QUALITY AND FEATURE OF MULTIPLE-CHOICE QUESTIONS IN EDUCATION

Bing Jia, Dan He, Zhemin Zhu
Beihua University, China
E-mail: jia.bing@foxmail.com, 15337614@qq.com, zhuzm485@nenu.edu.cn

Abstract

The quality of multiple-choice questions (MCQs) as well as the student's solve behavior in MCQs are educational concerns. MCQs cover wide educational content and can be immediately and accurately scored. However, many studies have found some flawed items in this exam type, thereby possibly resulting in misleading insights into students' performance and affecting important decisions. This research sought to determine the characteristics of MCQs and factors that may affect the quality of MCQs by using item response theory (IRT) to evaluate data. For this, four samples of different sizes from US and China in secondary and higher education were chosen. Item difficulty and discrimination were determined using item response theory statistical item analysis models. Results were as follows. First, only a few guessing behaviors are included in MCQ exams because all data fit the two-parameter logistic model better than the three-parameter logistic model. Second, the quality of MCQs depended more on the degree of training of examiners and less on middle or higher education levels. Lastly, MCQs must be evaluated to ensure that high-quality items can be used as bases of inference in middle and higher education.
Keywords: *higher education, item evaluation, item response theory, multiple-choice test, secondary education*

Introduction

In education exams, multiple-choice questions (MCQs) are commonly used in secondary and higher education because they can be easily and accurately scored and save significant manpower and time. Although some studies suggest that MCQs only focus on what students remember and do not assess the extent to which they understand, analyze, and apply course-related information (Walsh & Seldomridge, 2006), MCQs remain among the most common types of assessment questions extensively used in standardized tests (Bailey et al., 2012; DiBattista & Kurzawa, 2011; Zhu et al., 2018). On the one hand, MCQs can immediately cover instructional contents and be scored easily (Brown & Abdulnabi, 2017; DiBattista & Kurzawa, 2011; Nedeau-Cayo et al., 2013). On the other hand, MCQs provide students with a method based on their experiences. Examiners encourage examinees to guess whenever they can eliminate a wrong choice, which is a better strategy than completely blind guessing (Frary, 1988, p. 76). When the guessing process is not random, the success of the guessing process will be based on the examinees' abilities (San Martín et al., 2006; van der Maas et al., 2011; Zhu et al., 2018). Hence, the chosen options in response to MCQs provide information of the examinees' experiences. Only a few teachers have formal education on the rules of MCQ writing or MCQ assessment (Brown & Abdulnabi, 2017). Thus, when the quality of MCQs is not good owing to the lack of teacher training, test results may mislead the assessment of examinee achievement (Brady, 2005; Brown & Abdulnabi, 2017; Downing, 2005; Masters et al., 2001; Stagnaro-Green & Downing, 2006; Tarrant et al., 2006). Therefore, MCQ quality should be evaluated in the field of education.

This research was based on exam data from China and the US to obtain a general conclusion. Although several studies had shown a gap between the two countries in the teacher

https://doi.org/10.33225/pec/20.78.576 ISSN 1822-7864 (Print) ISSN 2538-7111 (Online)

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 78, No. 4, 2020

and student levels (Stevenson et al., 1990; Stevenson & Stigler, 1994), MCQs had been widely used in China and the US. However, China has more students and smaller proportion of teachers and students compared with the US. Therefore, MCQs could compensate for the heavy work pressure. Moreover, examinations in China and the US may provide different information. This research used item response theory (IRT) and selected four exams in the two countries to evaluate the quality of MCQs. First, this research briefly reviewed the methods of assessing MCQ quality. Second, the IRT models that were used to assess MCQs in this research are introduced. Third, four MCQ exam data sets were assessed by using IRT models.

Assessment of MCQ Quality

Three methods are used to assess the quality of MCQs. The first is a conventional method based on process control. This method has four steps in writing MCQ items, namely, (1) defining content, (2) choosing style and format, (3) writing items, and (4) writing options (Haladyna & Rodriguez, 2013); and applies five items to assess item quality: (1) questions are clearly stated, (2) questions are free of errors, (3) distractors are feasible, (4) accompanying explanations are good, and (5) specified answer is correct (Purchase et al., 2010). However, these criteria are more about prevention than evaluation.

The second method is based on classical test theory (CTT). CTT indicates that the sum of examinees' true scores based on theoretical ability and unobserved random error is equal to the scores of individual examinees in the test (observed score = true score + error). Thus, the examinees' actual levels of ability are assessed by the number of correctly answered items (de Ayala, 2009; Schaughency et al., 2012). Difficulty and discrimination are the most common CTT indicators of one MCQ item. The difficulty (p) of each item can be signed by the proportion of candidates who answered an item correctly. Items $p > .80$ or $p < .20$, which means too easy or too difficult, respectively, will be rejected from the test because they fail to provide sufficient useful information on the examinees' abilities (Brown & Abdulnabi, 2017). Discrimination (r) of each item can be signed using the Pearson product-moment correlation coefficient between the scores on the items and on the total test. Discrimination is a parameter that distinguishes between higher and lower ability examinees (Fan, 1998). Items that do not have a significantly positive value ($r > .20$) should be rejected (Ebel & Frisbie, 1991). However, CTT has two major limitations. (1) The observed score depends on the item sample and (2) the item statistical parameters depend on the examinee sample (Fan, 1998). These two limitations are summarized as circular dependency. Accordingly, estimating results in CTT depends heavily on samples because if the test is difficult, then students will obtain low scores. Hence, examinees will appear to be low achievers even if they have high levels of ability and vice versa (Hambleton et al., 1993). Similarly, the parameters of MCQs depend on the abilities of the sampled examinees, and changes in their abilities will affect the item parameters (Brown & Abdulnabi, 2017). When the class of examinees is changed, the same items will be assigned with different difficulty and discrimination values. From a statistical point of view, the reason for change is that CTT depends on the ability of the examinees and not on the distribution of student population abilities.

The third method is based on IRT. In 1911, Binet selected items for the Binet–Simon Intelligence Scale (Baker & Kim, 2004). The proportion of correct responses at each age (ability level) in the scale was obtained and presented in tabular form. Terman (1916) plotted the proportion of correct responses as a function of age and fitted a smooth line to these points. Given that the smooth line is S-shaped, the function is called item characteristic curve (ICC). IRT can be regarded as a series of statistical models used to fit ICC. These models that are based on the dual properties of items and the examinees' performance can be used to estimate their abilities (Hambleton & Jones, 1993). The item parameter estimation in IRT models is based on the distribution of examinees. Thus, item parameters (e.g., difficulties, discrimination, and guessing) may be different in the different examinee teams (Borsboom, 2005; Embretson & Reise, 2000; Hambleton et al., 1993). However, the parameters of one item for different examinee teams can be linked by linear invariance of the item parameters. Discrimination parameter (*a*) item and difficulty parameter (*b*) can be used to assess the quality of MCQs in IRT. The difficulty parameter is the point that equals examinees' abilities, in which the probability of answering items correctly is 50% (Embretson & Reise, 2000) and often has a range of $-4b4$. An item that is too difficult or too easy may lead the answers of the examinees to be correct or wrong. Consequently, the parameters of the IRT models may not be estimated.

In the *b* region, ICC is nearly linear with a slope of $a/\sqrt{2\pi}$ (Lord & Novick, 1968). Thus, an approximate interpretation of the discrimination parameter indicates the slope of ICC at the point on an ability scale corresponding to the difficulty. In IRT, answering additional questions correctly does not increase the examinees' abilities. In IRT, people's scores increase based more on the number of difficult questions the examinees answered than the number of questions the examinees answered correctly (Brown & Abdulnabi, 2017). The other difference with CTT is that the items, regardless of large or small discrimination, are significant in IRT, particularly in computer adaptive testing. However, items that have negative discrimination value remain undesirable. Negative discrimination value means that high-ability examinees will not easily obtain the correct answers, but low-ability examinees may easily determine the correct one. In addition, different models of IRT can be used to judge the behavior of examinees. CTT barely has such a function, which is the advantage of IRT.

Assessment of MCQs in Education

Assessment of MCQs is an important and necessary undertaking. MCQs are commonly used in secondary and higher education. In secondary education, MCQs can extensively cover instructional content. MCQs are the most common item types for some subjects, such as mathematics and physics. Hence, MCQs' quality must be guaranteed. In higher education, large-scale exams are administered for different majors or colleges. Teachers would have to exert effort to correct the papers. Therefore, teachers choose MCQs as the item types for the entire exam.

Assessment of MCQs in education follows two steps. The first step is designing MCQs. The design of the MCQ items must comply with guidelines for writing MCQs. Many guidelines are available (Brady, 2005; Burton, 2005; Downing & Yudkowsky, 2009; Haladyna, 2004; 2013). Research has claimed that teachers who trained in MCQ item writing can produce high-quality MCQs (Jozefowicz et al., 2002), and low-quality written MCQs may negatively impact examinees' performances or achievements (Clifton & Schriner, 2010; Downing, 2005; Tarrant et al., 2006). However, only a few academics have formal training in the principles of MCQ item writing (Brown & Abdulnabi, 2017). This lack of formal training causes the low quality of certain proportions of MCQs in exams (Downing, 2005; Ellsworth et al., 1990; Hansen & Dexter, 1997; Masters et al., 2001; Tarrant et al., 2006). Therefore, the second step, which is based on statistical method, is crucial.

The second step involves assessing the statistical properties of items by using a statistical method to exclude low-quality items. In general, teachers want to rule out items that are too difficult or simple, with worse discrimination, or easy to guess. In CTT, the discrimination parameter should be above +.20 (Ding & Beichner, 2009; Su et al., 2009; Thorndike, 2005). However, IRT is used as the assessment tool in the current research owing to the comparative disadvantages of CTT.

Models in IRT

The most popular model in IRT models is the two-parameter logistic (2PL) model (Birnbaum, 1968). The 2PL model can be used to calculate the probability of a correct item response with the item parameters and examinees' abilities. The probability of a correct response can be signed by Φ_{ij} and has the following form:

$$\Phi_{ij} = \frac{1}{1+\exp(-a_i(\theta_j - b_i)}, \quad (1)$$

where a_i is the item discrimination parameter, b_i is the item difficulty parameter with $\boldsymbol{i}$=1,…, $\boldsymbol{I}$ is the indexing items, $\boldsymbol{j}$=1,…, and $\boldsymbol{J}$ is indexing examinee subscript. When a_i equals 1, the 2PL model evolves to the one-parameter logistic (1PL) model. 1PL model has difficulty parameters only.

In an exam, the 2PL model can provide the item information by discrimination and difficulty parameters. However, the 2PL model does not include the guessing component. Given that the guessing process widely occurs in MCQs, many new models based on the 2PL model have been introduced by adding a guessing process *p* (*g*). The most popular one is the three-parameter logistic (3PL) model. The 3PL structure is an item response function with Φ_{ij} and a guessing component (San Martín et al., 2006). The structure is as follows:

$$P(\theta) = p(g) + (1 - p(g))p(r) . \quad (2)$$

In the 3PL model guessing function, is a constant based on an item, and the guessing function is reduced to a guessing parameter, such that, when p(r)=Φ_i.

Although the 3PL model is popular in IRT in the past 20 years, studies have found that the parameter recovery accuracy for a 3PL model depends on the extent of guessing presented in the data (Han, 2012; Holland, 1990; Pelton, 2002; San Martín et al., 2006). Thus, San Martín et al. (2006) suggested that the 3PL model can only be used when the sample size is extremely large. Birnbaum (1968) conjectured that low-ability examinees may select correct responses by chance. Accordingly, the guessing parameter should be the same as the chance level 1/*m*, where *m* is the number of response options in MCQs. Thereafter, the guessing function is reduced to a constant (p(g)=1/m), and this model is called 3PL with fixed lower asymptote (designated 3PL with FLA) when p(r)=Φ_i. The function is as follows:

$$P_i(\theta) = \frac{1}{m} + \left(1 - \frac{1}{m}\right)\Phi_i . \quad (3)$$

This model contains *g*-process and has more accuracy parameter recovery than the 3PL model.

MCQ Assessment by IRT

Assessment under IRT should continue to be considered with CTT. Lord and Novick (1968) found a contact of discrimination between CTT and IRT when the abilities of the examinees followed a normal distribution. The discrimination in IRT is slightly larger than in CTT (discrimination in IRT and CTT are .258 and .436 and .25 and .4, respectively). In CTT, the discrimination parameter should be above +.15 (Kehoe, 1995), +.20 (Ding & Beichner, 2009; Su et al., 2009; Thorndike, 2010), and +.25 (Considine et al., 2005). When discrimination is .3, ICC will lose the S-shape and behave similarly as a straight line. Therefore, the discrimination parameter in IRT in the current research is divided into three intervals, namely, $-\infty < a_i \leq$

0, $-\infty < a_i \leq .3$, and $.3 < a_i < +\infty$ which correspond to unacceptable, recommend delete, and acceptable, respectively.

When the difficulty of one item is extremely low or extremely high, its discriminatory parameter tends to suffer (DiBattista & Kurzawa, 2011). MCQs that are extremely difficult (difficulty per parameter <.30) or extremely easy (difficulty per parameter >.90) will have difficulty in discriminating high and low achievers in CTT (Ebel & Frisbie, 1991). No research has been conducted on the contact of difficulty between CTT and IRT. This lack of research may be caused by the essential difference of difficulty between the two theories. The former is concerned with how many people determine the correct answer, while the latter is concerned with the examinees' abilities. In general, the values of examinees' abilities are from −3 to +3. This research divides the difficulty in IRT into two intervals, $-2.4 < b_i \leq 1.2$ and otherwise, which correspond to unacceptable and acceptable, respectively (For the −3 to +3 interval, −2.4 and 1.2 correspond to 90% and 30%, respectively).

If the evaluation criteria for model selection (AIC and BIC) chose a guessing model, then items with a guessing parameter of >.25 are unacceptable when MCQs have four options (Brown & Abdulnabi, 2017). However, several studies have been convinced that guessing uses ability (San Martín et al. 2006; Zhu et al., 2018). The current research did not evaluate the quality of questions by guessing parameters. When a guessing model is selected by AIC or BIC, the examinees can be considered guessing in the exam. This method can be used to determine whether examinees were guessing in MCQs.

Model Size

Sample sizes should be considered because of the complexity of the IRT models (Hambleton & Jones, 1993). Research has suggested that the 1PL model will estimate accurately with $N < 100$ (Boone et al., 2014), and others are the opposite (Houts et al., 2016). IRT is more dependent on simple sizes (i.e., item and examinee sizes) than CTT (Akour & AL-Omari, 2013). Different estimation methods (e.g., MCMC, MLE, MH, and EM) have different requirements for sample size. Empirically, when the examinee sample is over 100 and the item size is above 50, 1PL, 2PL, and 3PL with FLA models will estimate accurately. However, the examinee samples should be above 500 for the 3PL model. Thus, the 3PL model may not converge estimates when the data are small.

Real Data

For exploring the normal nature of the MCQ exam, four samples of different sizes from US and China in secondary and higher education were chosen. All the data sets were analyzed using IRT methods to choose the best fitting model. And then, item parameters were estimated to assess the quality of MCQs.

Data Sets

The data sets comprised one from secondary education and one data set from higher education in the US (MEU and HEU, respectively); and one data set from middle education and one data set of higher education from China (MEC and HEC, respectively). All multiple-choice items contained four alternatives.

Table 1
Attributes of four data and writing teacher

Exam	Attribute			
	Attributes of Data		Attributes of Writing Teacher	
	Examinee Size	Item Size	Train	Company
MEU	2000	65	Trained	University Teacher
MEC	734	12	Not Trained	University Teacher
HEU	96	65	Not Trained	Middle School Teacher
HEC	1008	50	Being Trained	Middle School Teacher

The MEU data contained of 2,000 examinees. It was a state mathematics assessment that contained 65 MCQs. The data were used in an early research (Zhu et al., 2018). The HEU data were taken from 96 first-year student students and from a science freshman course in 2017 (spring). The exam consisted of 65 MCQs.The MEC data were from a mathematics simulated examination for a college entrance examination, which contained 12 MCQs. Different areas of China's college entrance examination use different exams in mathematics, but only 12 MCQs are included in each mathematics exam. In general, their difficulty parameters were increasing. HEC, which contained 50 MCQs, was from the final examination of Principles of Pedagogy. All registered students in the university need to take this exam. Therefore, a total of 1008 examinees from 11 majors (i.e., Biology, English, Physics, Mathematics, Chinese, Music, Art, History, Dance, Sports and Chinese Language and Literature) provided the data. The four MCQs were written by different teachers. MEU was written by the teacher who had passed MCQ writing training, and HEC was written by the teacher who participated in the training. HEU and MEC were written by teachers who were not trained in MCQ item writing. MEU and MEC were written by a university teacher, while HEU and HEC were written by middle school teachers.

Data Analysis

The data sets were analyzed using IRT. Four evaluation criteria for model selection were used in the real data: log likelihood (LL), −2LL, Akaike information criterion (AIC), and Bayesian information criterion (BIC). These are the most popular evaluation criteria in IRT. LL expresses the probability of a given set of observations for different values of statistical parameters. −2LL means −2 multiplied by LL. Given a set of competitive models for designated data, AIC estimates the quality of each model relative to the fitting between data and models. BIC is similar to AIC but adds a penalty term for the number of parameters. The model with the largest LL (lowest -2LL, lowest AIC, lowest BIC) value which means the best fitting model should be preferred. All exams were analyzed using the 1PL, 2PL, 3PL, and 3PL with FLA models, except for the science exam, because the number of examinees were below 100.

Research Results

Results of MEU

The MEU data came from 2000 examinees and had 65 MCQs that presented 4 options. The results of the MEU data are as follows.

Table 2
Goodness of Fit for MEU

Criterion	1PL	2PL	3PL	3PL with FLA
LL*	−67511	−66697	−66582	−66926
-2LL	135022	133394	133164	133850
AIC	135152	133654	133554	134110
BIC	135516	134382	134646	134838

*LL = Log Likelihood

Table 2 summarizes the fitting of the four models for the MEU data according to the LL, AIC, and BIC values. Although the goodness of fit of the 3PL model was better than the other models by −2LL and AIC, the fitting of the 2PL model was better than that of the other models by BIC. Given that BIC is more advanced than AIC, 2PL had the best fitting to the data. Table 3 shows the parameters.

Table 3
Item parameter by the 3PL model in MEU

Item	Discrimination	Difficulty	Item	Discrimination	Difficulty
1	2.1132	−.6976	34	.9820	.4315
2	1.8535	−.8335	35	1.7461	−1.6727
3	1.6916	−.1999	36	1.5949	−.5518
4	1.3193	.3296	37	1.2554	−.5831
5	1.6350	−.7062	38	1.7614	−.2562
6	1.0479	−.4203	39	.7282	−1.0174
7	.9722	.2679	40	.9490	−1.1574
8	1.6024	−1.2724	41	1.4250	−.2407
9	2.022	−1.1362	42	.7850	−1.1540
10	1.2668	.2788	43	1.5006	−1.2734
11	1.4010	−.2582	44	1.2423	.5989
12	1.2856	.7626	45	1.8518	−1.3875
13	1.3650	−.1249	46	1.3387	−.5634
14	1.4639	−.7244	47	.9081	−.1467
15	1.2111	−.5657	48	1.2582	−1.8985
16	1.7974	−1.6913	49	1.4789	−1.7205
17	1.600	−.8429	50	1.1996	−.8432
18	1.3297	−.4830	51	1.3667	.1572
19	1.5312	−.5796	52	1.0683	.6201
20	1.3853	−1.1751	53	.3278	−1.4600
21	1.0074	−.3351	54	1.0363	−.6617
22	2.3669	−.9952	55	1.5048	−1.5392
23	2.0074	−.1253	56	.6290	−1.3656
24	1.7307	−1.0988	57	.9458	−1.5654
25	1.3027	−1.5379	58	1.6192	−.5214
26	1.1391	−.2850	59	1.1100	.5149
27	2.1303	−.8613	60	.8932	−.2721
28	1.2974	−1.5938	61	1.1553	−.2585
29	1.2892	−.7045	62	1.4740	−1.3182
30	1.6441	−.2325	63	1.2890	−.8158
31	.8098	−.3194	64	1.2960	.4233
32	.7927	−.6782	65	.9902	−1.1348
33	.8030	−.1241			

The item parameters of MEU by the 2PL model include difficulty and discrimination parameters. The difficulty parameters are from −1.89853 to .7626612, and 10 difficulty parameters are positive. The discrimination parameters are from .3278112 to 2.36695, and all the values are positive. The discrimination parameters of $-\infty < a_i \leq 0$, $0 < a_i \leq .3$, $.3 < a_i \leq +\infty$ correspond to 0, 0, and 65 items, respectively. Table 4 shows the acceptability in MEU.

Table 4
Acceptability in MEU

	Discrimination			Difficulty	
Value	$-\infty < a_i \leq 0$	$0 < a_i \leq 0.3$	$0.3 < a_i \leq +\infty$	$-2.4 < b_i \leq 1.2$	$b_i \leq -2.4$ *or* $b_i \leq 1.2$
Advice	Unacceptable	Recommend Delete	Acceptable	Acceptable	Unacceptable
Number	0	0	65	65	0

Table 4 illustrates that the quality of this exam is perfect, and all items should be accepted. This rating can be related to the skills of examiners. The MEU exam is an evaluation test, which has been designed by teachers who have passed the MCQ design training.

Results of MEC

MEC data were obtained from 734 examinees and contained 12 MCQs that had 4 options. The result of the MEC data is as follows.

Table 5
Goodness of Fit for MEC

Criterion	1PL	2PL	3PL	3PL with FLA
LL*	−2498	−2477	−2477	−2484
-2LL	4996	4954	4954	4968
AIC	5020	5002	5026	5016
BIC	5075	5112	5191	5126

*LL = Log Likelihood

Table 5 summarizes the fitting of the four models for the MEC data according to LL, AIC, and BIC values. The fitting of the 2PL model was better than that of the other models by BIC, AIC, and −2LL. Hence, 2PL had the best fit to the data. Table 6 presents the parameters. Item 4 has unacceptable discrimination and difficulty parameters.

Table 6
Item parameter by 3PL model in MEU

Item	Discrimination	Difficulty	Item	Discrimination	Difficulty
1	1.3236	-4.1386	7	1.9850	-2.2485
2	.9038	-4.9618	8	2.2702	-1.8790
3	.1759	-18.4270	9	1.3433	-1.7954
4	-.1614	28.0032	10	.9451	-2.1729
5	.8311	-2.3700	11	.8603	-1.3057
6	1.0681	-2.9202	12	.9548	-.5241

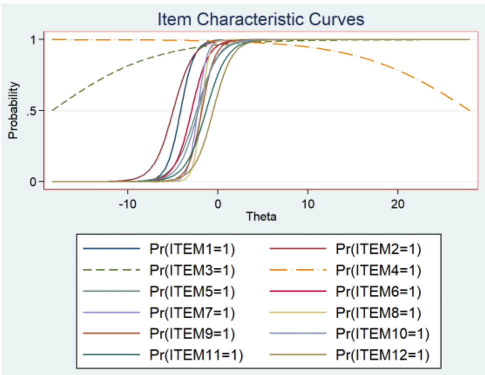
The item parameters of MEU by 2PL model in Table 6 include difficulty and discrimination parameters. The difficulty parameters are from −18.42705 to 28.0032, and one difficulty parameter is positive. The discrimination parameters are from −.1614428 to 2.270278, and one value is negative. The discrimination parameters, $-\infty < a_i \leq 0$, $0 < a_i \leq .3$, $.3 < a_i \leq +\infty$ correspond to 1, 1, and 10 items, respectively. Seven difficulty parameters of the items were acceptable. Only one item (item 4, shading) had unacceptable discrimination and difficulty. Acceptability in MEC is shown in Table 7, and the ICCs are shown in Figure 1.

Table 7
Acceptability in MEU

	Discrimination			Difficulty	
Value	$-\infty < a_i \leq 0$	$0 < a_i \leq 0.3$	$0.3 < a_i \leq +\infty$	$-2.4 < b_i \leq 1.2$	
Advice Number	Unacceptable 1	Recommend Delete 1	Acceptable 10	Acceptable 7	Unacceptable 4+1*

*4+1 means their 4 difficulty parameters of items small than -2.4, and 1 difficulty parameters larger than 1.2

Figure 1
ICCs of 2PL model in MEU



Negative discrimination for one item means that examinees with high abilities may not easily answer this item correctly, whereas examinees with low abilities may easily reply with the correct answer. This result is against the original intention of the exam. Figure 1 shows that the discrimination parameter of item 4 is negative. Thus, this item should be rejected. The discrimination of item 3 is .1759142, and the S-shape shown in Figure 1 is nearly lost. Hence, the recommendation is to delete the item owing to low discrimination. This action will not bring a significant increase in the probability of answering correctly when the significant change in the ability exists. This exam was written by a middle school teacher who was not trained in MCQ item writing. This background may be the reason for the low-quality test questions. The first four items were simple. Overall, the difficulty of the test gradually increased, which was consistent with the original intention of the MCQ design.

Results of HEU

The HEU data came from 96 examinees and used 65 MCQs that had 4 options. The results of the HEU data is as follows.

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 78, No. 4, 2020

Table 8
Goodness of Fit for the HEU

Criterion	1PL	2PL	3PL	3PL with FLA
LL*	-3095	-2946	Non-convergence	-2952
-2LL	6190	5892		5904
AIC	6320	6152		6164
BIC	6487	6485		6497

*LL=Log Likelihood

Table 8 summarizes the fitting of the four models for the HEU data according to the LL, AIC, and BIC values. The fitting of the 2PL model was better than that of the other models by BIC, AIC, and −2LL. Hence, 2PL had the best fit to the data. Table 9 shows the parameters. Items 8, 15, 23, 24, 32, and 65 have bad discrimination and difficulty parameters.

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 78, No. 4, 2020

Table 9
Item parameter by 2PL model in HEU

Item	Discrimination	Difficulty	Item	Discrimination	Difficulty
1	.5325	-3.4995	34	1.2987	-.4140
2	.3689	-3.5522	35	.9570	-1.8847
3	1.7733	-1.7545	36	.5510	-.3307
4	1.2311	-.7041	37	.3245	-1.2012
5	1.0981	-2.0392	38	.3920	-4.8675
6	.9033	-2.8422	39	.3082	-1.6966
7	.6651	-4.0879	40	1.4462	-1.1166
8	.1855	-7.2503	41	.5298	.3312
9	-.3739	.6922	42	1.3359	-.1507
10	.9818	-2.4201	43	1.7783	-1.4215
11	1.2173	-1.9038	44	.7474	-2.8634
12	1.5659	-1.6429	45	.7637	.1760
13	1.2349	-1.0009	46	.7958	-2.4814
14	2.8825	-.2132	47	.8398	-1.4324
15	-.3252	-3.6347	48	-.4064	-1.4231
16	.4889	-2.8691	49	1.4073	-1.8194
17	.6150	.1407	50	.2269	-4.4155
18	2.1514	.1039	51	1.1072	-1.4926
19	2.3433	-1.5597	52	1.2687	-.7880
20	.7637	-1.5410	53	1.4141	-.3569
21	2.9663	-1.7229	54	1.1807	-.6233
22	.1867	-.9029	55	1.7712	-.4076
23	-.1175	11.3908	56	2.2198	-1.9118
24	-.4200	3.1461	57	1.8666	-2.1910
25	-.4173	-.2116	58	1.3369	-2.3793
26	1.1451	-3.1916	59	2.0738	-1.9696
27	.8500	.3326	60	1.0608	-1.5356
28	.3131	.9587	61	.9791	-1.4132
29	2.1511	-1.1229	62	1.5173	-.4290
30	1.7528	-.5243	63	1.2799	-.5064
31	.4444	-6.2792	64	.8829	-.3466
32	.0085	-81.2328	65	-.1481	6.0183
33	.6749	-2.8436			

The item parameters of HEU in the 2PL model included difficulty and discrimination parameters. The discrimination parameters ranged from −.420011 to 2.966365, and seven values were negative. The discrimination parameters, $-\infty < a_i \leq 0$, $0 < a_i \leq 0.3$, $0.3 < a_i \leq +\infty$ correspond to 7, 3, and 55 items, respectively. The difficulty parameters ranged from −81.23288 to 11.39088, 10 difficulty parameters were positive, and 3 were above 1.2. Hence, 45 difficulty parameters of the items were acceptable. Table 10 shows the parameter acceptability.

ISSN 1822-7864 (Print) ISSN 2538-7111 (Online) https://doi.org/10.33225/pec/20.78.576

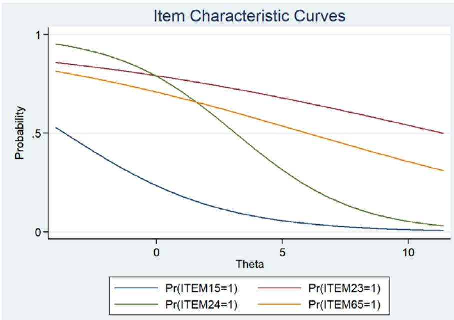
Table 10
Acceptability in HEU

	Discrimination			Difficulty	
Value	$-\infty < a_i \leq 0$	$0 < a_i \leq 0.3$	$0.3 < a_i \leq +\infty$	$-2.4 < b_i \leq 1.2$	
Advice	Unacceptable	Recommend Delete	Acceptable	Acceptable	Unacceptable
Number	7	3	55	45	17+3*

*17+3 means their 17 difficulty parameters of items small than -2.4, and 3 difficulty parameters larger than 1.2

Several items (i.e., 15, 23, 24, and 65) had unacceptable discrimination and difficulty parameters. Figure 2 shows that ICC of the items was decreasing, and all four items were extremely difficult.

Figure 2
ICCs of 2PL model in HEU



The discrimination parameters of items 15, 23, and 24 were negative, and the discrimination parameters of items 8, 22, and 32 were considerably small. The probability of a correct item response will become a straight line. A total of 20 items had difficulty parameters that were considerably large or small. A total of 4 items must be deleted, 19 items should be deleted, and 1 item is recommended to be deleted. This result accounted for 36.9% of the total items. Hence, the examination quality of the science freshman course was not ideal. The teacher who wrote this exam was not trained in MCQ writing.

Result of HEC

HEC is the final examination of Principles of Pedagogy administered in the provincial universities of China. The data came from 1008 examinees and the 50 MCQs had 4 options. Table 11 shows the result of the goodness of fit.

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 78, No. 4, 2020

Table 11
Goodness of Fit for the HEC

Criterion	1PL	2PL	3PL	3PL with FLA
LL*	-23363	-23288	-23273	-23553
-2LL	46726	46576	46546	47105
AIC	46826	46776	46846	47506
BIC	47072	47268	47583	47797

*LL=Log Likelihood

Table 11 summarizes the fitting of the four models for the HEC data according to the LL, AIC, and BIC values. The fitting of the 2PL model was better than that of the other models by BIC, AIC, and −2LL. Hence, 2PL had the best fit to the data. Table 12 shows the parameters.

Table 12
Item parameter by 2PL model in HEC

Item	Discrimination	Difficulty	Item	Discrimination	Difficulty
1	.6455	-2.4457	26	.8347	-.9127
2	1.2045	-1.6687	27	.7673	-2.7523
3	1.1753	-2.1569	28	1.1248	-1.4340
4	.8745	-1.9263	29	.9641	-2.0333
5	.9850	-1.7998	30	1.1592	-2.1377
6	.5271	-2.6933	31	1.1554	-1.2114
7	.9292	-1.7485	32	.7915	-1.2184
8	.9584	-1.4585	33	1.0578	-1.8038
9	.9954	-1.6352	34	1.0022	-1.7915
10	1.0638	-1.5479	35	.7843	-1.7383
11	.7309	-1.8956	36	1.1337	-1.2625
12	.7263	-1.6809	37	.8840	-1.8274
13	.7071	-2.1333	38	.8183	-1.4568
14	.8385	-1.3122	39	.6962	-2.4359
15	1.2246	-1.5387	40	1.1790	-1.7167
16	.7317	-2.5494	41	1.2298	-1.2589
17	.8780	-1.8374	42	1.2594	-1.3773
18	.8611	-1.8749	43	1.2097	-1.0256
19	.7361	-2.5030	44	1.3262	-1.4568
20	1.1752	-1.6403	45	.8642	-1.1404
21	.8700	-1.0667	46	.7598	-2.1071
22	1.1631	-1.6598	47	.9691	-1.7370
23	.8539	-2.0611	48	.8393	-1.7279
24	.4740	-1.4296	49	1.1735	-1.6492
25	.9604	-1.4494	50	.8920	-1.6899

The item parameters of HEC by 2PL model included difficulty and discrimination parameters. The discrimination parameters ranged from .4740901 to 1.326276, and no value was negative. All discrimination parameters were in $.3 < a_i \leq +\infty$. The difficulty parameters ranged from −2.752367 to −.9127278, and not one difficulty parameter was positive. Table 13 shows the parameter acceptability.

Table 13
Acceptability in HEU

	Discrimination			Difficulty	
Value	$-\infty < a_i \leq 0$	$0 < a_i \leq 0.3$	$0.3 < a_i \leq +\infty$	$-2.4 < b_i \leq 1.2$	
Advice	Unacceptable	Recommend Delete	Acceptable	Acceptable	Unacceptable
Number	0	0	50	44	6+0*

Note: *6+0 means their 6 difficulty parameters of items small than -2.4, and no difficulty parameters larger than 1.2

Table 13 illustrates that all discrimination of the items was acceptable. The 44 difficulty parameters of the items acceptable indicated 67% of the total. Only 6 difficulty parameters of the items were below −2.4, and the smallest one was −2.752367. All the discriminate parameters were acceptable. No items had bad different and difficulty parameters together. Hence, the exam quality was very good. This exam was written by an Education teacher, who is undergoing training in MCQ writing.

Discussion

This research showed that whether the examinees choose the guessing method was not related to the question or exam type. One of the disadvantages of MCQs is that students who do not know the correct answer may arrive at the correct one by guessing. However, this research showed that examinees may not choose to guess in an MCQ exam. This result is consistent with that in a previous research (Brown & Abdulnabi, 2017). A logical approach is to analyze items using a statistical model capable of detecting the effects of chance performance (Brown & Abdulnabi, 2017). The 2PL model, which has no guessing parameter, consistently has the best fit for MEU, MEC, HEU, and HEC. This result may be attributed to the four exams being general, and not competitive. The basic purpose of the four exams is to test whether the students have learned specified knowledge. Hence, the scope of knowledge in the exam is clear to the examinees. Thus, they may have prepared for the exam well, so that no guessing is needed for them.

Training teachers in MCQ writing was necessary because of the unideal quality of MEC and HEU. The other two data sets were better. On the one hand, the result showed that propositional techniques between higher and secondary education did not necessarily have differences. On the other hand, propositional techniques between China and the US were not necessarily different. The writer of MEU is a professional proposition technician, thereby explaining the perfect quality of the exam. The HEC writer is an education teacher. Although she is not a professional proposition technician, she receives good education training and is being trained when she was writing the exam. Her background could explain why the quality of HEC is good. Flawed MCQ items may result in misleading insights into student performances and contaminate important decisions. Hence, proposition teachers must receive the relevant proposition technical training. Unfortunately, the majority of Chinese and American teachers are untrained.

The low-quality examination found by research and analysis had shown the need to evaluate the quality of test questions before the examination. At this point, schools in China and the US should evaluate the quality of their exams. Moreover, the quality evaluation of mid-term examination questions could effectively improve the quality of the final examination (Brown & Abdulnabi, 2017). Subsequently, the effect of the evaluation of students' level could be improved. In China, schools seldom conduct quality evaluation of test questions, except for the national examination, based on statistical methods. The current research showed that the evaluation of test questions based on proposition rules was not reliable. Tarrant et al. (2006) evaluated 2,770 MCQs used over five years (from 2001 to 2005) and concluded that nearly half (46%) of the items were bad because of violation of item-writing guidelines. The quality of HEU showed that the items of the exam should be assessed as well.

In the whole world, MCQs are commonly used in secondary and higher education. Compared to the previous research, in this research the data of middle and higher education examinations from different countries were analyzed to give a more general description of MCQs. There are some low quality MCQs in both China and the US exams. The proposition teachers who are trained by MCQ writing can write high quality items whatever they are in middle school or higher school. So, the items of the education exam should be assessed, and proposition teachers should receive the relevant proposition technical training in statistics. In particular, the MCQs quality analysis method based on IRT should be one of the main contents of training.

The quality of MCQs was assessed by IRT. Four different models were used to fit exam data: 1PL 2PL 3PL 3PL with FLA. Four evaluation criteria for model selection were used in the real data: LL, −2LL, AIC, BIC. The feature of the most fitting model was used to interpret MCQs exams. All the data consistently has the best fit to 2PL model, which has no guessing parameter. That means teachers do not need to worry about students may guess in MCQs. Oppositely, in most MCQs exams, students resolve the items with their abilities. The limitation of this research is the use of simple models.

Conclusions

The research results showed the quality of MCQs was not ideal in some exams. It is necessary to eliminate the low-quality MCQs before the test. The effective way to improve MCQ quality is to train teachers in MCQ writing. Proposition teachers must receive the relevant proposition technical training. This can improve the accuracy of student assessment. The research results also showed the examinees could choose the correct answers without choosing a guessing method, even if they had a chance to guess. The future research about student's response pattern can be carried out by other researchers related to the results of this study. Complex IRT models can be used to analyze and explore the answer patterns of students under different examinations.

Acknowledgements

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was partially supported by the National Education Science of China (Grants CHA190269). Special thanks to fund team member Hongwei Wang and her friend Ph.D. Fernando G. Quintana for their help.

References

Akour, M., & AL-Omari, H. (2013). Empirical investigation of the stability of IRT item-parameters estimation. *International Online Journal of Educational Sciences, 5*(2). https://eis.hu.edu.jo/deanshipfiles/pub106314725.pdf

Bailey, P. H., Mossey, S., Moroso, S., Cloutier, J. D., & Love, A. (2012). Implications of multiple-choice testing in nursing education. *Nurse Education Today, 32*(6), e40-e44. https://doi.org/10.1016/j.nedt.2011.09.011

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed., revised and expanded). New York: Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examiner's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences*. Springer Science & Business Media.

Borsboom, D. (2005). Measuring the mind: *Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Burton*, R. F. (2005). Multiple-choice and true/false tests: myths and misapprehensions. *Assessment & Evaluation in Higher Education, 30*(1), 65-72. https://doi.org/10.1080/0260293042003243904

Brady, A. M. (2005). Assessment of learning with multiple-choice questions. *Nurse Education in Practice, 5*(4), 238-242. https://doi.org/10.1016/j.nepr.2004.12.005

Considine, J., Botti, M., and Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian, 12,* 19-24. http://dx.doi.org/10.1016/S1322-7696(08)60478-3

Brady, A. M. (2005). Assessment of learning with multiple-choice questions. *Nurse Education in Practice, 5*(4), 238-242. https://doi.org/10.1016/j.nepr.2004.12.005

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning, 2*(2), 4. https://files.eric.ed.gov/fulltext/EJ985723.pdf

Ding, L., and Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics–Physics Education Research, 5,* 020103. http://dx.doi.org/10.1103/PhysRevSTPER.5.020103

Downing, S. M. (2005). The effects of violating standard item-writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in health sciences education, 10(2), 133-143.* http://dx.doi.org/10.1007/s10459-004-4019-5

Drasgow, F, and Parsons, C K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement,* 7, 189-199. https://doi.org/10.1177/014662168300700207

Ebel, R. L., & Frisbie, D. A. (1972). Essentials of educational measurement (pp. 492-494). Prentice-Hall.

Ellsworth, R. A., Dunnell, P., & Duell, O. K. (1990). Multiple-choice test items: What are textbook authors telling teachers? *The Journal of Educational Research*, *83*(5), 289-293. https://doi.org/10.1080/00220671.1990.10885972

Embretson, S. E., & Reise, S. P. (2013). *Item response theory.* Psychology Press.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357–381. https://doi.org/10.1177/0013164498058003001

Finch, H., & Monahan, P. (2008). A bootstrap generalization of modified parallel analysis for IRT dimensionality assessment. *Applied Measurement in Education, 21*(2), 119-140. https://doi.org/10.1080/08957340801926102

Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice*, *7*(2), 33-38. https://doi.org/10.1111/j.1745-3992.1988.tb00434.x

Gavin T. L. Brown, and Hasan H. A. Abdulnabi. (2017). evaluating the Quality of higher education instructor-constructed Multiple-choice Tests: Impact on student grades. In *Frontiers in Education* (Vol. 2, p. 24). Frontiers. doi.org/10.3389/feduc.2017.00024

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Routledge.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Han, K. T. (2012). Fixing the c parameter in the three-parameter logistic model. *Practical Assessment, Research, and Evaluation*, *17*(Article 1). https://doi.org/10.7275/f0gz-kc87

Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing test banks. *Journal of Education for Business, 73*(2), 94-97. https://doi.org/10.1080/08832329709601623

Holland, P. W. (1990). The Dutch identity: A new tool for the study of item response models. *Psychometrika, 55*(1), 5-18. https://link.springer.com/content/pdf/10.1007/BF02294739.pdf

Houts, C. R., Edwards, M. C., Wirth, R. J., & Deal, L. S. (2016). A review of empirical research related to the use of small quantitative samples in clinical outcome scale development. *Quality of Life Research, 25*(11), 2685-2691. https://link.springer.com/content/pdf/10.1007/s11136-016-1364-9.pdf

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic Medicine*, *77*(2), 156-161. https://doi.org/10.1097/00001888-200202000-00016

Kehoe, J. (1994). Basic item analysis for multiple-choice tests. *Practical Assessment, Research, and Evaluation, 4*(1), 10. https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1053&context=pare

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* IAP.

Masters, J. C., Hulsmeyer, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying textbooks used in nursing education. *Journal of Nursing Education, 40*(1), 25-32. https://doi.org/10.3928/0148-4834-20010101-07

Nedeau-Cayo, R., Laughlin, D., Rus, L., & Hall, J. (2013). Assessment of item-writing flaws in multiple-choice questions. *Journal for Nurses in Professional Development, 29*(2), 52-57. https://doi.org/doi:10.1097/NND.0b013e318286c2f1

Pelton, T. W. (2002). *The accuracy of unidimensional measurement models in the presence of deviations from the underlying assumptions* (Doctoral dissertation). Brigham Young University Department of Instructional Psychology and Technology.

Purchase, H., Hamer, J., Denny, P., & Luxton-Reilly, A. (2010, January). The quality of a PeerWise MCQ repository. In *Proceedings of the Twelfth Australasian Conference on Computing Education-Volume 103* (pp. 137-146). http://citeseerx.ist.psu.edu/viewdoc/download?doi=https://doi.org/10.1.1.664.5139&rep=rep1&type=pdf

Stagnaro-Green, A. S., & Downing, S. M. (2006). Use of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education. *Medical Teacher, 28*(6), 566-568. https://doi.org/10.1080/01421590600711153

Stevenson, H. W., Lee, S. Y., Chen, C., Lummis, M., Stigler, J., Fan, L., & Ge, F. (1990). Mathematics achievement of children in China and the United States. *Child Development, 61*(4), 1053-1066. https://doi.org/10.1111/j.1467-8624.1990.tb02841.x

Stevenson, H., & Stigler, J. W. (1994). *Learning gap: Why our schools are failing and what we can learn from Japanese and Chinese education*. Simon and Schuster.

Martín, E. S., Del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement, 30*(3), 183-203. https://doi.org/10.1177/0146621605282773

Secolsky, C., & Denison, D. B. (Eds.). (2012). *Handbook on measurement, assessment, and evaluation in higher education*. Routledge.

Su, W. M., Osisek, P. J., Montgomery, C., & Pellar, S. (2009). *Designing multiple-choice test items at higher cognitive levels. Nurse Educator, 34*(5), 223-227. http://dx.doi.org/10.1097/NNE.0b013e3181b2b546

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today, 26*(8), 662-671. http://dx.doi.org/10.1016/j.nepr.2006.07.002

Thorndike, R. M., & Thorndike-Christ, T. M. (2010). *Measurement and evaluation in psychology and education.* Pearson.

Walsh, C. M., & Seldomridge, L. A. (2006). Critical thinking: Back to square two. *Journal of Nursing Education, 45*(6), 212-219.

van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review, 118*(2), 339. https://doi.org/10.1037/a0022749

Yudkowsky, R., Park, Y. S., & Downing, S. M. (Eds.). (2019). *Assessment in health professions education*. Routledge.

Zhu, Z., Wang, C., & Tao, J. (2019). A Two-Parameter Logistic Extension Model: An efficient variant of the Three-Parameter Logistic Model. *Applied Psychological Measurement, 43*(6), 449-463. https://doi.org/10.1177/0146621618800273

Received: *April 25, 2020*　　　　Accepted: *July 20, 2020*

Cite as: Jia, B., He, D., & Zhu, Z. (2020). Quality and feature of multiple-choice questions in education. *Problems of Education in the 21st Century*, *78*(4), 576-594. https://doi.org/10.33225/pec/20.78.576

Bing Jia	Assistant Professor, Teacher Education and Training Center, Beihua University, Jilin, China. E-mail: jia.bing@foxmail.com
Dan He	Associate Professor, School of Education Science, Beihua University, Jilin, China. E-mail: 15337614@qq.com
Zhemin Zhu *(Corresponding author)*	PhD, Associate Professor, School of Education Science, Beihua University, Jilin, China. E-mail: zhuzm485@nenu.edu.cn