Iranian Journal of Applied Linguistics (IJAL)
Vol. 24, No. 2, September 2021, 135-161

A Three-Parameter Logistic IRT Calibration of the Items of the TEFL MA Admission Test as a High-Stakes Test in Iran

Fateme Nikmard

Islamic Azad University, Karaj, Iran

Kobra Tavassoli*

Islamic Azad University, Karaj, Iran

Abstract

To explore the characteristics of the items of the Teaching English as a Foreign Language (TEFL) MA Admission Test (henceforth TMAAT) as a high-stakes test in Iran, the current research utilized a three-parameter logistic Item Response Theory (IRT) calibration of the test items. The three-parameter logistic IRT model is the most comprehensive among the three models of IRT for it takes into account all the three effective parameters of item difficulty, item discrimination, and guessing simultaneously. The data were a random selection of 1000 TMAAT candidates taking the test in 2020 collected from Iran's National Organization of Educational Testing (NOET). The software used to analyze the data was jMetrik (Version 4.1.1), which is the newest version so far. As the results indicated, the TMAAT worked well in discriminating the higher and lower ability candidates and preventing the candidates from guessing the responses by chance, but it was not much acceptable regarding the difficulty level of the items as the items were far too difficult for the test-takers. The most important beneficiaries of the present investigation are test developers, testing experts, and policy-makers in Iran since they are responsible to improve the quality of the items in such a high-stakes test.

Keywords: Guessing, High-stakes test, Item difficulty, Item discrimination, Three-parameter IRT model

Corresponding author: Department of ELT, Faculty of Literature and Foreign Languages, Karaj Branch, Islamic Azad University, Karaj, Iran *Email:* kobra.tavassoli@kiau.ac.ir

1. Introduction

The TEFL MA admission test (TMAAT) is one of the high-stakes university admission tests in Iran, which is taken by thousands of students each year. As stated by Firoozi et al. (2015), each year, a considerable number of applicants compete with each other with the hope of being admitted to the TEFL MA programs in different universities. Considering the impact high-stakes tests have on the lives of many students, it seems necessary to think of a way to guarantee the quality of such tests, which are developed and administered each year. Lin (2020) has pinpointed the significance such tests by stating that "high-stakes testing is ubiquitous, and it brings important consequences, intended or unintended, for stakeholders, particularly students, and teachers (p. 159)." Accordingly, language testers have recognized the prominence of the effect high-stakes tests have on test takers' lives and their own responsibilities in improving all testing-related activities (Ahmadi & Mousavi, 2017). Language testers and assessors face many challenges, one of the most important of which is related to making decisions about both individuals and organizations or institutions (Bachman, 2014). To make appropriate decisions for the test-takers' admission into a program, the candidates' ability level, the academic requirements specific to any course of study that could affect the difficulty level of the items in the test, and the availability of the learning facilities for the candidates should be considered (Farley et al., 2020). Also, item selection, test design procedures, and reporting test scores are significant matters to be taken into account seriously.

To overcome such challenges, Bachman (2014) suggested that language testers consider all the knowledge and skills in the field of language testing and try to apply them to the immediate practical testing needs of the educational systems, from kindergarten to university. As he stated, education experts should also willingly collaborate with measurement experts since both groups have the noteworthy knowledge and skill about the important issues related to the validity of interpretations and the consequences of using high-stakes tests.

In spite of the importance of high-stakes tests in general and the TMAAT in particular, the evaluation of this test has long been overlooked in Iran (Ahmadi & Mousavi, 2017). Moreover, the challenges test developers and policy makers face in making correct decisions about the test-takers' lives (Bachman, 2014) through the TMAAT make it vital to investigate the quality of the items in such a high-stakes test (Chapelle, 2020). Having good quality items makes it possible to have a rich item bank for future tests as well.

A common way of checking the quality of the items in a high-stakes test and identifying the problematic ones is the use of the Item Response Theory (IRT) (Bachman & Palmer, 2010). IRT was first introduced as a response to the shortcomings of the Classical Test Theory (CTT), which was previously utilized to diagnose the items not contributing to the internal consistency or overall quality of a test. Such diagnosis is especially important in a high-stakes test, which determines the future lives of a large number of test-takers including many (un)intended consequences (Lin, 2020).

To overcome this gap in the Iranian context, the present research aimed at evaluating the quality of the items included in the TMAAT (Year 2020) through the three-parameter Item Response Theory (IRT). For this purpose, we posed the following research question:

To what extent are the item characteristics of the TMAAT (i.e., item difficulty, item discrimination, and guessing) acceptable based on the three-parameter logistic IRT model?

2. Review of the Literature

To investigate the quality of the items in a test, especially in a high-stakes test, IRT models became popular. The different IRT models became widespread in the 1990s. The models were rooted in advanced psychometric theories to assess the latent attributes of each respondent such as their ability level or beliefs (di Vettimo, 2022). The models were developed based on strong theoretical foundations of item-free person measurement, sample-free item calibration, suitable items, and person identification (Ellis & Ross, 2014). MacDonald and Paunonen (2002) explained the theoretical benefits of the IRT over the CTT and mentioned that while item discrimination estimates based on the CTT are only accurate across some conditions, those of the IRT are accurate under almost all settings. The basic focus of IRT models is on the factors affecting the individuals' observed score on each item. That is, IRT models can make strong predictions about the test-takers' performance on individual items, their ability levels, and the characteristics of individual items (Ellis & Ross, 2014).

IRT models are capable of making the testing procedure more effective through modifying or adjusting tests while they are in progress (Bailey, 2020). In other words, it is possible to amend the items presented according to the individual's response to the preceding item. In addition, using IRT models, it is possible to decide whether an item has to be included in a set of items. Another benefit of IRT models is that they do not need presenting a

single number for each item (di Vettimo, 2022). That is, the data could be fed to the model in the form of positive responses and/or as a combination of neutral and positive responses.

The three main models of IRT are the one-parameter (or Rasch), two-parameter, and three-parameter logistic IRT models.

2.1. The One-Parameter Logistic (1-PL) IRT Model (or Rasch Model)

The one-parameter logistic (1-PL) IRT model (also called the Rasch model) estimates only the parameter of item difficulty (Barkaoui, 2014). It is the simplest among various types of IRT models. However, it takes into account a supplementary assumption that all the items have identical discrimination power since they are differentiated only by their difficulty level. The model is based on the *unidimensionality* assumption that for a measurement to be effective, it should examine only one attribute at a time (Brambor et al., 2020). A limitation of the model, according to MacDonald and Paunonen (2002), is that all the items are regarded as having the same and fixed discrimination power, and only the item difficulty is calculated.

Misfitting items, items not fitting the model (de Ayala, 2009), are those violating the 1-PL IRT model probabilistic expectations leading the respondents not to provide the correct answer. Possible reasons are different interpretations of the respondents and/or the high difficulty level of the items that cause the test-takers to guess the response (Ellis & Ross, 2014). IRT works based on the supposition that there is an unobserved (latent) influential element affecting the individuals' responses which is subsequently used as the basis for estimating the difficulty (or severity) level of the items (Bailey, 2020). The 1-PL IRT model could be utilized with items that are dichotomously scored (Szabó, 2008), are binary, or are ordinal (Brambor et al., 2020).

2.2. The Two-Parameter Logistic (2-PL) IRT Model

This model adds another item parameter, that is, item discrimination to the difficulty level of the items and evaluates both (MacDonald & Paunonen, 2002; Warnby et al., 2023). It measures item quality by the extent to which the test-takers possessing similar ability levels

have answered them consistently, which is called the discrimination power (Bailey, 2020). Accordingly, a low discrimination power is a sign of the existence of intruding factors other than the target latent variables indicating that the item is not well-constructed. That is, the discrimination parameter makes it clear how much an item is accurate in discriminating the respondents (Metsämuuronen, 2022). In fact, discrimination is the ability of each individual item to distinguish between individuals with different ability levels (MacDonald & Paunonen, 2002). A more difficult item probably receives a correct answer from the more proficient test-takers having more knowledge of the subject under investigation.

The 2-PL IRT model also works with dichotomous, binary, or ordinal scores (Brambor et al., 2020; Szabó, 2008). Moreover, it is the most frequently used IRT model for high-stakes multiple-choice items, especially those aiming to filter out the items not working properly before operating the test (Ellis & Ross, 2014). As Runzrat et al. (2019) claimed, it is a technique drawn from computer-aided assessment, aiming at assessing the test-takers' ability and forecasting the upshots of the responses they provide.

2.3. The Three-Parameter Logistic (3-PL) IRT Model

The 3-PL IRT model is known as the most complicated of the three IRT models (MacDonald & Paunonen, 2002), which proposes that the more difficult an item is, the more likely the candidates try to guess the response by chance. The 3-PL IRT model takes into account the guessing parameter (Ellis & Ross, 2014; Metsämuuronen, 2022) in addition to the two parameters of difficulty level and discrimination power. The 3-PL IRT model evaluates the chance a person, even with a very low ability level, has to provide the correct answer to a test item through only guessing. Adding this third parameter of guessing to the model makes it much more complicated than the other IRT models (Warnby et al. 2023).

Taking the guessing factor into account in the 3-PL IRT model's analysis seems essential if respondents can answer the item by mere guessing, which is a consistent problem in multiple-choice items (Majoros, 2022). Here too, the items could be scored dichotomously (MacDonald & Paunonen, 2002), could be binary or ordinal (Brambor et al., 2020).

Following either of the IRT models, to rank the candidates correctly, the items that are malfunctioning should be eliminated from the test (Ellis & Ross, 2014). However, if various IRT models show different malfunctioning items, a possibility is to choose the model (i.e.,

one-parameter, two-parameter, or three-parameter) that fits the data in the best way with the lowest number of item deletions (Ellis & Ross, 2014). Other alternatives to item removal are that of item rewording (de Ayala, 2009), providing answer keys with more than one correct answer, and providing a confirmation on the accuracy of the scoring key (Ellis & Ross, 2014).

2.4. TEFL MA Admission Test (TMAAT)

Developed by the Center of Educational Measurement (CEM), the Iranian TEFL MA Admission Test (TMAAT) is a high-stakes test based on whose results important decisions are made for many test-takers (Farhady & Hedayati, 2009). The test consists of general and specialized sections. The general section tests the test-takers' knowledge of structure, vocabulary, cloze passages, and reading comprehension whereas the specialized section evaluates their knowledge of linguistics, language teaching, and language testing. The total number of items in the test is 120, 60 for the general section and 60 for the specialized section. The scores candidates obtain in the two sections are added and counted as the total score assigned to each candidate. The total scores are then ranked and the test-takers are accepted at different universities in their related majors. In case the test taker's score is not high enough to enter their first requested university, their second requested university is considered. This goes on until the candidate's score matches the criterion in the requested university (Farhady & Hedayati, 2009).

Although there exist various investigations on different validity issues of high-stakes language tests of either national or international scope, not many studies focused on the university admission test for English-related majors, except to investigate their fairness, washback, or validity (Razavipur, 2014). Accordingly, examining the quality of the items of the TMAAT seemed necessary to the researchers in this study.

3. Method

3.1. Context of the Study

The data of the current inquiry were a random sample of the scores of 1000 applicants (both female and male) for the TEFL MA admission test (TMAAT) in Iran in 2020. The data were provided by Iran's National Organization of Educational Testing (NOET) on a random basis, and no specific demographic information of the 1000 applicants was provided for the confidentiality of the information and the test security reasons.

The TMAAT is a centralized admission test administered once a year to MA applicants of TEFL in Iran. The test consists of two sections. The first section, which includes 60 items,

measures the candidates' general knowledge of English including the *structures*, *vocabulary*, and *reading comprehension*. The second section, which also includes 60 items, measures the candidates' specialized knowledge of the three main domains of *linguistics*, *language teaching*, and *language testing*.

Since the first section is common to all candidates taking part in the MA admission test of English including Teaching English as a Foreign Language (TEFL), English Literature, and English Translation, for the purpose of this research, only the candidates' performance on the second section checking their specialized knowledge of TEFL courses was investigated. The reliability of the test scores was estimated through the person separation method, which is a common reliability estimation technique in IRT. The person separation method reliability values were 0.96 for linguistics, 0.94 for language teaching, and 0.96 for language testing domains. The results showed that the scores on all the three domains of the TMAAT were highly reliable.

3.2. Procedure

The scores of a randomly selected sample of 1000 TMAAT test-takers taking the test in Iran in 2020 were utilized in this research. The data were received from NOET following rigorous administrative procedures, moving back and forth from Islamic Azad University, Karaj Branch, to NOET. The test-takers' responses to 60 multiple-choice items in the three domains of linguistics, language teaching, and language testing were collected. For each of the 60 items, the test-takers had chosen the best answer from the four available options in multiple-choice items. To decide about the item characteristics and the fair decisions made on the basis of the test scores, the researchers ran three different 3-PL IRT models through jMetrik (Version 4.1.1) on the domains of linguistics, language teaching, and language testing, each including 20 multiple-choice items.

The researchers received a written consent from NOET to receive the TMAAT test scores and to run the required analyses on the scores. Further, to keep the test-takers' information confidential, no information about the candidates was provided to the researchers by NOET.

3.4 Data Analysis

To start analyzing the data, two primary assumptions of IRT, the *unidimensionality* and the *local item independence*, were checked. After ensuring these two assumptions, all the three IRT models were run on the three domains of linguistics, language teaching, and language testing to find out which model fits the data better. Comparing the three models, the 3-PL IRT model was chosen as it provided the best results. All the details are provided in the results section.

4. Results

To start the IRT analysis, two primary assumptions should be checked to ensure the appropriateness of the data sets for IRT analysis. The first prerequisite is the *unidimensionality* assumption. Unidimensionality exists as long as all the items of a test measure a single latent trait or ability (Ellis & Ross, 2014). Factor analysis is a proper way of checking the unidimensionality of a test. The second prerequisite is the *local item independence* assumption (Brambor et al., 2020; de Ayala, 2009; Ellis & Ross, 2014), which states item responses should not depend on each other, that is, the answers provided for any single item should not be related to the responses given to any other item within a test section. de Ayala (2009) also stated that the violation of this second assumption could lead to an overestimation of the total test information.

The first IRT assumption was met in the current research based on the results of the confirmatory factor analysis (CFA) on the TMAAT scores of 1000 candidates (See Appendix, Tables 1-7 and Figure 1). Both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) can be utilized to check the construct validity of the items in a test (Moradi et al., 2000). EFA is also common to explore the possible underlying factor structure of a set of observed variables (Pallant, 2020) while CFA is considered a logical way of checking the relevance of the hypothesized traits underlying the observed behavior of the respondents as identified previously (Szabó, 2008). In the case of the TMAAT, CFA was used since the aim was to check the TMAAT constructs (i.e., linguistics, language teaching, and language testing) as identified by NOET (Phakiti, 2014; Plakans, 2014). It should be noted that the items in each domain of the test were subject to a separate 3-PL IRT analysis, which means three IRT analyses were run on the three separate domains of the test each one consisting of 20 items. The second assumption was also met since the answers to the items in each domain of linguistics, language teaching, and language testing did not depend on each

other and each item questioned different issues. Therefore, running IRT was considered legitimate.

Moreover, from among the three models of IRT (i.e., one-parameter, two-parameter, and three-parameter), the 3-PL IRT model was utilized since compared with the other two models, it showed a better fit to the data, and it provided information on all the three parameters of item difficulty, item discrimination, and guessing. Each 3-PL IRT model was run through the jMetrik software (Version 4.1.1).

Before reporting the IRT results, the mean scores of the items were inspected since the higher a mean score is, the easier the item has been, and vice versa, the lower a mean score is, the more difficult the item has been (Aryadoust, 2021a). Tables 1 to 3 report the mean scores of the items included in each of the three domains of TMAAT.

Table 1. *Mean Scores of Linguistics Items*

Items	N	Mean	SD
61	1000	.20	.40
62	1000	.21	.41
63	1000	.07	.26
64	1000	.18	.38
65	1000	.08	.27
66	1000	.16	.37
67	1000	.14	.34
68	1000	.16	.37
69	1000	.10	.30
70	1000	.09	.29
71	1000	.04	.21
72	1000	.12	.33

Table 2. *Mean Scores of Language Teaching Items*

Items	N	Mean	SD
81	1000	.11	.31
82	1000	.10	.30
83	1000	.04	.21
84	1000	.07	.25
85	1000	.04	.19
86	1000	.18	.38
87	1000	.05	.23
88	1000	.14	.34
89	1000	.07	.26
90	1000	.09	.28
91	1000	.05	.23
92	1000	.07	.26

Table 3. *Mean Scores of Language Testing Items*

Items	N	Mean	SD
101	1000	.16	.37
102	1000	.03	.19
103	1000	.03	.18
104	1000	.04	.20
105	1000	.06	.24
106	1000	.11	.31
107	1000	.08	.27
108	1000	.06	.25
109	1000	.14	.34
110	1000	.08	.28
111	1000	.11	.32
112	1000	.18	.38

73	1000	.18	.38	93	1000	.03	.18	113	1000	.14	.35
74	1000	.08	.28	94	1000	.05	.23	114	1000	.11	.32
75	1000	.03	.18	95	1000	.06	.24	115	1000	.11	.32
76	1000	.12	.33	96	1000	.10	.30	116	1000	.15	.36
77	1000	.09	.29	97	1000	.10	.31	117	1000	.06	.24
78	1000	.06	.23	98	1000	.06	.23	118	1000	.08	.27
79	1000	.13	.34	99	1000	.12	.33	119	1000	.09	.29
80	1000	.14	.35	100	1000	.16	.37	120	1000	.18	.39

The mean scores reported in Tables 1, 2, and 3 make it clear that all the 60 items in the three domains of linguistics, language teaching, and language testing of TMAAT are somehow difficult as their mean scores are .21 at best, which is not a large value indicating the difficulty of the items in the test. Nevertheless, since the mean score is not an ultimate way of calculating the difficulty level of the items in a test, the difficulty level along with the discrimination power and the guessing parameter of the items were estimated through three 3-PL IRT analyses and the results are reported in Tables 4 to 9 below.

Before going into the analyses, we should mention the acceptable criteria for the three parameters of item discrimination, item difficulty, and guessing. de Ayala (2009) expressed that the values for the *discrimination power* of the items (known as Apar) can range from –∞ to +∞. Higher values demonstrate items that effectively discriminate test-takers holding different levels of ability assessed by the item (Aryadoust, 2021b; MacDonald & Paunonen, 2002). On the other hand, negative values show items that the lower ability test-takers have a higher probability of answering correctly, which should be discarded from the test since they behave in a counterintuitive fashion and are not consistent with the model (de Ayala, 2009). Szabó (2008, p. 32) provided the following ranges for item discrimination values:

If item discrimination ≥ .40, the item is functioning quite satisfactorily;

If .30 ≤ item discrimination ≤ .39, little or no revision is required;

If .20 ≤ item discrimination ≤ .29, the item is marginal and needs revision;

If item discrimination ≤ .19, the item should be eliminated or completely revised.

The ideal value for the *item difficulty* parameter (known as Bpar) is .50. Ellis and Ross (2014) further explained that at this difficulty level, the discrimination power of the item

would be at the maximum level. Therefore, in a well-developed test, almost all difficulty values are expected to be around .50. The higher difficulty values signal more difficult items, while the lower values show simpler ones (Aryadoust, 2021b; MacDonald & Paunonen, 2002). However, as Ellis and Ross (2014) stated, only items having difficulty values larger than +/− 2.95 are malfunctioning and should be removed from the test; otherwise, they could be revised.

Finally, the *guessing* parameter (known as Cpar) provides information about the extent to which test-takers have the chance of successfully guessing the correct answer even without knowing the answer (Szabó, 2008). The preferable values for the guessing parameter fall between .00 to .40 (Aryadoust, 2021b; Ellis & Ross, 2014). Higher guessing values indicate that the item can be answered correctly by chance and deviates from the expectations of the test.

Putting the information from the three parameters together, it is concluded that malfunctioning items that should be removed from the test are those with a discrimination parameter (Apar) value less than .30, a difficulty parameter (Bpar) value larger than +/− 2.95, and a guessing parameter (Cpar) value surpassing .40.

The first 20-item domain of the TMAAT is *linguistics*, which was subject to a 3-PL IRT model and the results are reported in Tables 4 and 5.

Table 4. *Marginal Maximum Likelihood Estimation (MMLE) Item Parameter Estimates of Linguistics*

Item	Apar	SE	Bpar	SE	Cpar	SE
61	1.63	.17	1.34	.08	.03	.01
62	1.47	.14	1.34	.09	.03	.01
63	1.37	.22	2.54	.24	.02	.01
64	1.49	.19	1.61	.11	.04	.02
65	1.36	.17	2.37	.19	.01	.01
66	1.12	.14	1.97	.17	.03	.01
67	2.03	.22	1.57	.08	.02	.01

68	1.58	.17	1.58	.10	.02	.01
69	1.53	.19	2.05	.14	.02	.01
70	1.55	2.09	2.05	.14	.02	.01
71	1.91	.35	2.58	.22	.02	.01
72	1.91	.19	1.63	.09	.01	.01
73	2.12	.20	1.31	.06	.02	.01
74	2.32	.23	1.79	.08	.01	.00
75	1.08	.25	3.85	.66	.01	.01
76	2.58	.19	1.48	.06	.01	.01
77	2.12	.22	1.78	.09	.01	.01
78	1.35	.21	2.74	.27	.01	.01
79	2.22	.19	1.47	.07	.01	.00
80	2.45	.19	1.38	.06	.01	.00

Checking the Apar values (showing item discrimination) for items 61 to 80, all of which assess the *linguistics* ability of the candidates, it becomes clear that all have values higher than .30, and therefore, larger than the point at which they have to be discarded from the test. Consequently, the discrimination parameter values could be regarded as adequate and capable of discriminating candidates possessing different levels of linguistics knowledge.

Concerning the Bpar values (showing the difficulty level of items), it becomes clear that although they are all within the range of +/− 2.95 specified for the acceptable difficulty value, they are very far from .50 which represents the optimal difficulty level. Therefore, the difficulty values could not be regarded as optimal. Moreover, since the Bpar values are large, it can be concluded that the items are more difficult than expected for a well-designed test. The difficulty level of the items reconfirmed the information previously demonstrated in Table 1.

Finally, the Cpar values (estimating the guessing parameter) are completely satisfactory and within the specified range of .00 to .40. That is, because the Cpar values are all between

.01 and .03, they all show that there is not a chance for the test-takers to correctly guess the answers without knowing the correct responses.

Table 5. *Item Fit Statistics of Linguistics*

Item	S-X2	df	*p*-value
61	9.4922	13.0000	.7348
62	21.7458	12.0000	.0405
63	18.8813	12.0000	.0914
64	20.4601	12.0000	.0589
65	27.2458	12.0000	.0071
66	15.9728	12.0000	.1925
67	7.6678	12.0000	.8105
68	16.1340	13.0000	.2420
69	25.4999	13.0000	.0198
70	13.9431	12.0000	.3044
71	9.0093	12.0000	.7021
72	23.9603	12.0000	.0206
73	14.1053	13.0000	.3665
74	17.0837	12.0000	.1465
75	19.9657	12.0000	.0677
76	10.6748	12.0000	.5570
77	8.9003	12.0000	.7114
78	16.5032	12.0000	.1693
79	16.2733	12.0000	.1790
80	16.1097	12.0000	.1863

Taking a close look at the *p*-values reported in Table 5 for the item fit statistics of the *linguistics* domain of the TMAAT, it becomes clear that only four items (62, 65, 69, and 72) do not fit the 3-PL IRT model.

To summarize the results obtained for the *linguistics* domain, we should mention that although the items were good enough to discriminate the higher and lower ability candidates and not to let the test-takers guess the answers by chance, they were not appropriate regarding the difficulty level of the items and they need some modifications. In other words, the items were more difficult than they should be, but they were good in discriminating candidates possessing different ability levels and in preventing the test-takers to guess the correct answers by chance. Therefore, they could be considered partially fair to the test-takers.

The next 3-PL IRT analysis was run on the 20 items included in the *language teaching* domain of the TMAAT whose outcomes are presented in Tables 6 and 7 below.

Table 6. *Marginal Maximum Likelihood Estimation (MMLE) Item Parameter Estimates of Language Teaching*

Item	Apar	SE	Bpar	SE	Cpar	SE
81	1.28	.18	2.25	.19	.02	.01
82	.98	.16	2.77	.32	.02	.01
83	2.05	.27	2.30	.15	.01	.00
84	1.81	.23	2.17	.14	.01	.01
85	1.12	.28	3.64	.62	.01	.01
86	1.89	.18	1.34	.07	.02	.01
87	1.11	.20	3.16	.41	.01	.01
88	2.01	.21	1.57	.08	.02	.01
89	1.69	.21	2.20	.15	.01	.01
90	1.78	.23	2.01	.13	.02	.01
91	2.31	.25	2.04	.10	.01	.00
92	1.37	.23	2.54	.24	.02	.01

93	1.88	.29	2.57	.21	.01	.00
94	1.88	.30	2.37	.18	.02	.01
95	1.57	.21	2.40	.19	.01	.01
96	1.77	.21	1.90	.11	.02	.01
97	2.24	.25	1.74	.08	.02	.01
98	1.53	.21	2.50	.21	.01	.01
99	1.97	.20	1.65	.08	.02	.01
100	1.43	.16	1.68	.11	.02	.01

The discrimination power values (i.e., Apar) of the items in the *language teaching* domain of the TMAAT (items 81-100), reported in Table 6, are also adequate as they are all above .30.

Regarding the difficulty values (i.e., Bpar) of the *language teaching* items, it can be seen that they all fall within the acceptable range of +/− 2.95. They are, however, far from .50 at which optimal difficulty can be guaranteed. As a result, they are more difficult than being considered good items in a well-developed test. This also reconfirms the information in Table 2 regarding the difficulty level of the items.

Concerning the guessing parameter (i.e., Cpar), however, all the values perfectly fall within the acceptable range of .00 to.40, which is an indication of not letting the respondents guess the correct answer without knowing it. All the Cpar values are between .01 and .02 and much lower than .40.

Table 7. *Item Fit Statistics of Language Teaching*

Item	**S-X2**	**df**	***p*-value**
81	12.1923	9.0000	.2027
82	22.3961	9.0000	.0077
83	17.2411	9.0000	.0451
84	6.5863	9.0000	.6801

85	9.8160	9.0000	.3656
86	13.4459	9.0000	.1434
87	20.7055	9.0000	.0140
88	11.7533	9.0000	.2276
89	8.8104	9.0000	.4550
90	7.5359	9.0000	.5815
91	12.1459	9.0000	.2052
92	5.3304	9.0000	.8046
93	11.8452	9.0000	.2222
94	6.2372	9.0000	.7160
95	9.7593	9.0000	.3703
96	9.7470	9.0000	.3713
97	12.6447	9.0000	.1794
98	11.9819	9.0000	.2143
99	6.3557	9.0000	.7039
100	12.0265	9.0000	.2118

As reported in Table 7, except for three items (82, 83, and 87), all the other *p*-values are indications of the good fit of the 3-PL IRT model for the analysis of the *language teaching* items as they are all higher than the .05 level of significance.

To sum up, the items in the *language teaching* domain of the TMAAT had good discrimination power being able to discriminate among the test-takers with different ability levels. However, the difficulty levels of the items were not satisfactory. That is, the items were more difficult than they should be in a high-stakes test. The guessing values were also at the acceptable range meaning that it was not possible for the test-takers to find the correct answers through mere guessing. Hence, it could be said that the items in this domain of the test were also partially fair to the test-takers.

A third 3-PL IRT analysis was run on the 20 items in the *language testing* domain of the TMAAT, and the results are demonstrated in Tables 8 and 9.

Table 8. *Marginal Maximum Likelihood Estimation (MMLE) Item Parameter Estimates of Language Testing*

Item	Apar	SE	Bpar	SE	Cpar	SE
101	2.53	.18	1.30	.05	.01	.01
102	1.84	.28	2.60	.21	.01	.00
103	1.49	.31	3.07	.36	.01	.01
104	1.19	.21	3.30	.42	.01	.01
105	1.52	.21	2.47	.20	.01	.01
106	2.29	.22	1.64	.07	.01	.01
107	1.72	.20	2.08	.13	.01	.01
108	1.55	.20	2.39	.18	.01	.00
109	2.46	.20	1.44	.06	.01	.01
110	2.62	.19	1.72	.06	.00	.00
111	1.76	.17	1.77	.10	.01	.00
112	2.46	.18	1.23	.05	.01	.01
113	2.13	.21	1.53	.07	.02	.01
114	1.34	.19	2.18	.17	.03	.01
115	1.48	.16	1.94	.13	.01	.01
116	2.68	.17	1.34	.05	.01	.00
117	1.65	.21	2.36	.17	.01	.01
118	1.23	.17	2.61	.24	.01	.01

119	1.47	.16	2.14	.15	.01	.01
120	2.23	.20	1.27	.06	.02	.01

In the case of the item discrimination values (i.e., Apar) for items 101 to 120 assessing the *language testing* ability of the test-takers, all items have values higher than .30, and are higher than the point at which they should be removed from the test. Consequently, the discrimination parameter values could be considered satisfactory and capable of discriminating candidates having different levels of language testing knowledge.

Concerning the difficulty level of the items (i.e., Bpar), except for items 103 and 104, all the difficulty values are within the acceptable range of +/− 2.95 specified for difficulty values. They are, however, not close to .50, the level at which the optimal difficulty value could be guaranteed. Therefore, it can be said that the difficulty values are not good. Moreover, since the Bpar values are very large, the items are considered more difficult than expected for a well-structured test. This also reconfirms the information in Table 3 regarding the difficulty level of the items. Moreover, in the case of items 103 and 104, they are considered very difficult and suggested to be discarded from the test.

All the values for the guessing parameter (i.e., Cpar), fall within the acceptable range of .00 to .40, meaning that there was not a chance for the test-takers to correctly guess the answer without knowing it. The Cpar values are all between .01 and .03, which are much lower than the specified range.

Table 9. *Item Fit Statistics of Language Testing*

Item	S-X2	df	*p*-value
101	11.1399	12.0000	.5170
102	8.7094	12.0000	.7275
103	25.6459	12.0000	.0120
104	11.1571	12.0000	.5155
105	12.8114	12.0000	.3829
106	11.2206	12.0000	.51.1

107	13.6587	12.0000	.3230
108	8.1582	12.0000	.7726
109	15.9610	12.0000	.1930
110	15.0492	11.0000	.1803
111	17.6395	12.0000	.1271
112	10.6855	12.0000	.5560
113	15.4644	12.0000	.2170
114	16.6493	12.0000	.1633
115	15.5750	12.0000	.2115
116	9.1842	12.0000	.6871
117	10.8715	12.0000	.5400
118	29.5764	12.0000	.0032
119	26.5079	12.0000	.0091
120	17.8233	12.0000	.1212

Finally, Table 9 shows that from among the *p*-values for the item fit statistics of the *language testing* domain of the TMAAT, only three items (103, 118, and 119) do not fit the 3-PL IRT model, which was the most suitable among the three IRT models for this study.

To wrap up the outcomes related to the *language testing* domain of the TMAAT, they all had good discrimination power values, while they were not satisfactory regarding their difficulty level. That is, the items were more difficult than they were supposed to be for the intended test-takers. The items in this section of the test had also acceptable guessing values not letting the test-takers guess the answer without knowing it. Therefore, the language testing items could also be claimed to be partially fair toward the test-takers.

To conclude whether the TMAAT items are appropriate and fair to the test-takers, we should mention that although the test items were more difficult than they should be, the test worked well regarding the item discrimination parameter by differentiating between the candidates possessing different ability levels in the three domains of linguistics, language teaching, and language testing. In addition, the test items had acceptable guessing parameter

values and did not let the test-takers guess the correct answer without knowing it. Therefore, the test is considered to be partially fair toward the test-takers. In other words, it is recommended to make the test easier so that it becomes more appropriate for the intended test-takers.

5. Discussion

To analyze the TMAAT and evaluate the items included in the test, three 3-PL IRT models were run on 1000 test-takers' scores on the three domains of linguistics, language teaching, and language testing. The results indicated that although the test did not provide good values in the case of the difficulty level of items, it worked well in the case of the discrimination power being capable of differentiating higher and lower ability level candidates, and the guessing parameter by not letting the test-takers guess the correct answers by chance. Therefore, it is recommended to make the test items easier for the test-takers by including less difficult items for the candidates to answer. This would increase the quality of the test and make the test more appropriate for the test-takers. The results of this study can be compared with similar research on high-stakes tests around the world.

Hilton (2021) delved into the test-takers' experiences of the national high-stakes test of Literacy and Numeracy Test for Initial Teacher Education students (LANTITE) and concluded that to have a higher-quality test, during the testing procedure, the test structure should be taken into consideration, an issue related to the guessing factor. In other words, a test has to be structured in a way that the test-takers could not guess the answers just by chance. The results are similar to the findings in this study since both studies considered mere guessing an undesirable feature of a high-stakes multiple-choice test.

Higher education admission through an entrance exam was the subject of Smirnios' (2022) research, in which the researcher sought English students' viewpoints about the exam, its processes, and the skills tested. Employing both quantitative and qualitative research methods, the findings demonstrated that although the university students entering the program through the admission test considered it a difficult test, they believed that the test is a good way of preparing them for the university courses and the different needed skills. That is, they stated that the admission test is a functional test and based on its results, it is possible to predict how a candidate will perform throughout the university years. The outcomes obtained are in agreement with those of the current study concerning the difficulty level of the TMAAT high-stakes test. In spite of their difficulty, high-stakes tests are perceived as

relevant procedures and based on their results, more capable university students could be chosen out of all the candidates taking part in the admission test. This also indicates the importance of the discrimination power of the high-stakes tests to differentiate test-takers with diverse ability levels.

Regarding the accessibility and security of high-stakes tests, an inquiry was also done by Coniam et al. (2021) who used a high-stakes test of the English language in an online mode to investigate the test-takers' attitudes toward the test. At the end of the study, they found that although delivering such a test online had its own challenges, it was beneficial to the test-takers in terms of accessibility and security matters, an issue less explored by researchers so far. It is apparent that to ensure the quality of an important test such as TMAAT, only focusing on item characteristics is not sufficient and other prominent and effective issues regarding the quality of a high-stakes test such as accessibility and security should also be investigated.

The major point the findings of the present research as well as some other inquiries' findings on high-stakes tests could add to the related literature is that administering high-stakes tests is necessary to make program-level decisions about the candidates' admission to different programs and universities. However, one of the most important matters such tests should take into account is that they should include items with acceptable discrimination powers, difficulty levels, and guessing to bear higher quality and to be considered more standardized.

6. Conclusion

To investigate the quality of the items in the TMAAT in Iran, we used three different 3-PL IRT calibrations of the items of this high-stakes test. The three 3-PL IRTs ran on the three domains of linguistics, language teaching, and language testing in the TMAAT revealed that the test did a good job regarding the discrimination power and guessing, whereas the parameter of the difficulty level of the items was not much satisfactory. Thus, the TMAAT seems not to have a perfect quality. A fundamental issue of high-stakes tests is that such tests could affect the lives of millions of test-takers whose future lives and careers depend on the outcomes of such tests. Therefore, developing a high-quality test, which ensures that correct decisions will be made based on their outcomes, would help many students' wishes come true. Moreover, since the current research implemented a new software (i.e., jMetrik, Version 4.1.1) to analyze the test items, it could be regarded as an innovation to be considered more seriously by other scholars. Compared to other software such as Winsteps, BILOG, etc., the

jMetrik software provide similar outputs in a more cohesive manner through more understandable tables. Furthermore, the findings of this research may make policymakers and authorities more responsible toward high-stakes tests and make them ponder more about the quality of the items while they design and develop admission tests since they have to weigh the applicants' attributes and the quality of the tests as influential issues. Moreover, a vital issue for test developers and measurement specialists, especially regarding high-stakes tests, is to be able to prepare item banks for subsequent uses in future tests. As the outcomes of the present inquiry indicated, IRT is a valuable type of analysis for test design, item selection, and equating and reporting test scores. Therefore, its use could be considered influential in developing and evaluating high-stakes tests.

One of the major limitations of the present inquiry was that the researchers could just receive 1000 anonymous TMAAT candidates' answer sheets without having access to their demographic information. That is, only the candidates' responses to the TMAAT items (Year 2020) in the three domains of linguistics, language teaching, and language testing were available, not the candidates' demographic information such as their gender, age, etc. If such demographic information were available, we could have run Differential Item Functioning (DIF) as a further confirmation of the outcomes obtained through IRT analyses and could identify whether there were any biased items in the test or not. Then, the issue of test fairness could also be investigated. Another limitation of the study was that the analyses were run on only one administration of the TMAAT (Year 2020) since the researchers could not have access to other administrations of the test as part of the policy of the NOET. Thus, the researchers are cautious to generalize the findings of the study.

Based on the outcomes of this research, we recommend the TMAAT test developers to be more cautious about the difficulty level of the designed items and to develop test items that are easier for test-takers. The TMAAT authorities as another beneficiary group are recommended to invite professional test developers to work collaboratively and design more appropriate TMAATs. Furthermore, policymakers should know that in high-stakes tests where making the right decisions about the test-takers are extremely important, serious attempts should be made to have more appropriate items. The authorities and policymakers should also publicize the results so that researchers can do more investigations on such high-stakes tests and improve their quality.

Eventually, we recommend interested researchers to replicate the inquiry with different groups of test-takers participating in other high-stakes tests at various levels and fields to see whether they find the same or different results. Another recommendation for future

researchers is to utilize other software packages such as the SPSS, Winsteps, R Packages, etc. to analyze their data to compare and contrast the obtained results. Lastly, prospective researchers can use other ways of analyzing the quality of high-stakes tests, like using the Generalizability Theory (GT), along with IRT, in a single study to compare and contrast the results to improve the situation.

References

Ahmadi, A., & Mousavi, S. A. (2017). Aspects of the impact of language tests on students' lifeworld: An analysis of the Iranian B.A. university entrance exam based on Habermas's Social Theory. *Journal of Modern Research in English Language Studies, 4*(2), 31-51.

Aryadoust, V. (2021a, January 29). Rasch measurement using user-friendly jMetrik [Video]. YouTube. https://www.youtube.com/watch?v=DeM3iINcftY.

Aryadoust, V. (2021b, February 8). Item response theory made easy with user-friendly jMetrik [Video]. YouTube. https://www.youtube.com/watch?v=86nI_hfmvHY.

Bachman, L. F. (2014). Ongoing challenges in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment: Approaches and development, Volume III* (pp. 1586-1604). John Wiley & Sons, Inc.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Bailey, N. (2020). Measuring poverty efficiently using adaptive deprivation scales. *Social Indicators Research, 149*(3), 891-910. https://doi.org/10.1007/s11205-020-02283-1.

Barkaoui, K. (2014). Multifaceted Rasch analysis for test evaluation. In A. J. Kunnan (Ed.), *The companion to language assessment: Approaches and development, Volume III* (pp. 1299-1320). John Wiley & Sons, Inc.

Brambor, T., Goenaga, A., Lindvall, J., & Teorell, J. (2020). The lay of the land: Information capacity and the modern state. *Comparative Political Studies*, *53*(2), 175-213. https://doi.org/10.1177/0010414019843432.

Chapelle, C. A. (2020). An introduction to language testing's first virtual special issue: Investigating consequences of language test use. *Language Testing*, *37*(4), 638-645. https://doi.org/10.1177/0265532220928533.

Coniam, D., Lampropoulou, L., & Cheilari, A. (2021). Online proctoring of high-stakes English language examinations: A survey of past candidates' attitudes and perceptions. *English Language Teaching*, *14*(8), 58-72. https://doi.org/10.5539/elt.v14n8p58.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Publications, Inc.

di Vettimo, M. S. (2022). Measuring public support for European integration from population-level data using a Bayesian IRT model. *European Union Politics*, *23*(2), 171-191, https://doi.org/10.33774/apsa-2020-47ldr.

Ellis, D. P., & Ross, S. J. (2014). Item response theory in language testing. In A. J. Kunnan (Ed.), *The companion to language assessment: Approaches and development, Volume III* (pp. 1262-1279). John Wiley & Sons, Inc.

Farhady, H., & Hedayati, H. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics*, *29,* 132-141. https://doi.org/10.1017/S0267190509090114.

Farley, A., Yang, H. H., Min, L., & Ma, S. (2020). Comparison of Chinese and Western English language proficiency measures in transnational business degrees. *Language, Culture and Curriculum*, *33*(3), 319-334. http://dx.doi.org/10.1080/07908318.2019.1630423.

Firoozi, T., Rasooli, A., & Zandi, H. A. (2015). *Critique of MA TEFL national entrance exam using a reverse engineering approach.* The 13th International TELLSI Conference. Lorestan University, Iran.

Hilton, A. L. (2021). *The heart, the wallet and the cookie cutter: Student and stakeholder experiences of undertaking LANTITE, the high-stakes test in Australian initial teacher education.* PhD Dissertation. Murdoch University, Perth, Western Australia.

Karimnia, A., & Kay, E. (2015). An evaluation of the undergraduate TEFL program in Iran: A multi-case study. *International Journal of Instruction*, *8*(2), 83-98. http://dx.doi.org/10.12973/iji.2015.827a.

Lin, D. (2020). Book review: Teacher involvement in high-stakes language testing. *Language Testing*, *37*(1), 159-162. https://doi.org/10.1177/0265532220932481.

MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*(6), 921-943. https://doi.org/10.1177/0013164402238082.

Majoros, E. (2022). *Linking recent and older IEA studies on mathematics and science*. Acta Universitatis Gothoburgensis.

Metsämuuronen, J. (2022). Essentials of visual diagnosis of test items: Logical, illogical, and anomalous patterns in tests items to be detected. *Practical Assessment, Research, and Evaluation*, *27*(1), 1-31. https://doi.org/10.7275/n0kf-ah40.

Moradi, B., Tokar, D. M., Schaub, M., Jome, L. M., & Serna, G. S. (2000). Revisiting the structural validity of the gender role conflict scale. *Psychology of Men & Masculinity, 1*(1), 62. https://psycnet.apa.org/doi/10.1037/1524-9220.1.1.62.

Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS.* Routledge.

Phakiti, A. (2014). Questionnaire development and analysis. In A. J. Kunnan (Ed.), *The companion to language assessment: Approaches and development, Volume III* (pp. 1098-1115). John Wiley & Sons, Inc.

Plakans, L. (2014). Written discourse. In A. J. Kunnan (Ed.), *The companion to language assessment: Approaches and development, Volume III* (pp. 1098-1115). John Wiley & Sons, Inc.

Razavipur, K. (2014). On the substantive and predictive validity facets of the university entrance exam for English majors. *Journal of Research in Applied Linguistics (RALs), 5*(1), 77-90.

Runzrat, S., Harfield, A., & Charoensiriwath, S. (2019). *Applying item response theory in adaptive tutoring systems for Thai language learners.* In 2019 11th International Conference on Knowledge and Smart Technology (KST) (pp. 67-71). IEEE.

Smirnios, M. (2022). *Matriculation exam vs. entrance exam – first-year English students' experiences on university admission.* Master's Thesis. Helsingin yliopisto. http://hdl.handle.net/10138/344906.

Szabó, G. (2008). *Language testing and evaluation: Applying item response theory in language test item bank building.* Peter Lang.

Warnby, M., Malmström, H., & Hansen, K. Y. (2023). Linking scores from two written receptive English academic vocabulary tests – The VLT-Ac and the AVT. *Language Testing*, 1-28, https://doi.org/10.1177/02655322221145643.