90Validation of School-Based College English Listening Test: Evaluating Construct Validity Using the 3PL Item Response Theory Model

Meng Lyu(✉)

Xi'an Jiaotong University, Xi'an, China
mariannelyu@stu.xjtu.edu.cn

Abstract. More and more universities are implementing school-based English exams designed and administered by their professors, yet there is limited research on the quality and reliability of school-based exams. This study aims to evaluate the validity of the school-based college English listening test items according to the three-parameter logistic model in the Item Response Theory (IRT). The test consisted of 25 items, and 944 male and female college students from a university in Western China were selected in a simple random manner in this study. The statistical program (SPSS) and the (Bilog-mg3) program were used to analyze the responses of these participants. In this research, the validity of the school-based college English listening test was assessed on two levels: whether the items conformed to the questioning norms and whether they effectively represented the construct validity. The results showed that this set of listening items were high in difficulty, ideal in differentiation, high in guessing, reasonable in information content, and with a certain degree of constructive validity, which could accurately reflect the actual listening proficiency of the intermediate and high-level test takers.

Keywords: Item Response Theory · Construct validity · Listening test

1 Introduction

Language testing is an important basis for test-taker proficiency, and has a wide range of social implications. Whether the test items are comprehensive in terms of difficulty and differentiation, and whether they measure candidates' abilities objectively, fairly, and scientifically have been issues of concern in the language testing.

The application of IRT in language testing is mainly used for analyzing the quality of subjective questions in terms of test reliability, but not enough for objective questions, which account for a large proportion of test papers. The purpose of this paper is to evaluate the validity of the school-based college English listening test items according to the three-parameter logistic model in the Item Response Theory (IRT).

X. Yuan et al. (Eds.): ICEKIM 2023, AHCS 13, pp. 2043–2054, 2023.
https://doi.org/10.2991/978-94-6463-172-2_228

2 Item Response Theory

2.1 Construct Validity

Validity is a constant theme in the field of language testing. Since Messiek [1] theoretically proposed a holistic view of validity, many scholars have conceived different validity frameworks, such as Mislevy et al.'s [2] evidence-centered approach, Kane's [3, 4] explanatory arguments, and Bachman & Palmer's [5] Assessment Use Argument.

Construct Validity refers to the extent to which performance on a test matches the predictions we make based on ability or conceptual theory [5]. Overall validity theory suggests that construct validity is the most fundamental aspect of validity and is the core of validity theory, which is the basis for determining the end of test scores and the validity of the use of the results.

2.2 Logistic Model of Item Response Theory

The one-parameter logistic model (1-PL) (see below equation) is considered to be one of the most commonly used models in the item response theory, which assumes that all items differ from each other only by the item difficulty parameter.

$$p_i(\theta) = 1/(1 + e^{(-D(\theta - b_i))}),\ i = 1, 2, 3 \ldots, n$$

The two-parameter logistic model (2-PL) (see below equation), which is proposed by Harrington [6], includes an item discrimination parameter in the model, which could be obtained from the 3-PL IRT model when the pseudo-chance parameter is assumed to be zero.

$$p_i(\theta) = 1/(1 + e^{(-Da_i(\theta - b_i))}),\ i = 1, 2, 3 \ldots, n$$

The three-parameter logistic model (3-PL) (see below equation) is an extension of the two-parameter logistic model, which is derived by adding the item guessing parameter. 3-PL contains the three possible parameters of the item, difficulty, discrimination, and guessing (bi, ai, ci). These parameters measure the probability of the answer of the examinee of the ability of (θ) to the item (i).

$$p_i(\theta) = c_i + (1 - c_i)1/(1 + e^{(-Da_i(\theta - b_i))}),\ i = 1, 2, 3 \ldots, n$$

Pi (θ) refers to the probability that an examinee with ability θ answers item i correctly. To be specific, bi is about item difficulty parameter, ai stands for item discrimination parameter, ci item is a guessing parameter, and θ represents the trait level of person i.

In practice the 3PL model fits the data better than the 2PL model, and the empirical data indicate that the 3PL model provides more stable and more accurate item parameters and test reliability.

Table 1. Factor Eigenvalue and Percentage (N = 944) (Table Credit: Original)

Factor	1	2	3	4	5	6
Eigenvalue	4.12	1.28	1.17	1.11	1.07	1.04
Percentage	16.5	5.1	4.7	4.6	4.3	4.2

3 Methodology

This is an experimental study using quantitative analysis. The random sample comprised 944 non-English major freshmen in a university in northwestern China and counterbalanced in gender. This normative sample were included to examine the construct validity of a school-based listening test. 25 objective items were adopted on the listening test, all using the 0/1 scoring method. There are three sections in the listening test. Section A contains 15 multiple choice items including 8 short conversations and 2 long conversations. Section B contains 10 items with 3 short passages. Section C is a short text fill-in-the-blank, so it is not included in the analysis of this test. In this paper, a three-parameter IRT model is adopted and Marginal Maximum Likelihood Method is used for item parameter estimation.

4 Data Analysis

4.1 Unidimensionality

In the IRT framework, unidimensionality refers to the fact that all items in a test measure the same type of ability or latent trait of the test taker. First, to test the unidimensionality of the listening test, i.e., whether the test items examined only the students' listening ability. We did a factor analysis of the listening test using SPSS 26.0. It was found that the KMO value >0.6. This meant that Bartlett's test of sphericity was significant, so the data from the listening section could be used for factor analysis. After rotation, a total of six factors were extracted (Table 1 and Fig. 1).

The factor analysis results showed that the first-factor loading was 3–4 times higher than the figure of second factor, and the scree plot indicated a clear break between the first and second factors. Therefore, the listening items of this test are unidimensional and can be analyzed using the IRT model (Table 2).

4.2 Item Parameters and Quality Analysis

a) Difficulty level.

The average difficulty of the 25 items in the listening test was 0.44, with an overall high difficulty level. There were two questions with low difficulty ($b < 0$): items 1 and 22, 16 items with moderate difficulty ($0 < b < 0.9$), and seven questions with difficulty ($b > 0.9$): items 10, 13, 14, 15, 20, 23, and 24. There was a certain degree of variation in the difficulty of the test questions, and all difficulty bands were covered. But they

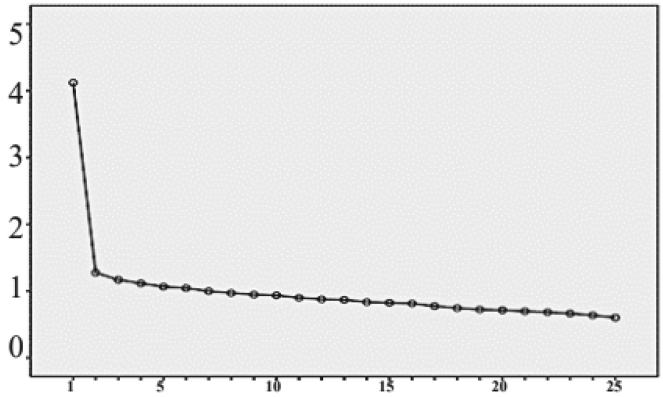

Fig. 1. Scree plot (Picture Credit: Original)

were mainly concentrated in the middle and high difficulty bands, and there were fewer questions in the low difficulty band.

b) Degree of differentiation.

The mean discrimination of the 25 items in the listening test was 0.96, which was good discrimination, and there were no questions with less than 0.5 discrimination.

c) Guessing.

The average guessing of the 25 items in the listening test was 0.35, slightly higher than 0.25, which indicated that the guessing of the test questions was relatively high. Specifically, 24 items with a guessing that is greater than 0.25, 7 items with guessing close to or greater than 0.5, including items 1, 8, 9, 16, 18, 19 and 22.

d) Information Content.

The listening test contained 25 items, accounting for 25% of the total score. When the standard error was 0.20, the items of the listening test should be 25. Then the information content of the listening items should be 6.25 or more, with 0.25 per question. The quality of the listening items was considered good when the information content of each item was more than 0.25. The items with less than 0.25 information in the listening test were: items 1, 16, 20, and 22, those closer to 0.25 are items 2, 8, 12, and 14. Therefore, the deletion of items 1, 16, 20, and 22 should be considered (Figs. 2, 3, 4, 5).

At the same time, the total test information function of the listening test is shown below (see Fig. 6):

As can be seen from the Fig. 6, when the test taker's ability value was close to 0.6, and the amount of information reached the maximum that was close to 7, it could best reflect the test taker's actual listening level. Meanwhile, when the total information

Table 2. Summary of item parameters (Table Credit: Original)

Item	Correct rate	Discrimination	Difficulty	Guessing	Load Value
1	46.4	0.931	-0.765	0.5	0.682
2	17.9	0.9	0.167	0.344	0.669
3	42	1.068	0.172	0.251	0.73
4	18.3	1.052	0.178	0.35	0.725
5	40.3	1.013	0.046	0.3	0.712
6	19.9	0.953	0.161	0.26	0.69
7	47.8	1.281	0.311	0.303	0.788
8	20.8	0.951	0.41	0.47	0.689
9	48.8	1.027	0.225	0.45	0.716
10	14.7	0.862	0.933	0.299	0.653
11	34.1	0.896	0.12	0.298	0.667
12	19.2	0.83	0.513	0.363	0.639
13	43.9	1.304	1.615	0.387	0.793
14	19.1	0.726	1.201	0.304	0.587
15	41.2	0.83	1.552	0.23	0.639
16	13.4	0.738	0.158	0.459	0.594
17	31.5	0.83	0.539	0.354	0.638
18	15.5	0.917	0.125	0.429	0.676
19	31.4	1.213	0.222	0.433	0.771
20	9.8	0.647	0.969	0.275	0.543
21	25.2	0.936	0.139	0.29	0.683
22	21.5	0.856	-0.356	0.482	0.65
23	49.3	1.433	0.963	0.286	0.82
24	17.7	0.822	1.37	0.235	0.635
25	41.0	0.88	0.085	0.358	0.661

content was more significant than 6.25, the candidate's ability range was about 0.2–1.2, this set of listening items was suitable for candidates with above-average test levels.

e) Loading value.

The mean loadings of the 25 items on the listening test were 0.68, which meant that the construct validity of the listening test consisting of 25 items was 0.68, which was relatively ideal. If items 14, 16, and 20 are deleted, the construct validity would be improved to 0.69.

To summarize, based on the comprehensive assessment of the four dimen-sions of difficulty, discrimination, guessing, and information content of each item, the researcher

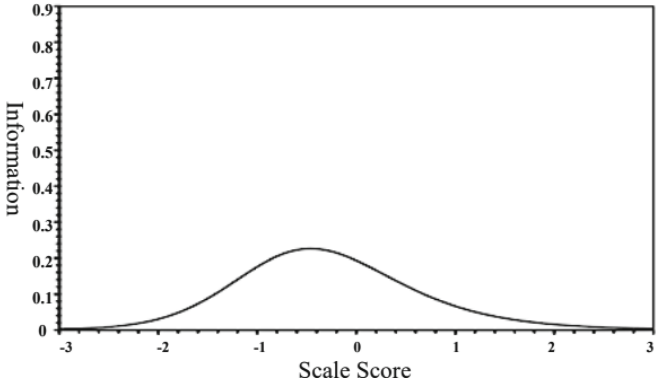

Fig. 2. Item Information Curve: ITEM001 (Picture Credit: Original)

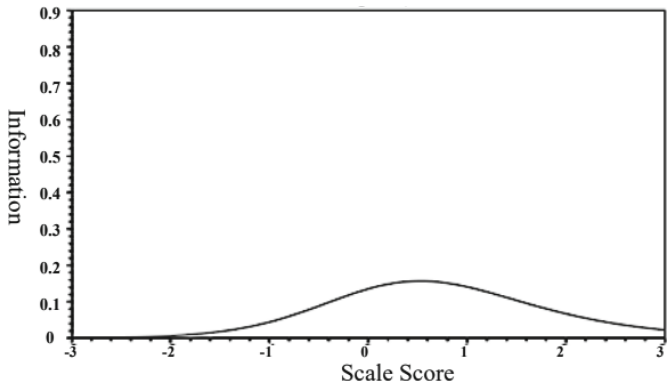

Fig. 3. Item Information Curve: ITEM0016 (Picture Credit: Original)

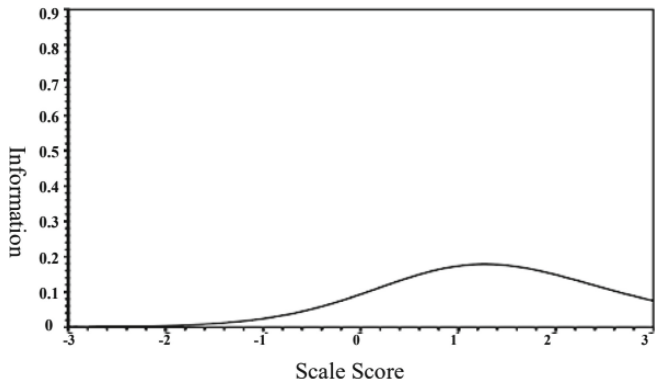

Fig. 4. Item Information Curve: ITEM0020 (Picture Credit: Original)

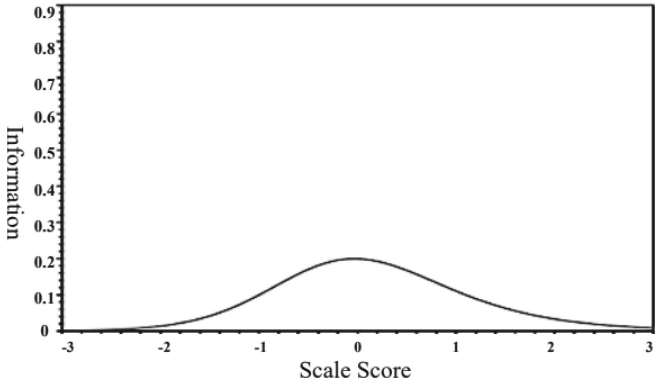

Fig. 5. Item Information Curve: ITEM0022 (Picture Credit: Original)

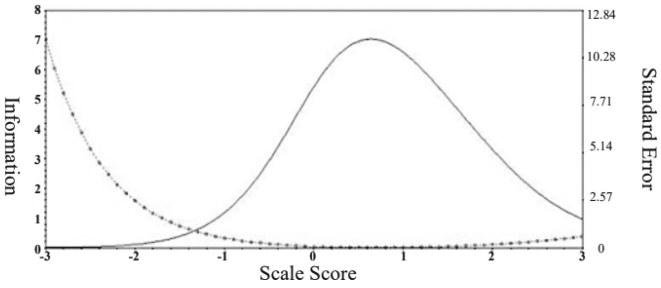

Fig. 6. Total test information curve (Picture Credit: Original)

has listed the quality items that meet the questioning norms and the items that do not meet the norms that should be considered for deletion.

4.3 Analysis of High-Quality and Non-standard Items

a) High-quality items.

The researcher considers six quality questions that meet the specifications of the questions, including questions 3, 5, 7, 11, 15, and 21. Among them, questions 3, 5, 7, 11, and 21 are moderately difficult, reasonably differentiated, relatively low in guessing and contribute a reasonable amount of information. The following analysis is based on the example of question 3, whose item characteristic curves are shown below.

As can be seen from Figs. 7 and 8, Item 3 has a moderate difficulty value, reasonable differentiation, a guess rate close to 0.25, and a significant contribution of information. In the dialogue, even if candidates do not understand the vocabulary word "vegetarian", they can guess it through the context. The question is moderately complex and can reflect the listening level of the candidate.

Meanwhile, Item 15 (Figs. 9 and 10), although complicated, has ideal discrimination, a low guessing rate, and a reasonable amount of information contribution, so it can be used to select high-level candidates.

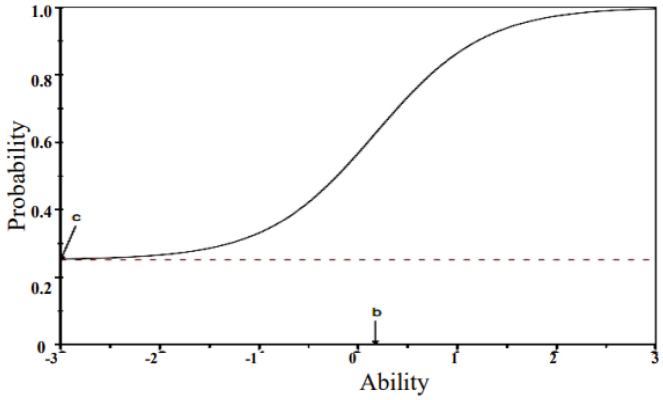

Fig. 7. Item Characteristic Curve: ITEM003 (Picture Credit: Original). (a = 1.068, b = 0.172, c = 0.251)

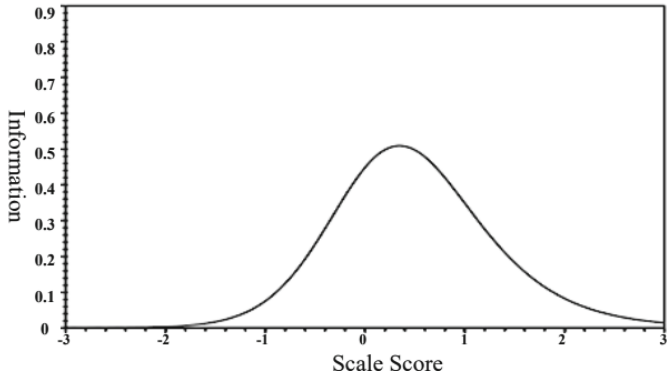

Fig. 8. Item Information Curve: ITEM003 (Picture Credit: Original)

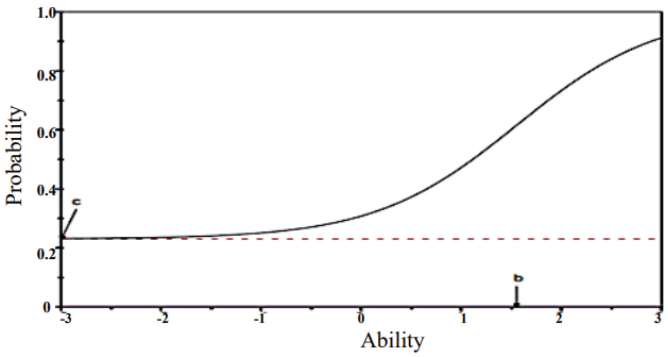

Fig. 9. Item Characteristic Curve: ITEM0015 (Picture Credit: Original). (a = 0.830, b = 1.552, c = 0.230)

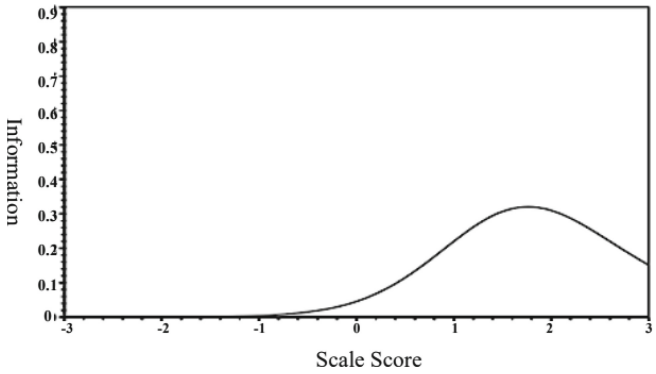

Fig. 10. Item Information Curve: ITEM0015 (Picture Credit: Original)

Item 15 asks about the speaker's views on rail systems in other countries, again which is an inference about the speaker's implied opinion based on the dialogue. In conjunction with Item 20, which will be mentioned below, both of these items have a high difficulty factor, indicating that inferring the speaker's implied opinion is a weakness to candidates. To exclude the other options, it is necessary to listen to the context. Railroads in other countries rely on nationalized support to survive, while rail services in other countries are not mentioned. In other words, it is very likely that the answer is wrong just by the keywords heard corresponding to the options, and the candidate must clearly understand the intention of the speaker's example. Simultaneously, since the item has a low guessing rate, which can clearly distinguish between high-level candidates (Figs. 11 and 12).

b) Non-standard items.

Item 22 (Figs. 13 and 14) is less complicated, less distinguishable, has a higher guessing rate, and contributes less information, so that it should be considered for deletion. Item 22 examines the candidate's ability to understand the main idea information.

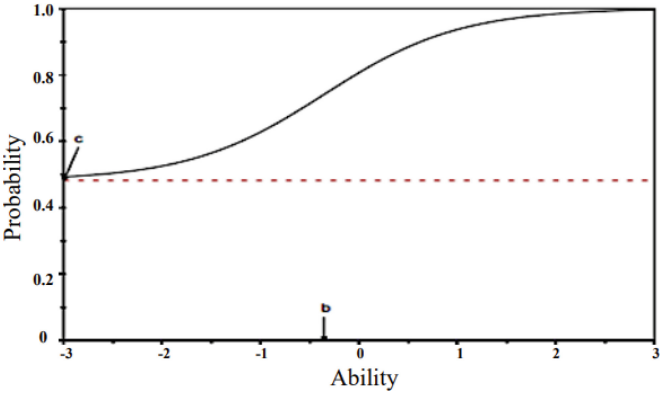

Fig. 11. Item Characteristic Curve: ITEM0022 (Picture Credit: Original). (a = 0.856, b = −0.356, c = 0.482)

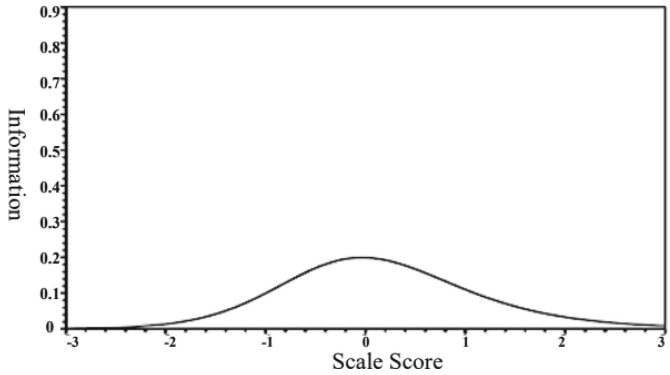

Fig. 12. Item Information Curve: ITEM0022 (Picture Credit: Original)

The candidate can eliminate the irrelevant distractions according to the keywords and directly correspond to the correct option. Even if candidates do not hear the keywords clearly, they can use listening strategies to guess the primary responsibility of the female reporter based on their understanding of the whole text.

In contrast to the Item 22 above, the item characteristic curves for Item 20 were as followed.

As shown in Figs. 15 and 16, Item 20 is more complex and moderately differentiated but has a high guessing rate and contributes less information, which is also considered for deletion. A possible reason for the item's relatively high guessing rate and low information contribution is that candidates can infer the difficulty of funding research with respect to their own research experiences. To some extent, it shows that the candidate's listening level does not fully determine the probability of answering the question correctly.

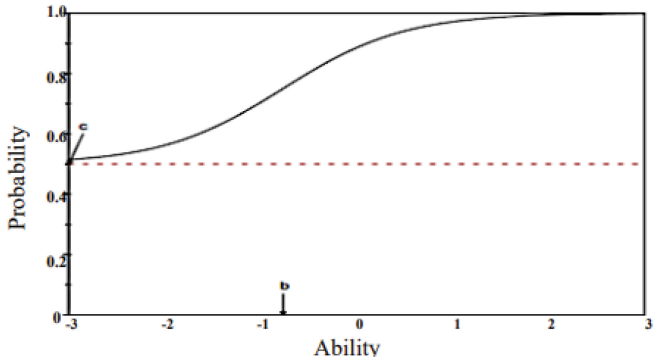

Fig. 13. Item Characteristic Curve: ITEM0022 (Picture Credit: Original). (a = 0.931, b = −0.755, c = 0.500)

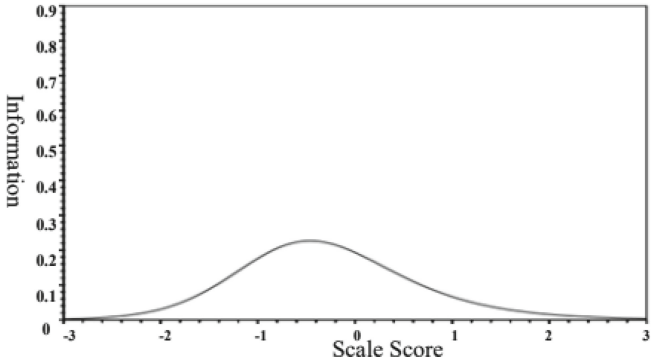

Fig. 14. Item Information Curve: ITEM0022 (Picture Credit: Original)

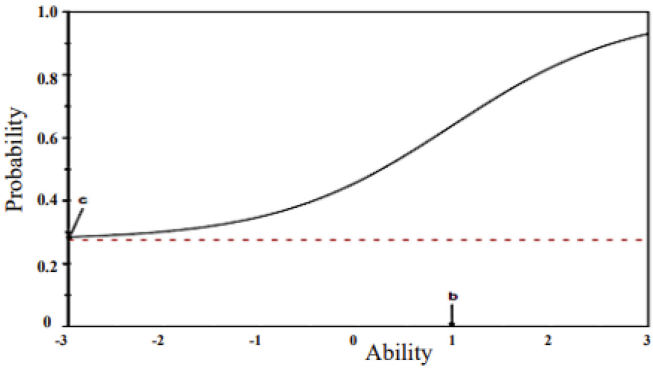

Fig. 15. Item Characteristic Curve: ITEM0020 (Picture Credit: Original). (a = 0.931, b = −0.755, c = 0.500)

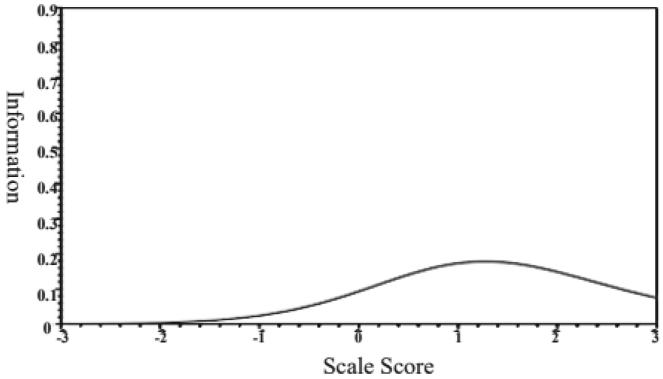

Fig. 16. Item Information Curve: ITEM002 (Picture Credit: Original)

5 Conclusions

This study evaluates the validity of the school-based college English listening test items according to the 3PL model. In general, this set of listening items were high in difficulty, ideal in differentiation, reasonable in information content, and with a certain degree of validity, which can accurately reflect the actual listening proficiency of the intermediate and high-level test takers. Regarding specific items, the candidates could generally understand the main idea of these questions, while inferring the speaker's opinion and intention was their weak points.

References

1. Messick, S. Validity [A]. In R. L. Linn (ed). Educational Measurement (3rd ed) [C]. New York: Macmillan, 1989: 13–103.
2. Mislevy, R. J., Steinberg, L. S. & Almond, R. G. On the structure of educational assessments [J]. Measurement: Interdisciplinary Research and Perspectives, 2003(1): 3–62.
3. Kane, M. T. An argument-based approach to validity [J]. Psychological Bulletin, 1992, 112(3): 527–535.
4. Kane, M. T. Validating the interpretations and uses of test scores [J]. Journal of Educational Measurement, 2013, 50(1): 1–73.
5. Bachman, L. & Palmer, A. Language Assessment in Practice [M]. Oxford: Oxford University Press, 2010.
6. Harrington, R.F. (1968). Field computation by moment methods. CR [1] Picture credit: Original