Original Article

Dimensionality, discrimination power and difficulty of English test items: The case of a graduate exam for healthcare applicants

Seyyed Samad Sajjadi[1], Nematullah Shomoossi[2], Enayat Shabani[3], Abdurrashid Khazaei Feizabad[4*], Giti Karimkhanlooei[5]

[1]Department of Applied Linguistics, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[2]Department of Applied Linguistics, Sabzevar University of Medical Sciences, Sabzevar, Iran.
[3]Department of Applied Linguistics, Tehran University of Medical Sciences, Tehran, Iran.
[4]Department of Applied Linguistics, Zahedan University of Medical Sciences, Zahedan, Iran.
[5]Department of Applied Linguistics, School of Medicine, Zanjan University of Medical Sciences, Zanjan, Iran.

Article info

Article history:
Received 25 Feb. 2023
Accepted 20 Dec. 2024
Published 10 Sep. 2024

***Corresponding author:**
Abdurrashid Khazaei Feizabad, Department of Applied Linguistics, Zahedan University of Medical Sciences, Zahedan, Iran.
Email: arkhazaei@yahoo.com

How to cite this article:
Sajjadi SS, Shomoossi N, Shabani E, Khazaei Feizabad A, Karimkhanlooei G. Dimensionality, discrimination power and difficulty of English test items: The case of a graduate exam for healthcare applicants. J Med Edu Dev. 2024;17(55):108-119.

Abstract

Background & Objective: Administered by the Iranian Center for the Measurement of Medical Education, national university entrance exams are administered nationwide where English constitutes a vital section. This study aimed to assess dimensionality, discrimination power and difficulty of English test items in this graduate entrance exam.

Material & Methods: This quantitative study examined 160 English test items administered to 41633 test-takers applying for graduate studies in Iranian universities of medical sciences in 2021, and reported the characteristics of test takers during three successive years (2019, 2020, and 2021). NOHARM software (version 4.0) was used to analyze the data by examining dimensionality of the tests reporting a two-parameter model.

Results: Generally, female participants outnumbered the male, with a similar pattern among the admitted participants (70% females vs. 30% males). A positively significant correlation was found between participants' Grade Point Average and English test scores ($p < 0.05$). In 2021, the results of four administration sessions with a high reliability (i.e. 0.92, 0.88, 0.90 and 0.91) were analyzed separately. Two dimensionality parameters (i.e., difficulty & discrimination) fitted the model while the guessing parameter did not. English tests proved to be "difficult", with either "high" or "very high" discrimination power. Neither "easy" nor "very easy" items were found. No items were associated with "no" or "very low" discrimination power.

Conclusion: Overall, the tests functioned well; however, more research is required to rigorously evaluate the exams. Improvements concerning the social and long-term effects of these tests are suggested.

Keywords: English language, testing, national exam, graduate level, medical education

Introduction

A variety of English examinations are administered to screen out university applicants in healthcare majors across the world. For instance, the United States Medical Licensing Examination (USMLE) is regarded as one of the toughest exams in the world; it measures candidates' clinical abilities, medical knowledge, and English language ability (1). Also, the Medical College Admission Test (MCAT) is a multiple-choice examination for admission to medical schools in the USA (2). Another well-known test is the Occupational English Test (OET), which evaluates the language communication abilities of healthcare professionals seeking to register and practice in an English-speaking workplace (3). In India, the Foreign Pre-Medical Entrance Test is administered (4, 5). In Iran, similar tests are designed by the Center for the Measurement of Medical Education in order to assess the language abilities of applicants in the healthcare majors.

National university entrance exams, publicly known as Konkour in Iran, are administered at undergraduate, graduate, and postgraduate levels. While hundreds of thousands of high school graduates participate in the undergraduate nationwide exam (6), participants in the graduate and postgraduate exams just amount to tens of

thousands each year; in fact, all exams are quite competitive. The Center for the Measurement of Medical Education, subordinate to the Ministry of Health and Medical Education, administers the graduate and postgraduate level exams, which are vitally important for higher education applicants of healthcare and medicine.

Testing packages for each major (or set of similar majors) include a set of 40 English test items, together with tests of specialized courses. This Foreign Language Test consists of 20 vocabulary and 20 reading comprehension items (all multiple-choice), which must be completed in 40 minutes. The test is a norm-referenced test designed to assess applicants' ability in reading comprehension and vocabulary knowledge of academic English for healthcare and medical students; however, the assessment of the four skills is not in perspective. The participants are all educated in Farsi, Iran's official language. The items are normally developed with varying degrees of difficulty, and administered once a year in four sessions in two consecutive days, normally at a weekend. Furthermore, another 40-item English test is developed for the applicants of 'medical journalism', which is considered a more difficult test than the other four tests because a higher proficiency level is expected of its applicants who are mostly graduates of English and medicine. The tests are normally based on test takers' academic needs and test items reflect their undergraduate courses of English.

As far as it is known, English level requirements upon entering MSc. programs is not determined through a centralized test of English in other countries. In fact, even non-native-English speking countries do not administer an English test for medical MSc. applicants as we do in Iranian universities. For instance, in most universities in Indonesia, Brunei Darussalam, the Philippines, Malaysia, Singapore, and other ASEAN countries, they require International English Language Testing System (IELTS) Band score 6.5 or TOEFL score of 550 at entry point In Türkiye, international English language proficiency tests such as PTE academic, TOEFL iBT (Score of at least 70) and IELTS (minimum score 5.0) are required for master's programs; for PhD programs, the minimum score 6.0 is acceptable In general, Turkish universities demand a prerequisite entry language score on international language tests ranging between B1 and C1 level of proficiency according to the CEFR levels Therefore, this unique test of English requires specid consideration.

Due to its exclusive emphasis on a vocabulary and reading comprehension, this test is not regarded as a true test of English proficiency; rather, it is meant to measure applicants' language performance to some extent (7). Furthermore, owing to the restricted number of seats available in medical universities, they are highly competitive and serve two purposes: as a gatekeeper to weed out the less qualified students and as a guarantee of the admitted applicants' future academic abilities (8, 9). Nevertheless, despite its high-stakes nature and its evident impact on a significant number of test takers' future academic and professional prospects, to our knowledge, no reliable reports have been published on its effectiveness, reliability and validity; even thecnical reports are unavailable on the web. Therefore, the present study aimed to evaluate these characteristics and dimensionality of English test items in this nationwide medical graduate entrance exam during three successive years. In fact, it was carried out to evaluate these English exam items in light of statistical computational approaches in order to reflect a technical evaluation of these test items. The findings should aid in revising the construction and administration procedures.

Material & Methods

Design and setting(s)

This quantitative study was designed aiming to investigate dimensionality, discrimination power and difficulty of English test items in the graduate entrance exam for healthcare applicants in Iran. The test includes a set of 40 English test items, which are administered on the same day as tests of specialized courses. These test items consists of 20 vocabulary and 20 reading comprehension items, in the multiple-choice format, which must be completed in 40 minutes.

Participants and sampling

The study was carried out on English test items (n = 160) administered to 14,827 test takers applying for graduate studies at Iranian universities in the medical sciences in 2021, and test takers' characteristics were reported during three successive years (2019–2021). Normally, months prior to exam administration, the applicant enrolls for the exam and prepares for the exam. The applicants' gender and details of their registration, absenteeism, and admission details are reported below in Table 1.

Data collection method

The study data were obtained under confidentiality requirements from the Center for the Measurement of Medical Education directed by the Ministry of Health and Medical Education, but the test takers' personal

information (e.g., name or identity information) were not included. The obtained data included the test takers' performance on 160 English test items administered in 2021, together with their characteristics during three successive years (2019, 2020, and 2021). Using Excel and Word software from the Microsoft Office Package, we tabulated the data in different tables and organized them into different categories so that the analyses could be performed.

Table 1. Participants in the graduate entrance examination (2019-2021)

Candidates	Gender	2019		2020		2021	
		Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Registrants	Male	21584	26.64	18233	26.01	21364	26.53
	Female	59433	73.76	51861	73.99	59168	73.47
	Total	81017	100	70094	100	80532	100
Participants	Male	12619	26.95	12940	25.53	14718	26.1
	Female	34206	73.05	37739	74.47	41633	73.9
	Total	46825	100	50679	100	56381	100
Absentees	Male	8965	26.22	5293	27.26	6646	27.52
	Female	25277	73.78	14122	72.74	17505	72.48
	Total	34192	100	19415	100	24151	100
Allowed to choose a major	Male	11203	26.26	3522	26.58	8358	27.11
	Female	31458	73.74	9730	73.42	22473	72.89
	Total	42661	100	13252	100	30831	100
Admitted	Male	1298	30.27	1376	30.5	1856	30.79
	Female	2990	69.73	3136	69.5	4171	69.21
	Total	4288	100	4512	100	6027	100

Data analysis

Statistical analyses were conducted using descriptive and inferential statistics; additionally, fitting into a dimensionality model was examined using Noharm version 4.0. Correlations between test-takers' English test scores and Grade Point Average (GPA) were also examined. NOHARM software (version 4.0) was further used to analyze the data by examining the dimensionality of the tests and reporting a two-parameter model.

Dimensionality

In order to apply the item-response theory for item analysis, it is essential that each test undergoes unidimensionality evaluation. Unidimensionality is one of the two assumptions in item-response theory. It denotes that only one single dominant factor affects a testee's performance, i.e. the test taker's ability which is being tested and measured. Another assumption is local independence, which means that responding to a single item will be independent of other items if the dominant factor (i.e. ability) is controlled (10).

Different models are suggested for the item-response theory, which are labelled by the scoring model (e.g. two-parameter, multi-parameter, and nominal) and number of parameters (e.g. difficulty and discrimination parameters and guessing effect) (11). For determining the number of parameters of an item, all three fitting types must be examined with the data, and the most appropriate one should be selected. In the present analysis, likelihood indexes were used for comparing and choosing the right model.

Dimensionality parameters

Difficulty parameter in IRT is similar to item difficulty in its classical counterpart but the difference is that in IRT as the values increase the item becomes more difficult, and test takers need a higher ability to get the item right. It ranges from -4 to +4, and it becomes more difficult as we move from -4 towards +4. While this value may fluctuate between 0 and 1 in the classical test theory, its IRT values may even exceed 1. Guessing parameter estimates that to what extent an individual test taker with a very low ability can correctly answer an item. Low values of this parameter (below 0.1) is acceptable but above that is unsatisfactory. Items with guessing parameter above 0.25 are inappropriately constructed items due to higher guessing likelihood. Values below 0.1 are considered optimal items in a test. While references on the IRT models do not present clear-cut classifications for the parameters in question, Baker (13) developed and suggested a scale for difficulty and discrimination parameters, which is the basis of our analysis (Tables 2 and 3) too.

Item-Response Theory (IRT) Models

A variety of IRT models are available to accommodate different measurement situations. In a one-parameter model or the Rasch model, it is assumed that the discrimination parameter remains the same for all items but for each item, a difficulty parameter can be specified. An advantage of the Rasch model is its capacity to be

used with smaller samples sizes. However, if equal discrimination is not assumed, the two-parameter model is applied where two parameters affecting an individual's response to a particular test item are considered (i.e. difficulty level and item discrimination). Therefore, a difficulty level and a discrimination power value are separately reported for each item, as reported below.

Table 2. Levels of the difficulty parameter

2.001 to 3	1.001 to 2	0.001 to 1	-1 to 0.001	-1.001 to -2	-2.001 to -3
Very Difficult	Almost Difficult	Difficult	Almost Easy	Easy	Very Easy

Table 3. Levels of the discrimination parameter

0 to 0.009	0.01 to 0.34	0.35 to 0.64	0.65 to 1.34	1.35 to 1.69	≥ 1.70
None	Very low	Low	Medium	High	Very High

Item difficulty

Item difficulty is the total percentage of testers who score a certain item right and is represented by P. As the following formula indicates, P is computed by the number of testees who correctly answered a certain item (R) divided by the total number of test takers (T) multiplied by 100.

$$P = R/T \times 100$$

Item discrimination

Represented by D, item discrimination power is an index that indicates how well an item is able to distinguish between high and low achievers. It is computed from equal-sized high and low-scoring groups on a test by subtracting the number of successes of the low-achievers on the item from the number of successes of the high-achieving group and then dividing this difference by the size of a group using D = (UG - LG)/ n formula. It may range from + 1 to -1. The higher the discrimination index, the better the test item can discriminate between students with higher test scores and those with lower test scores. For instance, D = 0 means the item has no discriminatory power, while D = 1 means the item has the highest perfect discrimination power.

Formula 2.

D = (upper group right answers - lower group right answers) ÷ number of group members (upper or lower)

Finally, when guessing is plausible, the three-parameter logistic model applies and three parameters affecting an individual's response to a particular test item are reported (difficulty level, discriminating power and the guessing effect) (12). But the decision to use one model over another depends on several factors, including the response format, whether the discrimination parameter can be kept constant across items, whether guessing is plausible, and whether different category response parameters must be estimated for each item on a scale (10).

Results

Analysis of the collected data showed that the enrollment of female participants outnumbered that of male ones throughout the three years; however, the proportion varied from one-fourth to almost one-third (Table 1). The ratio remained almost the same when we considered the total number of test takers by gender. As for absentees, 26% of them were male and the rest (74%) were female. Admitted participants consisted of 70% females and 30% males. Considering the number of admitted candidates, 8.7% the total female participants and 10.3% of the total male participants were admitted (Table 1).

between English test scores and GPA

In the present study, due to the large sample size, and the quantitative nature of English language scores and GPA, Pearson's correlation coefficient was applied to investigate the possible correlation. In the present study, preliminary analyses were performed to ensure no violation of the assumptions of normality, linearity and homoscedasticity. Small size but significant correlation was observed ($p < 0.05$) ($r = 0.260$; confidence interval 95%). In other words, the higher an applicant's GPA, the higher his/her English test score.

Evaluation of dimensionality

At first, NOHARM software (version 4.0) was used to check the dimensionality of the test (four sessions in 2021, each session containing 40 items). The Tanaka index values in the output of the software confirmed the unidimensionality of the test (e.g. for Session 1, Tanaka index of goodness of fit = 0.9853312, and Root Mean Square of Residuals (RMSR) or lower off-diagonals = 0.0090737; details of other three sessions are available on demand). If the Tanaka index value is greater than 0.90, the fit is acceptable, and if it is greater than 0.95, the fit is good. Considering that the value of the obtained indexes in all four sessions were above 0.95, the four tests were considered unidimensional. In addition, the

very low value of RMSR was another proof of the suitability of the unidimensional model, leading to the enhanced dependability of the tests; dependability is seen as the extent to which test results reflect the level of the construct we are meant to measure (14). In other words, only one dominant factor had an effect on the subjects' performance and, here, this dominant factor was the desired ability (i.e. language knowledge) of the individual.

Model selection

To choose the right model, the significance of the difference between the likelihood indices of the two models should be examined. Here, the difference between the likelihood indices between the one- and two-parameter models was greater than the value of the Chi-square table. As a result, the null hypothesis of no difference between the one- and two-parameter models was rejected. On the other hand, the value of this difference between the two- and three-parameter models was lower than the value of the Chi-square table, which confirmed the null hypothesis that there was no difference between these two models; therefore, the two-parameter model was used for analysis (Table 4).

Because the data we acquired for the present study were the result of four administrations each year, test items are analyzed separately and reported below. Before entering the exam analysis, it is necessary to mention that in the graduate exam, five parallel sets of questions are given to candidates who take the exam at the same time. In other words, the candidates of a series of similar fields take the exam at the same time (except for the medical journalism, which has its own set of questions). Accordingly, the answer sheets of all the candidates were subject to analysis.

A) The 2021 graduate exam (Session 1)

The first session of the 2021 graduate exam was conducted with 40 questions, administered to 13,290 participants. The maximum and minimum scores of the exam were obtained at 38.31 and 0.11, respectively; in addition, the reliability was calculated at 0.92 Table 5.

B) The 2021 graduate exam (Session 2)

The second session of the 2021 graduate exam was conducted with 40 questions, administered to 15,422 participants. The maximum and minimum scores of the exam were 34.55 and 0.03, respectively. The reliability value was obtained at 0.88 (Table 6).

C) The 2021 graduate exam (Session 3)

The third session of the 2021 graduate exam was conducted with 40 questions, administered to 9,441 participants. The maximum and minimum scores of the exam were 37.38 and 0.09, respectively. The reliability value was obtained at 0.90 (Table 7).

D) The 2021 graduate exam (Session 4)

The fourth session of the 2021 graduate exam was conducted with 40 questions, administered to 9,262 participants. The maximum and minimum scores of the exam were 38.12 and 0.15, respectively. The reliability value was obtained at 0.91 (Tables 8–10).

Table 4. Likelihood indices among the models (2021–Sessions 1, 2, 3, 4)

Year/Session	One-parameter	Two-parameter	Three-parameter
2021 - Session 1	-244675.2	-242027.2	-242029.9
2021 - Session 2	-244675.2	-242027.2	-242029.9
2021 - Session 3	-161145.5	-159467.2	-159462.4
2021 - Session 4	-165435.3	-163290.5	-163276.1

Notes: Lower values of likelihood indices indicate better fit of the model to the data, aiding in model comparison and selection for the analysis of test items. The selection of the appropriate model is crucial for accurate analysis and interpretation of the test data.

Table 5. Item difficulty and discrimination (2021–Session 1)

Question	Discrimination	Discrimination Power	Difficulty	Level of difficulty
q121	1.999	Very high	-0.32	Almost easy
q122	1.712	Very high	0.97	Almost difficult
q123	2.051	Very high	1.44	Difficult
q124	1.741	Very high	1.05	Difficult
q125	1.744	Very high	1.56	Difficult
q126	2.047	Very high	1.29	Difficult
q127	1.561	High	1.41	Difficult
q128	2.346	Very high	1.07	Difficult
q129	1.689	High	1.77	Difficult
q130	2.594	Very high	1.99	Difficult
q131	2.075	Very high	2.01	Very difficult
q132	1.693	High	0.91	Almost difficult
q133	2.247	Very high	1.63	Difficult

q134	1.422	High	2.78	Very difficult
q135	2.184	Very high	1.91	Difficult
q136	1.881	Very high	2.07	Very difficult
q137	1.285	Medium	1.13	Difficult
q138	2.527	Very high	0.97	Almost difficult
q139	1.932	Very high	1.55	Difficult
q140	2.425	Very high	0.33	Almost difficult
q141	1.477	High	2.02	Very difficult
q142	1.391	High	-0.53	Almost easy
q143	1.422	High	0.19	Almost difficult
q144	2.285	Very high	1.39	Difficult
q145	1.676	High	-0.17	Almost easy
q146	1.743	Very high	2.18	Very difficult
q147	1.995	Very high	1.70	Difficult
q148	1.424	High	1.10	Difficult
q149	2.424	Very high	1.35	Difficult
q150	2.264	Very high	1.23	Difficult
q151	2.2	Very high	2.09	Very difficult
q152	2.153	Very high	0.43	Almost difficult
q153	1.991	Very high	1.36	Difficult
q154	1.223	Medium	1.62	Difficult
q155	1.002	Medium	1.45	Difficult
q156	2.523	Very high	0.56	Almost difficult
q157	2.808	Very high	0.33	Almost difficult
q158	1.472	High	1.34	Difficult
q159	2.119	Very high	0.54	Almost difficult
q160	0.686	Medium	2.58	Very difficult
Average difficulty			1.26	Difficult
Average discrimination			1.89	Very high

Notes: Discrimination values indicate the ability of an item to differentiate between high and low performers, with higher values suggesting stronger discrimination. Difficulty values represent the level of difficulty for each item, with negative values indicating easier items and positive values indicating more difficult items.

Table 6. Item difficulty and discrimination (2021–Session 2)

Question	Discrimination	Power of Discrimination	Difficulty	Level of difficulty
q121	1.543	High	1.72	Difficult
q122	2.059	Very high	2.48	Very difficult
q123	1.62	High	1.52	Difficult
q124	2.118	Very high	1.86	Difficult
q125	1.853	Very high	2.54	Very difficult
q126	1.843	Very high	2.77	Very difficult
q127	1.664	High	2.49	Very difficult
q128	2.276	Very high	2.70	Very difficult
q129	1.496	High	2.70	Very difficult
q130	1.964	Very high	1.56	Difficult
q131	1.108	Medium	2.77	Very difficult
q132	1.944	Very high	2.31	Very difficult
q133	2.411	Very high	2.72	Very difficult
q134	1.624	High	2.48	Very difficult
q135	1.941	Very high	3.14	Very difficult
q136	2.237	Very high	2.49	Very difficult
q137	1.093	Medium	2.16	Very difficult
q138	2.048	Very high	2.92	Very difficult
q139	2.089	Very high	1.22	Difficult
q140	1.717	Very high	0.78	Very difficult
q141	1.064	Medium	1.94	Difficult
q142	1.785	Very high	1.81	Difficult
q143	2.22	Very high	1.74	Difficult
q144	1.927	Very high	1.66	Difficult
q145	1.978	Very high	0.41	Very difficult
q146	1.796	Very high	1.91	Difficult

q147	2.086	Very high	1.18	Difficult
q148	2.295	Very high	2.18	Very difficult
q149	2.344	Very high	3.60	Very difficult
q150	2.942	Very high	2.18	Very difficult
q151	1.846	Very high	0.88	Very difficult
q152	2.076	Very high	1.03	Difficult
q153	2.198	Very high	0.22	Very difficult
q154	0.785	Medium	1.94	Difficult
q155	2.489	Very high	2.79	Very difficult
q156	1.922	Very high	3.10	Very difficult
q157	1.571	High	3.03	Very difficult
q158	1.845	Very high	2.52	Very difficult
q159	1.556	High	0.58	Very difficult
q160	2.204	Very high	3.42	Very difficult
Average difficulty			1.89	Difficult
Average discrimination			2.09	Very high

Note: Discrimination values reflect the ability of each item to discriminate between high and low performers, with higher values indicating stronger discrimination. Difficulty values represent the level of difficulty for each item, with higher values indicating more difficult items.

Table 7. Item difficulty and discrimination (2021–Session 3)

Question	Discrimination	Power of discrimination	Difficulty	Level of difficulty
q121	1.761	Very high	1.65	Difficult
q122	1.012	Medium	1.82	Difficult
q123	1.775	Very high	1.59	Difficult
q124	2.187	Very high	0.57	Almost difficult
q125	0.934	Medium	2.44	Very difficult
q126	1.697	High	1.73	Difficult
q127	2.048	Very high	2.19	Very difficult
q128	2.228	Very high	1.66	Difficult
q129	1.59	High	2.30	Very difficult
q130	1.868	Very high	0.63	Almost difficult
q131	2.204	Very high	2.82	Very difficult
q132	2.507	Very high	1.30	Difficult
q133	1.272	Medium	2.46	Very difficult
q134	2.226	Very high	1.59	Difficult
q135	1.716	Very high	1.98	Difficult
q136	1.678	High	2.55	Very difficult
q137	0.978	Medium	3.24	Very difficult
q138	1.82	Very high	2.30	Very difficult
q139	2.247	Very high	2.29	Very difficult
q140	1.935	Very high	1.29	Difficult
q141	1.324	Medium	1.53	Difficult
q142	1.471	High	-0.49	Almost easy
q143	1.488	High	0.36	Almost difficult
q144	1.637	High	0.25	Almost difficult
q145	1.644	High	1.55	Difficult
q146	1.756	Very high	0.84	Almost difficult
q147	1.807	Very high	1.35	Difficult
q148	2.999	Very high	2.76	Very difficult
q149	1.972	Very high	1.98	Difficult
q150	2.344	Very high	1.42	Difficult
q151	2.368	Very high	1.83	Difficult
q152	2.236	Very high	1.02	Difficult
q153	2.468	Very high	1.62	Difficult
q154	3.205	Very high	2.31	Very difficult
q155	2.71	Very high	1.23	Difficult
q156	1.894	Very high	-0.23	Almost easy
q157	1.714	Very high	2.43	Very difficult
q158	2.384	Very high	2.29	Very difficult

q159	1.792	Very high	2.16	Very difficult
q160	1.898	Very high	1.55	Difficult
Average difficulty		1.92		Difficult
Average discrimination		1.65		High

Notes: Discrimination values indicate the extent to which each item distinguishes between high and low performers, with higher values suggesting stronger discrimination. Difficulty values represent the level of difficulty for each item, with higher values indicating greater difficulty.

Table 8. Item difficulty and discrimination levels (2021 – Session2021–Session 4)

Question	Discrimination	Power of Discrimination	Difficulty	Degree of difficulty
q121	1.638	High	-0.79	Almost easy
q122	1.541	High	0.83	Almost difficult
q123	1.418	High	1.99	Difficult
q124	1.452	High	1.61	Difficult
q125	1.366	High	0.91	Almost difficult
q126	1.811	Very high	0.06	Almost difficult
q127	1.652	High	2.10	Very difficult
q128	1.404	High	2.18	Very difficult
q129	1.586	High	1.30	Difficult
q130	1.071	Medium	2.26	Very difficult
q131	1.414	High	1.41	Difficult
q132	1.241	Medium	1.47	Difficult
q133	0.829	Medium	2.33	Very difficult
q134	2.124	Very high	0.71	Almost difficult
q135	0.813	Medium	2.07	Very difficult
q136	2.351	Very high	2.12	Very difficult
q137	0.574	Few	2.75	Very difficult
q138	1.321	Medium	2.47	Very difficult
q139	1.094	Medium	1.22	Difficult
q140	1.945	Very high	1.60	Difficult
q141	1.795	Very high	0.45	Almost difficult
q142	1.559	High	0.36	Almost difficult
q143	1.785	Very high	1.00	Difficult
q144	0.519	Few	2.26	Very difficult
q145	2.062	Very high	0.01	Almost difficult
q146	2.19	Very high	2.19	Very difficult
q147	1.78	Very high	2.54	Very difficult
q148	1.758	Very high	2.49	Very difficult
q149	2.351	Very high	2.30	Very difficult
q150	2.038	Very high	1.68	Difficult
q151	1.823	Very high	-0.02	Almost easy
q152	1.807	Very high	2.26	Very difficult
q153	2.688	Very high	1.21	Difficult
q154	2.414	Very high	2.22	Very difficult
q155	1.951	Very high	1.48	Difficult
q156	1.69	High	0.22	Almost difficult
q157	2.002	Very high	1.46	Difficult
q158	1.265	Medium	1.08	Difficult
q159	1.484	High	0.85	Almost difficult
q160	2.314	Very high	1.12	Difficult
Average difficulty		1.44		Difficult
Average discrimination		1.65		High

Notes: Discrimination values indicate the extent to which each item distinguishes between high and low performers, with higher values suggesting stronger discrimination. Difficulty values represent the level of difficulty for each item, with higher values indicating greater difficulty.

Table 9. Reliability and summary of average item difficulty and discrimination (2021–Sessions 1, 2, 3, 4)

Year	Sessions	Parameter	Values	Interpretation	Reliability
2019	Session 1	Average difficulty	1.26	Difficult	0.92
		Average discrimination	1.89	Very high	

Session 2	Average difficulty	1.89	Difficult	0.88
	Average discrimination	2.09	Very high	
Session 3	Average difficulty	1.92	Difficult	0.90
	Average discrimination	1.65	High	
Session 4	Average difficulty	1.44	Difficult	0.91
	Average discrimination	1.65	High	

Notes: Average difficulty values represent the average level of difficulty across all test items for each session, with higher values indicating greater difficulty. Discrimination values indicate the discriminatory power of test items, with higher values suggesting better discrimination between high and low performers. Reliability coefficients measure the consistency and stability of test scores, with values closer to 1.00 indicating higher reliability.

Table 10. Status summary of questions (2021–Sessions 1, 2, 3, 4)

Power of discrimination	Frequency				Level of difficulty	Frequency			
	Session 1	Session 2	Session 3	Session 4		Session 1	Session 2	Session 3	Session 4
No	0	0	0	0	Very easy	0	0	0	0
Very low	0	0	0	0	Easy	0	0	0	0
Low	0	0	0	2	Almost easy	3 (7.5%)	0	2 (5%)	2 (5%)
Medium	4 (10%)	4 (10%)	5 (12.5%)	7 (17.5%)	Almost difficult	9 (22.5%)	5 (12.5%)	5 (12.5%)	9 (22.5%)
High	10 (25%)	7 (17.5%)	7 (17.5%)	12 (30%)	Difficult	21 (52.5%)	13 (32.5%)	19 (47.5%)	14 (35%)
Very high	26 (65%)	29 (72.5%)	28 (70%)	19 (47.5%)	Very difficult	0	22 (55%)	14 (35%)	15 (37.5%)
Negative	0	0	0	0					

Notes: Power of discrimination indicates the ability of questions to distinguish between high and low performers, with higher values representing better discriminatory power. Frequency counts show the number of questions falling into each category of difficulty and discrimination level for each session. Interpretations of difficulty levels, ranging from "Very easy" to "Very difficult," aid in understanding the distribution of questions based on their perceived difficulty. The absence of questions in certain difficulty or discrimination categories is denoted by "0" frequency counts.

Discussion

This study aimed to examine the dimensionality of English test items on the nationwide graduate entrance exam for healthcare applicants and to report test-takers' characteristics. The characteristics of participants in three test packages belonging to three successive years (2019–2021) were described; four sessions in 2021 were analyzed for fitting a dimensional model. The results showed that female participants outnumbered male participants throughout the three years, and the proportion of admitted participants was similar (70% females vs. 30% males).

Additionally, a positively high correlation between participants' GPA and English test scores was observed ($p<0.05$); in fact, the higher the participants' GPA was, the greater the English test scores at the master's entrance examination. While these findings highlight the importance of English language teaching in healthcare education, complementary views stress the significance of entrance test results as a predictor of test takers' success and excellence in their majors (18). In simpler terms, this reciprocal influence underscores the pressing necessity to incorporate English proficiency assessments into master's entrance exams. Doing so acts as a gateway, granting an edge to individuals with advanced English skills, and serves as a predictor of their prospective success in their chosen fields of study.

In addition, the results of four administrations in 2021 were analyzed separately and reported in detail as a sample. In fact, all four tests had high reliability indices (i.e., 0.92, 0.88, 0.90, and 0.91). In other words, 92%, 88%, 90% and 91% of the variation among test measures was reliable, and only 8%, 12%, 10%, and 9% (applicable to the four tests) of the variance was attributed to measurement errors (19).

An important finding was the suitability of a unidimensional model to some extent, leading to the enhanced dependability of the tests. In fact, dependability tests revealed that only one dominant factor had an effect on the subjects' performance (i.e., language knowledge) (17, 20) Similar findings from Oman are reported in favor of psychometrically sound test items to attain satisfactory levels of unidimensionality to bridge the difficulty level of a test and participants' ability (21). A further proof comes from the reliability coefficients (e.g., the four test reliability indices: 0.92, 0.88, 0.90, and 0.91). Similarly, the MHLE was reported to have a reliability of 0.862 (10) which is considered an acceptable reliability index (22). These two tests are both designed and administered by the

Center for the Measurement of Medical Education to assess the language abilities of healthcare major applicants.

Another important aspect explored in the present study was the examination of dimensionality parameters, where only two parameters (i.e., the difficulty parameter and the discrimination parameter) fit the model and the third dimension (i.e., the guessing parameter) did not apply. The study analysis revealed that English language test results in all four sessions were "difficult," with either "high" or "very high" discrimination power. In fact, neither "easy" nor "very easy" items were found in the tests; also, none of the items were associated with "no" or "very low" discrimination power. While we did not find studies on graduate exams for healthcare applicants, a similar study was conducted on the Ministry of Science, Research and Technology (MSRT), which is a high-stakes English language proficiency test (23). Analysis of the difficulty and discrimination indices of the total test revealed that 14% of the test items were either easy or very easy, 38% were medium, and 48% were either difficult or very difficult. This finding is not in line with our findings because they examined the whole test, including other sections (listening comprehension, structure, and written expressions, along with reading comprehension); however, the present study examined only vocabulary and reading comprehension items. They classified 14% of the total items as nonfunctioning, which discriminated negatively or did not discriminate at all; however, this was not the case in the present study. In their study, 38% of the items displayed satisfactory difficulty, but low discriminating power was reported because the items were too easy (14%) or too difficult (48%) (23).

While concerns about jeopardizing validity due to the difficulty of such tests have been raised (24), Table 10 shows a balance between degrees of difficulty and discrimination power. For instance, session 1 results indicated a 75% degree of either "almost difficult" or "difficult" items, while in the same session, "high" or "very high" discrimination power was observed for 80% of the items. For session 2, all the items were difficult, while 90% of the items were discerning. Similar patterns can be seen in the other two sessions.

Conclusion

In brief, the present study showed that language knowledge was indeed tested during the entrance examination since unidimensionality was observed. Additionally, difficulty and discrimination indices were evident in perspective, with no traces of the guessing effect. We found that the four sessions functioned well enough, with high reliability indices and good quality test items in terms of difficulty and discrimination. Overall, a good balance was observed between the two parameters (i.e., difficulty and discrimination) (see Table 9 for details). Additionally, acceptably high reliability indices (i.e., 0.92, 0.88, 0.90, and 0.91) were observed in all four administration sessions. For dimensionality parameters, the four tests proved to show acceptable levels of difficulty, with either "high" or "very high" discrimination power as a nationwide exam. No "easy" or "very easy" items were found. In addition, no items were associated with "no" or "very low" discrimination power.

With respect to the study limitations, the main concern was the confidentiality of the test takers' performance. Additionally, obtaining the study data from the examination board required special arrangements that took a long time. In addition, we recommend that other researchers test the validity of the scale in future studies. We further feel that the concept of academic English was not fully operationalized as a construct due to practical limitations; for developing valid high-stakes tests, the inclusion of listening, writing, and speaking sections is suggested for future administrations. Another serious challenge may concern the consequential validity and occupational requirements for healthcare students, given the current trends (1, 25). A stronger emphasis should be placed on washback to bring about positive changes in teaching English to students in healthcare domains at the undergraduate level (11, 26), as well as on the revision of instructional systems at the graduate level. Future studies may focus on interviews with test developers and test takers to explore unheard voices.

Ethical considerations

The required data were obtained under confidentiality requirements from the Center for the Measurement of Medical Education directed by the Ministry of Health and Medical Education, without test takers' personal information (e.g., name or identity information). The study was approved (Ethics code: IR.SBMU.RETECH.REC.1399.1222) by the Ethics Committee of Shahid Beheshti University of Medical Sciences, Tehran, Iran.

Artificial intelligence utilization for article writing

No AI was employed to draft the present article, so the manuscript was written by the authors without using AI.

Acknowledgments

We are grateful to the research committee of Shahid Beheshti University of Medical Sciences, Tehran, Iran, for approving the research proposal. Also, we would like to thank the Center for the Measurement of Medical Education for provision of the relevant data.

Conflict of interest statement

None.

Author contributions

Seyyed Samad Sajjadi has supervised the study; Nematullah Shomoossi has designed and drafted the proposal and manuscripts; Enayat Shabani has obtained the data and assisted in the data analysis; Abdurrashid Khazaei Feizabadi has designed and drafted the proposal; all authors have read, revised and approved the final version of the article.

Supporting resources

The study was funded and approved (Ethics Code: IR.SBMU.RETECH.REC.1399.1222) by the Ethics Committee of Shahid Beheshti University of Medical Sciences, Tehran, Iran.

Data availability statement

Relevant data are reported in the present article. Further data will be accessible for researchers only with the permission of the Iranian Center for the Measurement of Medical Education.

References

1. Dillon G, Boulet J, Hawkins R, Swanson D. Simulations in the United States medical licensing examination™(USMLE™). BMJ Quality & Safety. 2004;13(suppl1):i41-i5. [https://doi.org/10.1136/qshc.2004.010025]

2. Schwartzstein RM, Rosenfeld GC, Hilborn R, Oyewole SH, Mitchell K. Redesigning the MCAT exam: Balancing multiple perspectives. Academic Medicine. 2013;88(5):560-7. [https://doi.org/10.1097/ACM.0b013e31828c4ae0]

3. Séguis B, McElwee S. Assessing clinical communication on the Occupational English Test®. Global perspectives on language assessment: Research, theory, and practice. 2019:63-79. [https://doi.org/10.1177/07342829211050537]

4. Anjali S, Sanjay Z, and Bipin B. India's foreign medical graduates: an opportunity to correct India's physician shortage. Educ Health (Abingdon). 2016;29(1):42-6. [https://doi.org/10.4103/1357-6283.178932]

5. Thappa DM. Jawaharlal Institute of postgraduate medical education and research, pondicherry, India. Journal of Postgraduate Medicine. 2001;47(2):147.

6. Khodi A, Alavi SM, Karami H. Test review of Iranian university entrance exam: English Konkur examination. Language Testing in Asia. 2021;11:1-10. [https://doi.org/10.1186/s40468-021-00125-6]

7. Lotfie MM. Language policy and practices in indonesian higher education institutions. Intellectual Discourse, 2018.26(2):p. 683–704-683–704.

8. Karakas, A. Turkish lecturers' and students' perceptions of English in English-medium instruction universities. 2016.

9. Özdemir-Yılmazer M. Direct Access to English-Medium Higher Education in Turkey: Variations in Entry Language Scores. Dil Eğitimi ve Araştırmaları Dergisi, 2022.8(2):p.325-345. [https://doi.org/10.31464/jlere.1105651]

10. Marandi SS, Tajik L, Zohali L. On the construct validity of the Iranian Ministry of Health Language Exam (MHLE). Journal of Language Horizons. 2020;4(2):9-36. [https://doi.org/10.22051/lghor.2020.28036.1180]

11. Hekmati N, Davoudi M, Zareian G, Elyasi M. English for medical purposes: An investigation into medical students' English language needs. Iranian Journal of Applied Language Studies. 2020;12(1):151-76. [https://doi.org/10.22111/IJALS.2020.5648]

12. ShayesteFar P. A model of interplay between student English achievement and the joint affective factors in a high-stakes test change context: Model construction and validity. Educational Assessment, Evaluation and Accountability. 2020;32(3):335-71. [https://doi.org/10.1007/s11092-020-09326-8]

13. Nguyen T, Han H, Kim M, Chan K. An introduction to item response theory for patient-reported outcome measurement. Patient. 2014;7(1):23–35. [https://doi.org/10.1007/s40271-013-0041-0].

14. Deng S, Bolt DM. A sequential IRT model for multiple-choice items and a multidimensional extension. Applied Psychological Measurement. 2016;40(4):243-57. [https://doi.org/10.1177/0146621616631518]

15. Baker FB. The basics of item response theory. 2nd ed. ERIC Clearinghouse on Assessment and Evaluation; College Park, MD, USA: 2001.
16. Kim S-H, Kwak M, Bian M, et al. Item response models in psychometrika and psychometric textbooks. Frontiers in Education. 2020 Jun 9 (Vol. 5, p. 63). Frontiers Media SA. [https://doi.org/10.3389/feduc.2020.00063]
17. Sheybani E, Zeraatpishe M. On the dimensionality of reading comprehension tests composed of text comprehension items and cloze test items. International Journal of Language Testing. 2018;8(1):12-26.
18. Ocbian MM, MP. Gamba, and J.D. Ricafort, Admission Test as Predictor of Performance of Students in the English Subject. JPAIR Institutional Research, 2015;6(1):p.34-45.
19. Abdellatif H, Al-Shahran i AM. Effect of blueprinting methods on test difficulty, discrimination, and reliability indices: cross-sectional study in an integrated learning program. Advances in medical education and practice, 2019:p.23-30. [https://doi.org/10.2147/AMEP.S190827]
20. Baghaei, P. and V. Aryadoust. Modeling local item dependence due to common test format with a multidimensional Rasch model. International Journal of Testing, 2015.15(1):p.71-87. [https://doi.org/10.1080/15305058.2014.941108]
21. Abdellatif, H., Test results with and without blueprinting: Psychometric analysis using the Rasch model. Educación Médica, 2023.24(3):p.100802. [https://doi.org/10.1016/j.edumed.2023.100802]
22. Hughes, A. Testing for language teachers. 2020: Cambridge university press. [https://doi.org/10.1017/9781009024723]
23. Ghahraki S, Tavakoli M, Ketabi S. Applying a two-parameter item response model to explore the psychometric properties: The case of the Ministry Of Science, Research And Technology (MSRT) high-stakes English language proficiency test. Two Quarterly Journal of English Language Teaching and Learning University of Tabriz. 2022;14(29):1-26. [https://doi.org/10.22034/ELT.2021.46325.2396]
24. Bazvand AD, Kheirzade S, Ahmadi A. On the statistical and heuristic difficulty estimates of a high stakes test in Iran. International Journal of Assessment Tools in Education. 2019;6(3):330-43. [https://doi.org/10.21449/ijate.546709]
25. Shomoossi N, Rad M, Fiezabadi M, Vaziri E, Amiri M. Understanding the research process and historical trends in English for medical purposes using scientometrics and co-occurrence analysis. Acta Facultatis Medicae Naissensis. 2019;36(3):235-47. [https://doi.org/10.5937/afmnai1903236S]
26. Shomoossi N, Rad M, Rakhshani MH. Efficacy of English language programs as judged by nurses and students of nursing: Do nurses in Iran need to know English? Acta Facultatis Medicae Naissensis. 2013;30(3):137. [https://doi.org/10.2478/afmnai-2013-0005]