Across Disciplines, Cultures and Technologies: An Item Response Theory Approach to Assessment of Learning

Jenny Mei Yiu Huen, The University of Hong Kong, Hong Kong
Yue Zhao, The University of Hong Kong, Hong Kong
Paul Siu Fai Yip, The University of Hong Kong, Hong Kong

The Asian Conference on Technology in the Classroom 2017
Official Conference Proceedings

Abstract
In the process of educating for change, we must strategically design assessment to examine how well our students are learning. This subject is important but easily neglected by educators or misrepresented in the education field.
This study applied the Item Response Theory (IRT), a contemporary psychometric approach that models the relationship between the response to individual items on a test and the underlying abilities being measured, to examine the psychometric properties of binary (true-or-false) question items designed to assess how much a sample of 209 Hong Kong Chinese students have learned in a web-based learning program accompanied with classroom teaching. The IRT analysis procedure was illustrated, from checking model assumptions, calibrating items to assessing goodness-of-fit.
Principal results of this study offered information for estimating item discrimination and item difficulty for each question item and provided test information to indicate how well the test contributes to the assessment of learning along a continuum ranging from low to high abilities. In this direction, the IRT approach offers useful information for design, diagnosis and revision of test items. For example, items with high discrimination are particularly useful and should be retained, whereas items with low difficulty are not particularly useful and could be considered for removal.
In conclusion, this study put forward an IRT approach that can be widely applied to design and modify test items such that assessment of learning can be better suited to the discipline, culture and technology in context.

Keywords: Item Response Theory; Assessment; Learning

Introduction

With the emerging technologies in teaching and learning, such as digital instructional materials, holographic technologies, Massive Open Online Courses, online platforms, Radio Frequency Identification technology and social media (IAFOR, 2017), educators have been focusing the discussion on the usage of these technologies in the classrooms. A pertinent subject seldom being discussed and implemented rigorously by educators is the assessment of students' learning in response to the technology being used, the discipline being taught, and the culture in context. Assessment of learning informs educators how much students have learned or mastered and identifies the gap between intended learning goal and current level of learning (e.g. Mok, 2012). Students' abilities or learning have been measured through various forms of assessment, usually through tests (Chatterji, 2003). In this direction, educators should ensure that tests, especially those constructed by teachers for assessment and evaluation purposes, are rigorously constructed. Several psychometric approaches have been well established in the assessment literature which can enhance the psychometric properties of tests. Among them, Item Response Theory (IRT) is a contemporary psychometric approach that models the relationship between the probability of response to individual items on a test and the underlying abilities being measured by the test (e.g. Hambleton, Swaminathan, & Rogers, 1991). IRT could not only provide validation to large-scale standardized tests (such as GRE tests), but also provide valuable information for educators to enhance small-scale and locally-constructed tests.

Built on the above background, the purposes of this study were to apply IRT to examine and enhance the psychometric properties of the test questions of a small-scale knowledge test constructed for a web-based learning program accompanied with classroom teaching for Primary 4 and Primary 5 students in Hong Kong.

Methodology and Methods

The program

The program was designed to promote positive attitudes and values to local Primary 4 and Primary 5 students in Hong Kong. The contents covered in the program include brain-based learning strategies, emotional management skills, thinking errors identification, problem-solving skills, effective communication skills, active listening, thankfulness towards others, etc. The program was a combination of a web-based program (11 e-lessons in total) and classroom teaching (8 classroom lessons in total). The e-lessons were used to prepare students for the content of upcoming classroom lesson. During classroom teaching, interactive activities such as role play, group discussion and card games were used to facilitate the teaching and learning.

The test

The test consisted of 10 true-or-false questions assessing students' knowledge on the contents covered in the program. The test was administrated to students in a classroom session upon their completion of the whole program. The students were asked to decide whether each of the ten statements was true or not. One sample statement, translated from Chinese to English, is "Negative emotions, such as anger

and anxiety, are something which we should not have even to a mild extent.". For any correct answer, students would score 1 point and 0 for incorrect answers. Hence, they would be able to score a minimum of 0 point to a maximum of 10 points.

Participants

Participants in this study consisted of 150 Primary 4 students (M = 9.33 years, SD = 0.72 years) and 59 Primary 5 students (M = 10.10 years, SD = 0.30 years) who participated in the program, completed all the questions in the test, and gave consent to the use of their test data for research purpose. Thus, the sample size was 209 (Male = 99; Female = 110) for the analyses.

Analyses

In an IRT analysis procedure, the analyses conducted include: 1) checking model assumptions on test dimensionality, 2) fitting the two-parameter IRT model and calibrating item parameters, and 3) assessing goodness-of-fit. These analyses were conducted using SPSS Statistics Version 23.0 (IBM Corp., 2015) and IRTPRO (SSI Inc., 2015).

Results and Discussion

The descriptive statistics of test scores are reported in Table 1.

Table 1: Test Score Descriptive Statistics

Test Question	Descriptive Statistics (N = 209)			
	Response frequencies (%)[a]			
	1	0	*M*	*SD*
Question 1	94.3	5.7	.94	.23
Question 2	89.5	10.5	.89	.31
Question 3	85.2	14.8	.85	.36
Question 4	90.0	9.1	.91	.29
Question 5	42.6	57.4	.43	.50
Question 6	28.2	71.8	.28	.45
Question 7	50.7	49.3	.51	.50
Question 8	82.3	17.7	.82	.38
Question 9	96.7	3.3	.97	.18
Question 10	60.8	39.2	.61	.49

Note. N = sample size; M = mean; SD = standard deviation.
[a] Response score categories are: 1 = *correct response* and 0 = *incorrect response*.

The test score distribution for the 209 students had a mean score of 0.72 (1 point as maximum) with a standard deviation of 0.25. Over 90% of the students got Question 1, Question 4 and Question 9 correct. From the descriptive statistics above, some questions may be easy for students.

Unidimensionality of data is an important model assumption to be checked before fitting the data with the unidimensional IRT model. Results from exploratory factor analysis on our test data showed that the first eigenvalue and its percentage of total

variance explained were greater than that of the second eigenvalue (see Table 2), suggesting one dominant factor/dimension which accounted for about 20% of the variability. Further, from the scree plot showing the eigenvalues on y-axis and the number of factor(s)/dimensions(s) on the x-axis (Figure 1), the slope has its greatest level-off moving from 1 factor/dimension to 2 factors/dimensions, indicating that one factor/dimension should be generated by the analysis. The above findings provide evidence that the unidimensionality assumption has been met to a reasonable degree level by our test data. In other words, we can assume that there is a single unidimensional ability in common underlying students' test performance in a set of 10 test questions.

Table 2: Eigenvalues and its Percentage of Variance Accounted for

Factor(s)/ Dimension(s)	Eigenvalue	% of variance accounted for
1	1.94	19.42%
2	1.32	13.21%
3	0.93	9.31%
4	0.87	8.66%
5	0.95	9.51%
6	0.91	9.06%
7	0.77	7.66%
8	0.77	7.74%
9	0.80	7.99%
10	0.74	7.44%

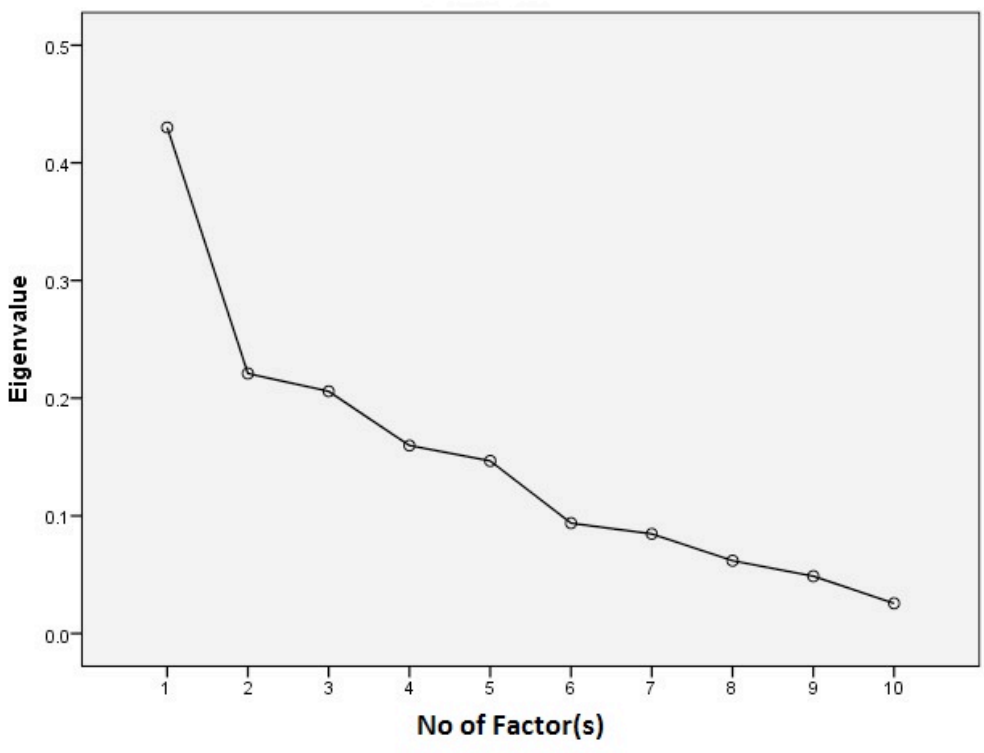

Figure 1: Scree Plot of Eigenvalues against Number of Factors.

The two-parameter IRT model (item discrimination and item difficulty as the two parameters being estimated) was then fit to the test data. The estimates obtained after item calibration are reported in Table 3. Item discrimination (denoted by *a* values in Table 3) addresses how well a test item can differentiate between students at different ability levels. The higher the item discrimination, the better the item can differentiate between students at different ability levels (such as low, medium and high abilities). From the results, the discrimination estimates range from -0.13 to 1.91. Item 1, Item 2, Item 4 and Item 9, are particularly low in discrimination, that they cannot differentiate well between students at different ability levels. Item difficulty (denoted by *b* values in Table 3) taps the difficulty level of a test item to the students. From the results, Item 1, Item 4, and Item 9 are very low in difficulty that these test items can be expected to be answered correctly by most students. This is consistent with the observation that over 90% of the students got Question 1, Question 4 and Question 9 correct (refer to Table 1). In summary, there are items in our test identified to be low in discrimination and difficulty. Considerations could be made to remove these items or revise them accordingly to enhance the psychometric properties of the test.

Table 3: Item Parameter Estimates.

Test Item	Item Parameter Estimates	
	a	*b*
Item 1	0.19	-14.78
Item 2	-0.13	16.58
Item 3	1.40	-1.66
Item 4	0.34	-7.01
Item 5	0.99	0.36
Item 6	0.78	1.35
Item 7	1.20	-0.03
Item 8	1.73	-1.31
Item 9	0.34	-9.94
Item 10	1.91	-0.36

Note. *a* = item discrimination; *b* = item difficulty.

An item level fit ($S{-}\chi^2$ statistic) statistic for each item are reported in Table 4, which served as evidence of model fit. Results showed that model fit (*p* at the .01 level) was good except for Item 8. This item could also be considered for removal or revision for a better model fit.

Table 4: Model Item Fit Statistics.

Test Item	Model Item Fit Statistics		
	χ^2	*df*	*p*
Item 1	4.69	4	0.32
Item 2	4.64	5	0.46
Item 3	11.81	4	0.02
Item 4	6.75	5	0.24
Item 5	8.70	4	0.07
Item 6	6.97	4	0.14
Item 7	2.90	4	0.58
Item 8	15.84	3	0.00
Item 9	2.44	3	0.49
Item 10	4.53	4	0.34

Note. χ^2 = Chi-Square; *df* = degree of freedom; *p* = probability.

The test characteristic curve (Figure 2) and the test information function with standard error of measurement (Figure 3) serve as a good summary of the test information. As observed from the test characteristic curve, the average student (theta = 0) could be expected to achieve a score about 70% on the test, whereas students of lower ability (theta ≈ -1.5) could be expected to achieve a score about 50% on the test. The test was thus generally easy for students. From the test information function with standard error of measurement, it can be seen that the test was providing a good level of measurement (i.e., high total test information and low standard error of measurement) for students performing from about .5 *SD* to 1 *SD* below the mean. The test may be improved in the future by replacing easy items with more difficult items which can provide good discrimination for students in the upper half of the test score distribution.

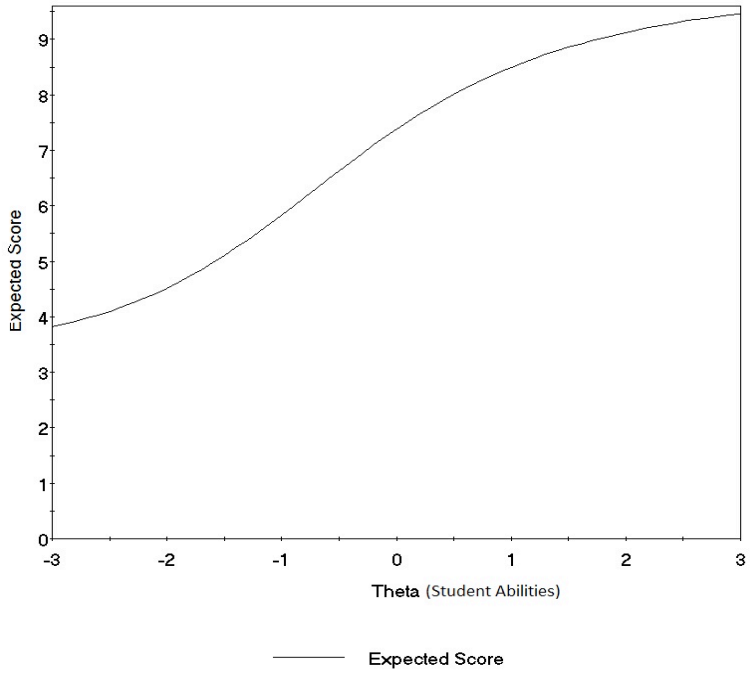

Figure 2: Test Characteristic Curve.

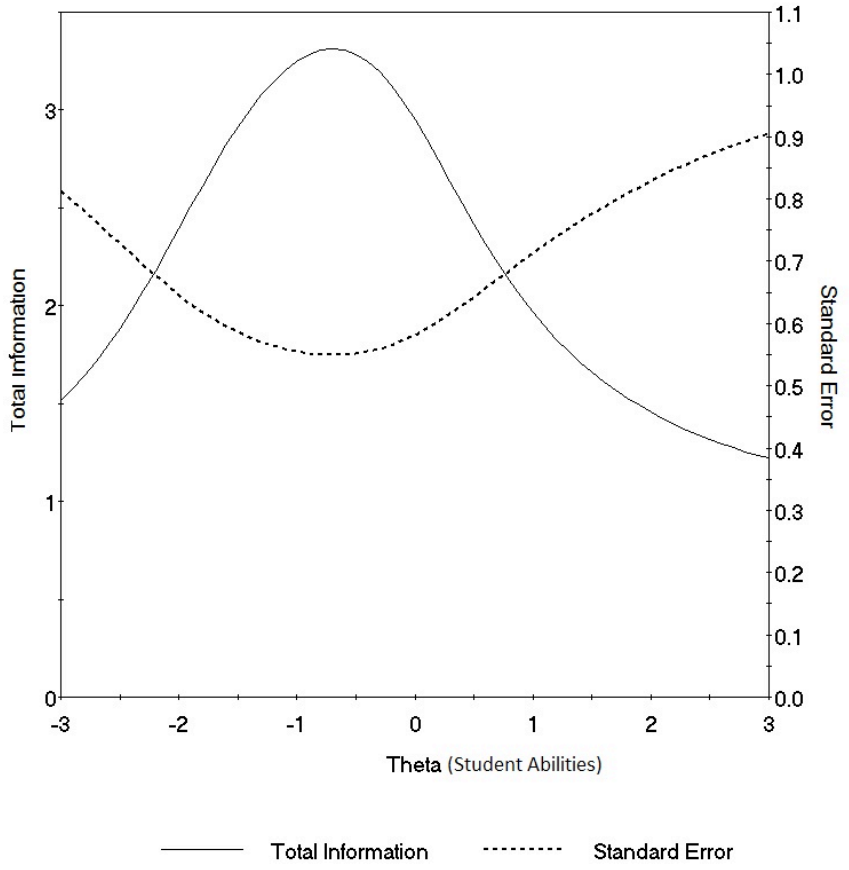

Figure 3: Test Information Function with Standard Error of Measurement.

Conclusion and Limitation

This study made use of an IRT approach to examine the psychometric properties of the test questions of a locally-constructed and small-scale test. It was noted that very few studies in the education field have focused the study on this subject before, especially in the local context. One exception was a study by Yao and Mok (2013) which applied the Rasch Model, a special kind of IRT model that estimates discrimination parameter for each item while keeping discrimination equal across all items, to develop and improve a locally-constructed Mathematics test to inform teaching and learning.

In this study, the IRT analysis procedure was illustrated with sufficient detail to make it possible for replication in other similar tests. Principal results offered useful information for design, diagnosis and revision of question items. Items with low discrimination and low difficulty were identified. These items were not particularly useful and could be considered for removal for revision. However, given the limitation of small number of test items in this test (only 10 items in total), loss of items could not be afforded. Yet, should there be revision to any of the test items, another iteration of IRT analysis procedure is required.

To end with, this study puts forward that the IRT can be widely applied to assessment of learning, across academic disciplines, cultural populations and emerging technologies.

Acknowledgements

We would like to acknowledge that the data used in this study was collected from a project "Quality Education Fund Thematic Network on Developing Students' Positive Attitudes and Values" funded by the Quality Education Fund of the Government of the Hong Kong Special Administrative Region. The project was administrated by the colleagues at the Hong Kong Jockey Club Centre for Suicide Research and Prevention at The University of Hong Kong. In particular, we thank Ms. Eliza Lai and Ms. Michelle Leung for their assistance in administrating and implementing the project and Mr. Sam So for administrating the web-based platform.

References

Chatterji, M. (2003). *Designing and using tools for educational assessment*. Boston, MA: Allyn and Bacon.

Hambleton, R. K., Swaninathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

IAFOR (2017). Special issue on "Technology in the Classroom". Retrieved from: https://iafor.org/journal/iafor-journal-of-education/

IBM Corp. (2015). *IBM SPSS statistics for Windows, Version 23.0*. Armonk, NY: IBM Corp.

Mok, M. M. C. (2012). *Self-directed learning oriented assessments in the Asia-Pacific*. New York, NY: Springer.

SSI Inc. (2015). *IRTPRO for Windows, Version 3*. Lincolnwood, IL, Scientific Software International.

Yao, J., & Mok, M. M. C. (2013). Implementing assessment for learning: An application of the Rasch model for the construction of a mathematics assessment to inform learning. *Assessment and Learning*, *2*, 29-50.

Contact email: jennymyhuen@connect.hku.hk