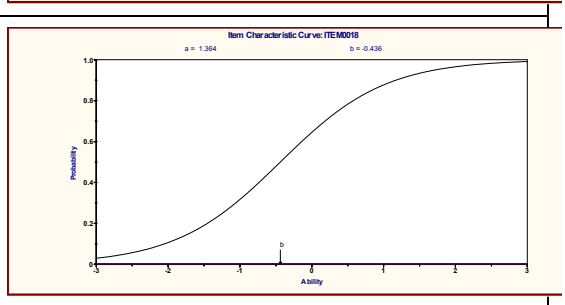
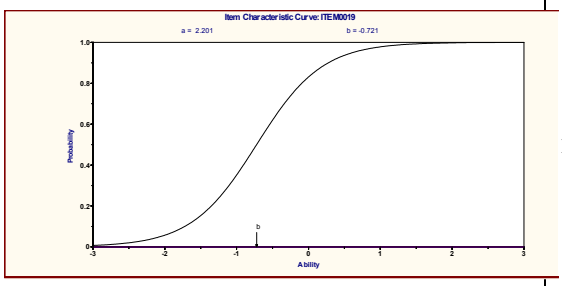
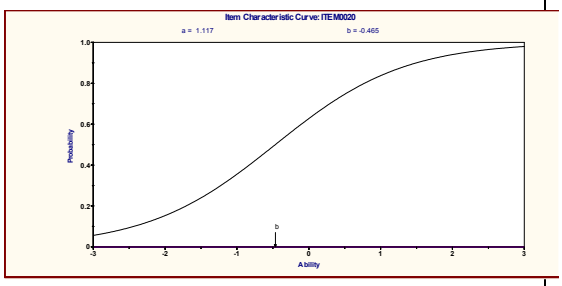
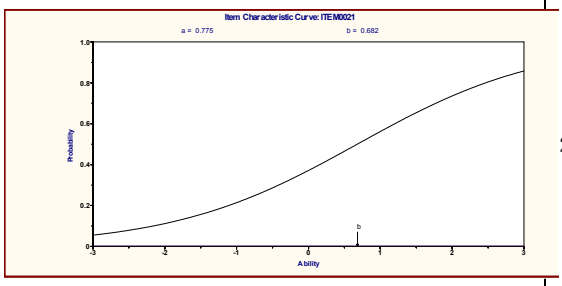
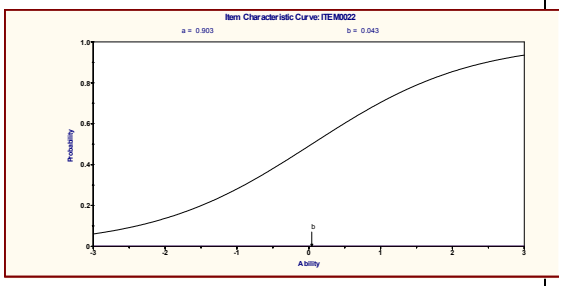
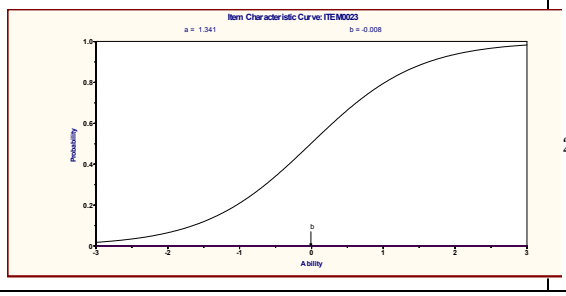
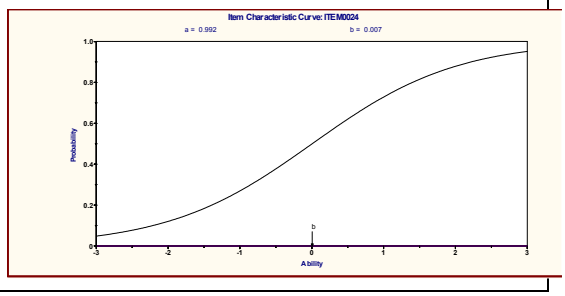
Smith, S. (2013). *Early childhood mathematics* (5th ed.). Pearson Education, Inc.

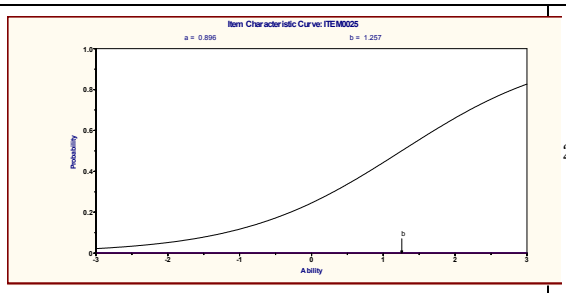
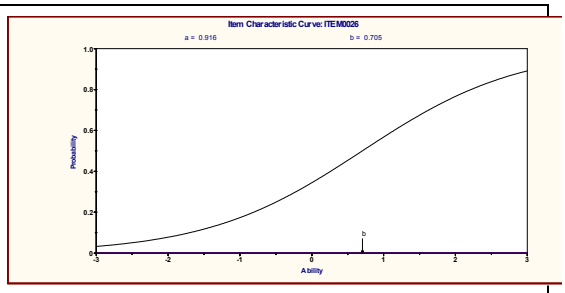
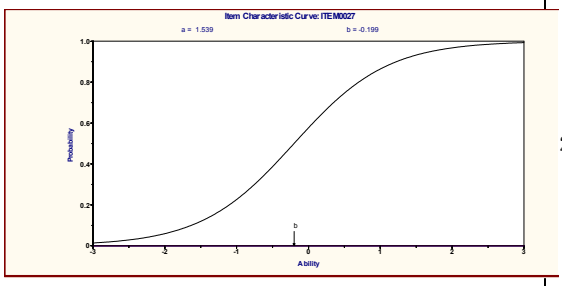
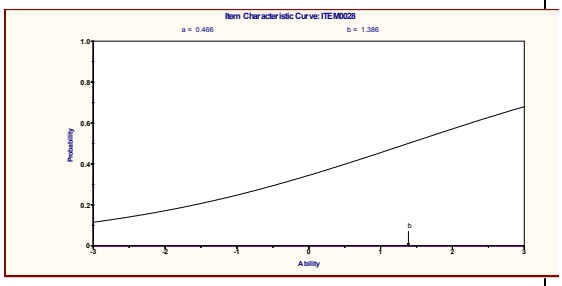
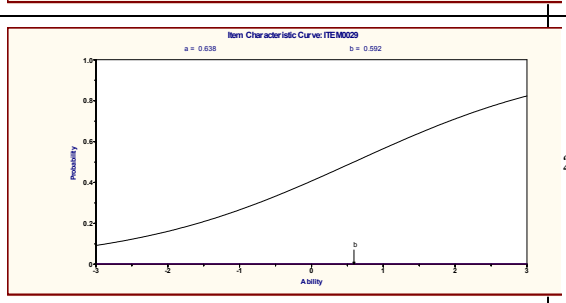
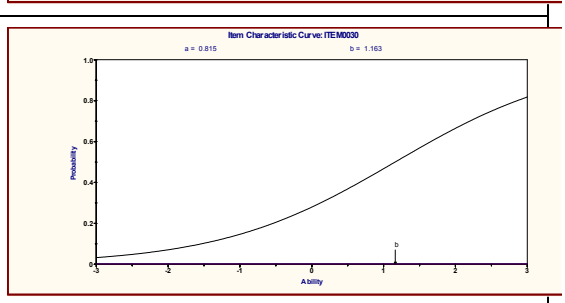
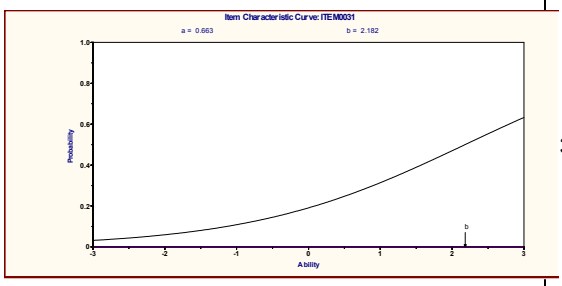
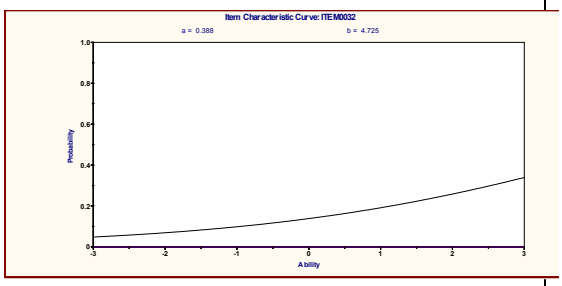
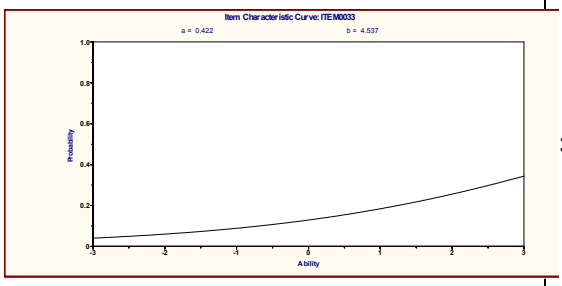
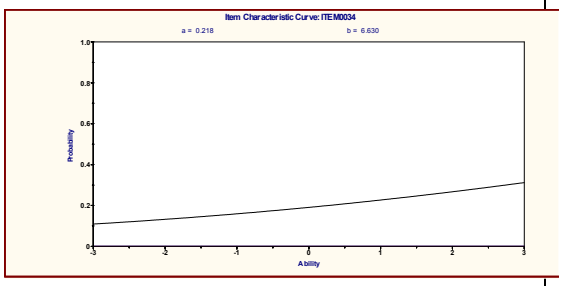
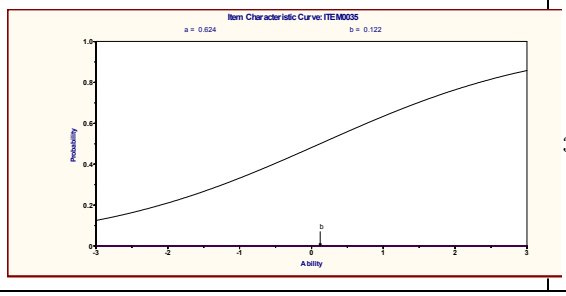
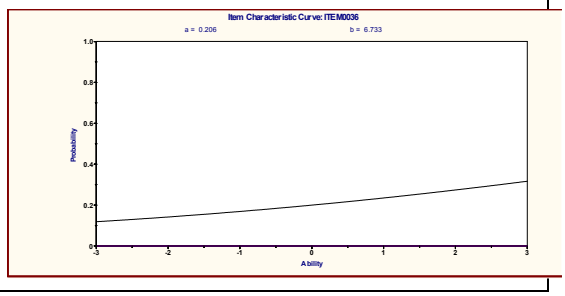
Thissen, D. (2000). *Reliability and measurement precision*. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 159-183). Mahwah, NJ: Lawrence Erlbaum Associates.

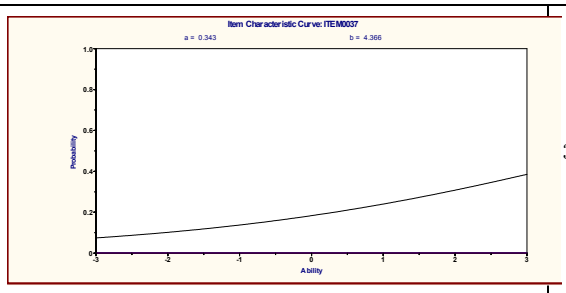
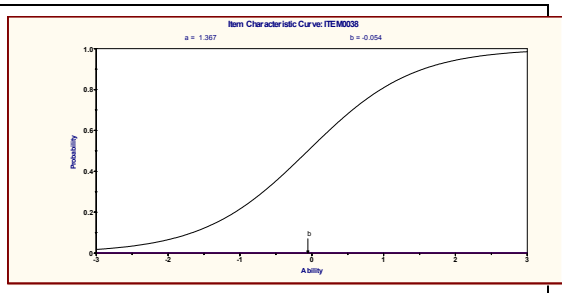
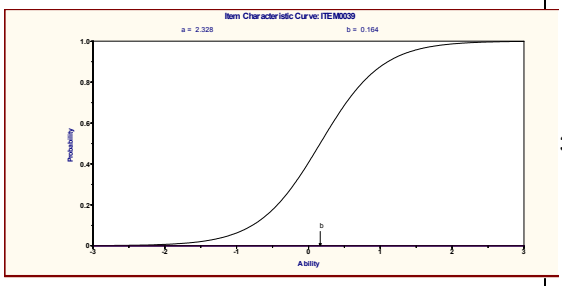
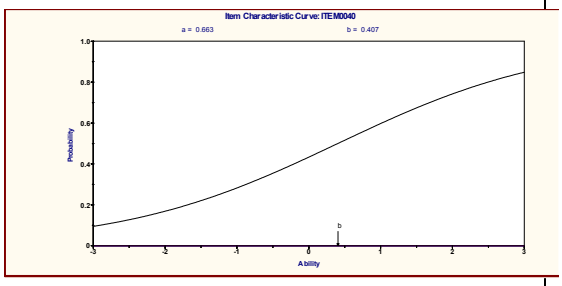
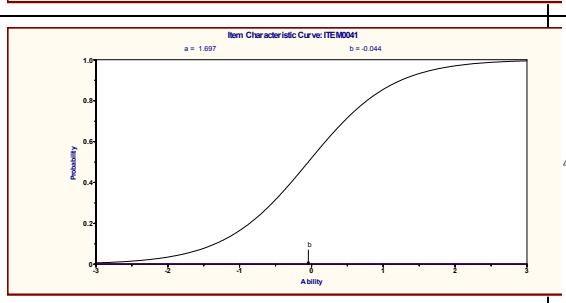
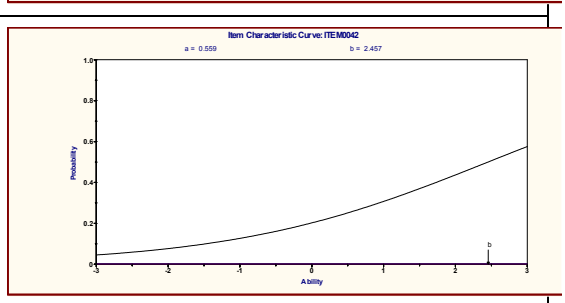
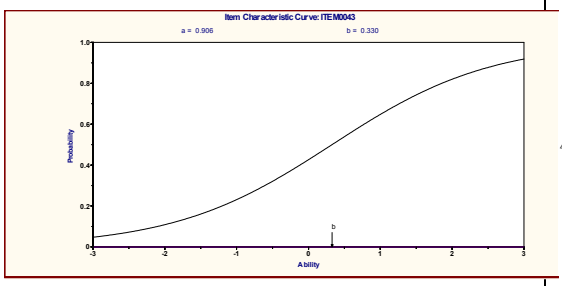
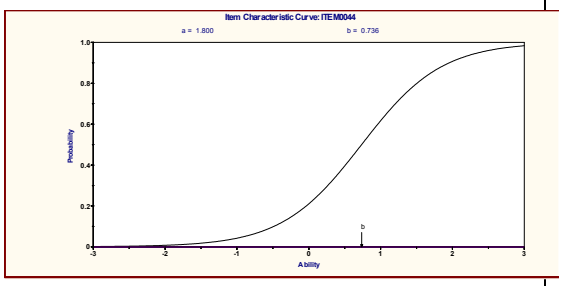
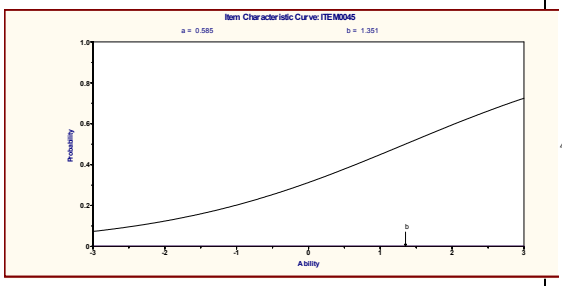
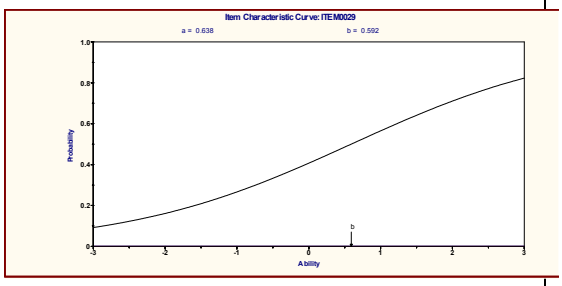
Wei, W., Yuan, H., Chen, C., & Zhou, X. (2012). Cognitive correlate of performance in educated mathematics. *British Journal of Education Psychology*, *82*(1), 157-181. http://dx.doi.org/10.1111/j.2044-8279.2011.02049.x

Appendix

# item	ICC	# item	ICC
1		2	
3		4	
5		6	
7		8	
9		10	

11		12	
13		14	
15		16	
17		18	
19		20	
21		22	

23		24	
25		26	
27		28	
29		30	
31		32	
33		34	

35		36	
37		38	
39		40	
41		42	
43		44	
45