

Item Analysis: Concept and Application

Assad Ali Rezigalla

Abstract

In the field of medical education, Item analysis is a statistical analysis of student's responses on exam items and the relationship between them. It provides constructive feedback about items quality, increases the effectiveness of the exam, and supports exam validity and reliability. The decision of adding or removing an item should depend mainly on the result of the item analysis. The feedback of item analysis can support modification of instruction methods. This chapter provides a comprehensive review of item analysis (psychometric analysis) and also can be used as methodological guidance to perform an informative analysis. The chapter discusses item analysis under the following headings, importance of item analysis, factors affecting item analysis, parameters of item analysis and application. The parameters of item analysis include the indices of the internal consistency, difficulty, discrimination, and distractor analysis.

Keywords: item analysis, difficulty index, reliability, discrimination index, KR20

1. Introduction

“Assessment is a central component of the teaching and learning” process. It is defined as “the systematic collection and analysis of information to improve student learning” [1]. Test (exam) is a part of student assessment and should be “An objective and standardized measure of a sample of behavior” [2]. Item analysis is a post-examination evaluation and can provide information about the quality of tests.

Item analysis is a statistical analysis of the student's responses on a test. Collection and summarization of students' responses can provide quantitative objective information that is useful in deciding the quality of the test items and increasing the assessment's efficiency [3, 4]. Also, Item analysis “investigates the performance of items considered individually either in relation to some external criterion or the remaining items on the test” [5].

2. Importance of psychometric analysis

Any educational test should measure students' achievement in content material. Also, it leads to an overall assessment of students' development to decide their academic status [6, 7].

The importance of item analysis is determined by the objective of the assessment [8]. In summative assessment, the assessment results should be reliable and valid because incorrect decisions about the academic status will lead to negative

consequences [9]. While for the formative evaluation where the target is students learning, the item analysis has no much importance in giving feedback about items construction to test composers.

In literature, many reasons were reported for the conduction of item analysis, including examining if the item is functioning as intended, did it assess the required concepts (content)?, did it discriminate between those who master the content material and those who were not? was it within the acceptable level of difficulty?, whether the distracters are functioning or not? [10, 11].

3. Factors affecting item analysis

Many factors can affect item analysis and hence its interpretation [8]. Difficulty and discrimination indices were constantly changing per administration and influenced by the ability and number of the examinee, the number of items, and the quality of instructions [8, 12].

Whatever the exam or test blueprinting (item selection) method, exam items remain a sample of the needed content material. The number of items (item sampling) carries excellent importance because one cannot ask about all contents. With a too-small number of items, the results may not be enough to reflect true student ability [8]. Technical item flaws are divided into two major types, test wiseness, and irrelevant difficulty. Test wiseness flaws can result in more easy items. Faults related to irrelevant difficulty can result in more challenging items unrelated to the content under assessment. It was reported that item analysis of exam with 200 examinees is stable, and with fewer than 100 examinees should be interpreted with caution (item difficulty or item discrimination index). While Downing and Yudkowsky described that even for a small number of the examinee (e.g., 30) still, the item analysis can provide a piece of a helpful information to improve item [13, 14].

4. Parameters of item analysis

The item or psychometric analysis parameters include difficulty index, reliability, discrimination index, distractor efficiency [2]. The descriptive statistics of the exam are important and can provide helpful generalized information [2]. The descriptive statistics include scores frequency, the mean, the mode, the median, and the standard deviation.

5. Cronbach's alpha (Index of reliability)

Cronbach's alpha (KR20) is widely accepted and used estimate of test reliability (the internal consistency) and reported to be superior to the split-half estimate [15, 16]. Although validity and reliability are closely associated, the reliability of an assessment does not depend on its validity [16, 17]. Coefficient alpha is known to be equal to Kr-20 if the item has a single answer, such as in the case of type A MCQs or binary [18–21].

Coefficient alpha reflects the degree to which item response scores correlate with total test scores [15]. It also describes the degree to which items in the exam measure the same concept or construct [22]. Therefore, it is connected to the inter-relatedness and dimensionality of the items within the exam [16, 20]. Cronbach's alpha is affected by exam time, the number and inter-relation of the items (dimensionality) and easy or hard, poorly written or confusing items, Variations in examinee

responses, curriculum content not reflected in the test, Testing conditions, and Errors in recording or scoring [22–24]. The value of alpha is decreased in the exam with fewer items and increased if items assessing the same concept (unidimensionality of the exam) [16]. Other factors were reported to impact alpha value, such as item difficulty, number of the examinee, and student performance in the exam time. It was argued that very high alpha values could indicate lengthy exams, parallel items, or a narrow coverage of the content material [22].

The alpha value of the exam can be increased by increasing the number of items with a high p-value (difficulty index). It was reported that items with moderate difficulty could maximize alpha value and while those with zero difficulties or 100 can minimize it [15]. In the same way, deletion of faulty items can increase alpha value. It should be considered that repetition of items in the same exam or using items assessing the same concept can increase alpha value.

6. Interpretation of Cronbach's alpha

The interpretation of reliability is the correlation of the test with itself. When the estimate of reliability increases, the portion of a test score related to the error will decrease. Wise interpretation of alpha needs an understanding of the inter-relatedness of items and whether the items measure a single latent trait or construct. Exam or test with different content materials such as integrated courses, for example, in the musculoskeletal system course, although is dominated by anatomy it contains other subjects of basic medical and clinical sciences that have different contents. Therefore, interpretation of such a course exam needs deep looks beyond the alpha figure. It was reported that KR20 of 0.7 is acceptable to short test (less than 50 items) and KR20 of 0.8 for an extended test (more than 50 item-test) [25]. Moreover, it was documented that a multidimensional exam does not have a lower (Table 1) alpha value than a unidimensional one [30].

A low alpha value can be due to a smaller number of items, reduced interrelatedness between items, or heterogeneous constructs [22]. A high value of alpha can suggest exam reliability, and some items are non-functional as they are testing the same content but in a different guise or repeated ones [16, 22]. Also, a high value indicates items with high interrelatedness, indicating a limited coverage of the content materials [22].

7. Improving Cronbach's alpha

Adding new items with an acceptable difficulty index, high discrimination power and distractor efficiency can increase the test reliability [22, 31, 32]. In addition, deletion of faulty items or those with low or very high p-value can improve Cronbach's alpha. Items with poor correlation or are not related should be revised or discarded from the exam.

8. Distractor analysis

Commonly are formed of a stem with or without leading question and five or four alternatives (type A MCQs). Among item's alternatives, only one is the key answer and others are called distractors [4]. Distractors should carry or convey a miss concept about the key answer and appear plausible. The distractors should appear similar to the key answer in terms of the used words, grammatical form,

Author	Interpretation of Cronbach's alpha (KR20)
Robinson, Shaver et al. [26]	≥0.80 Exemplary
	0.70–0.79 Extensive
	0.60–0.69 Moderate
	<0.60 Minimal
Cicchetti [27]	<0.70 Unacceptable
	0.70–0.80 Fair
	0.80–0.90 Good
	< 0.90 Excellent
Axelson and Kreiter [28]	>0.90 is needed for very high stakes tests (e.g., licensure, certification exams)
	0.80–0.89 is acceptable for moderate stakes tests (e.g., end-of-year summative exams in medical school, end-of-course exams)
	0.70–0.79 would be acceptable for lower stakes assessments (e.g., formative or summative classroom-type assessments created and administered by local faculty)
	<0.70 might be useful as one component of an overall composite score.
Obon and Rey [12]	>0.90 Excellent reliability
	0.80–0.90 Very good for a classroom test
	0.70–0.80 good for a classroom test
	0.60–0.70 Somewhat low (The test needs to be supplemented by other measure)
	0.50–0.60 Suggests need for revision of test (unless it is quite short, ten or fewer Items).
	0.50 < Questionable reliability.
Hassan and Hod [29]	> 0.7 is excellent
	0.6–0.7 is acceptable
	–0.5–0.6 is poor
	< 0.5 is unacceptable
	< 0.30 is unreliable

Table 1. Reference values and interpretation of Cronbach's alpha (KR20).

style, and length [19]. Distractor efficiency (DE) is the ability of incorrect answers to distract the students [12].

A functional distractor (FD) the distractor that is selected by 5% or more of the examinee [4, 33]. At the same time, those chosen by less than 5% of the examinee are considered non-functional (NFD) [4]. In comparison, other authors reported 1% of the examinee as the demarcation of functional distractors [34, 35]. Commonly items are categorized based on the numbers of NFDs in the item (**Table 2**) [12, 26, 36, 37].

The occurrence of NFD makes the item easier and reduces its discrimination power, while FD distractors are making it more difficult [36, 38]. It was reported that non-functional distractors are negatively correlating with reliability [38]. The presence of non-functional distractors can be related to two main causes. First is the training and construction ability of the item writer or composer. Second, the miss-match between the target content and the possible number of a distractor created. Thus, training and more effort in item writing and construction can decrease NFDs [36]. Other causes were related to NFDs, including the low

cognitive level of the item, irrelevant or limited number of plausible distractors, or presence of logic cues [39]. Another possibility of NFDs is mastering the content material of the item, and students can identify the distractor as the wrong one. If no other cause (s) for NFDs, they should be removed or changed with a more plausible option because it has no contribution to the measurement of the test [12]. If a distractor is selected more frequently than the key answer by a higher-scoring examinee, this may indicate poor constructions or a misleading question or miss or double-keyed [12, 40]. In this, concerning the use of three options is more practical than four, does not affect reliability, and does not affect the discrimination index significantly [26, 35–37].

Furthermore, it was reported that there is no psychometric reason that all items in the exam should have the same number of distractors [26, 41]. The required number of options in an item should be considered according to the content material from which plausible distractors can be developed [33, 40, 42].

Reducing the number of options/distractors will result in other important benefits such as reducing the answering time of the test and safe time can be used to cover more content material, reduce the burden on item composers, and have items with more acceptable parameters [43, 44].

Puthiaparampil et al. reported a non-high significant negative and positive correlation between the number of functional distractors and difficulty and discrimination indices, respectively [34]. While a significant positive correlation was reported between the DIF and the number of NFDs [45].

Many authors concluded that no predictable relationship between DE and difficulty index and discrimination index [26, 31, 40, 46, 47]. In addition Licon-Chávez et al. did not find a parallel performance between DE and other parameters of item analysis including Cronbach alpha [46]. In contrast, some authors claimed that low DE decreases the difficulty index [47, 48].

9. Improving distractor analysis

Restoring the optimal DE of the item can be achieved by identifying flaws related to the NFDs and correcting them or removing the NFDs from the item [39].

Number of NFD	Percentage	Interpritation
3	0	Poor
2	33.3	Moderate
1	66.6	Good
0	100	Excellent

Table 2.

Classification of items according to thee number of the nonfunctional distractors (NFD).

10. Difficulty index

The item difficulty (easiness, facility index, P-value) is the percentage of students who answered an item correctly [6, 40]. The difficulty index ranges from 0 to 100, whereas the higher the values indicate, the easier the question and the

Number of options	The ideal (optimal) difficulty level	
	[46, 47]	[24]
2	—	0.75
3	0.77	0.67
4	0.74	0.63
5	0.70	0.60

Table 3.
The ideal (optimal) difficulty level (for tests with 100 items).

low value represents the difficulty of hard items. The ideal (optimal) difficulty levels for type A MCQs is varying according to the number of the options (Table 3) [49, 50]. The range of items difficulty can be categorized into difficult, moderate, and easy. Easy and difficult items were reported to have very little discrimination power [48]. Item difficulty is related to the item and the examinee that took the test in the given time [24]. Thus, reusing of the item depending on its difficulty index should be controlled. Some authors found that difficulty indices of items assessing high cognitive levels in Bloom’s taxonomy such as evaluation, explanation, analysis, and synthesis are lower than those assessing remembering, understanding, and applying [51, 52].

During item or exam construction, the constructor should aim for acceptably level of difficulty [6]. Sugianto reported that items within the exam could be distributed according to difficulty to moderate level (40%), easy and challenging levels (20%), and easier and more challenging levels (10%) [6]. Other authors reported that most items should be of moderate difficulty or 5% should be in the difficult range [50, 53]. Some authors found that difficulty indices of items assessing high cognitive levels in Bloom’s taxonomy such as evaluation, explanation, analysis, and synthesis are lower than those assessing remembering, understanding, and applying [51, 52]. Regarding the general arrangement of test or examination, easy items start first then are followed by difficult ones. At the same time, in the case of diagnostic assessment, the sequence of the learning material is more important [6, 7].

Easy and difficult items affect the item’s ability to discriminate between students and show low discrimination power. Some reports described a negative correlation between exam reliability and difficult and easy items [38]. Oermann et al. reported that educationalists must be careful in deleting items with poor DIF because the number of items has more effect on test validity [54]. It is recommended that difficult items should be reviewed for the possible technical and content causes [50]. Possible causes of low difficulty index include uncovered (taught) content material, challenging items, missed key or no correct answer among the item options [55]. Easy items (high *P*-value) can be due to technical causes, or the concerned learning objective (s) were achieved or revisited in coverage that is more superficial [55].

11. Interpretation of difficulty index

In literature including medical education, many ranges of difficulty indices were reported (Table 4).

Author	Difficulty index	Interpretation
Uddin et al. [50]	>80%	Easy
	30–80%	Moderate
	<30%	Difficult
Kaur, Singla et al. [56]	>80	Easy
	40–80	Moderate
	<39	Difficult
Sugianto [6]	90	Easy
	50	Moderate
	10	Difficult
Date, Borkar et al. [37] and Kumar, Jaipurkar et al. [36]	<30	Too difficult
	>70%	Too easy
	50–60%	Excellent/ideal
	30–70%	Good/acceptable/average
Obon and Rey [12]	> 0.76	Easy (Revise or Discard)
	0.26–0.75	Right difficult (Retain)
	0–0.25	Difficult (Revise or Discard)
Bhat and Prasad [57]	>70%	Easy
	30–70%	Good
	<30%	Difficult

Table 4.
Reference values and interpretation of difficulty index (*p*-value).

12. Discrimination index (Power)

Item discrimination (DI) is the ability of an item to discriminate between higher achiever (good) students and low ones. It was defined as “stated that item discrimination is a statistic that indicates the degree to which an item separates the students who performed well from those who did poorly on the test as a whole” [6]. The discrimination power of an item is calculated by categorizing the examinee into upper 27% and lower 27% according to their total test score. The difference between the upper and lower group is divided by the number of the examinee in the upper group or the larger group or by half of the total number of the examinee or even by the total number [4, 6, 58, 59]. Obon and Rey [12] calculated the discrimination index as the difference of difficulty index between the upper and lower groups [12]. In literature, both 25 and 27% were reported as possible percentages of examinee categorization [60, 61]. The 27% is commonly used to maximize differences in normal distributions and increase the number of examinees in each category. The discrimination index range from 1.0 to –1.0. The positive discrimination index indicates that high achievers answer the item correctly more than those in the lower ones, which is desirable. The negative discrimination index reflects that lower achiever examinees answer the item more correctly, while zero discrimination indicates equal numbers of students in the upper and lower groups [36, 37]. Negative discrimination is thought to be due to

item flaws or inefficient distractors, miss keys, ambiguous wording, gray areas of opinion, and areas of controversy [12, 62]. Nevid and McClelland [52] reported that items assessing evaluation and explanation domains could discriminate between high and low performers, while Kim et al. [51] comments that items assessing remembering and understanding levels have low discrimination power [52, 54].

It was reported that discrimination indices are positively associated with difficulty index and distractor efficiency [39, 63]. The discrimination power of the item is reduced by the increased number of non-functional distractors [36].

A test with poor discriminating power will not provide a reliable interpretation of the examinee's actual ability [6, 64]. In addition, discrimination power will not indicate item validity, and deletion of items with poor discrimination power negatively impacts validity due to a decrease in the item number [65].

13. Discrimination coefficients

Discrimination coefficients can evaluate item discrimination. The discrimination coefficients include point biserial correlation, biserial correlation, and phi coefficient. Although point biserial correlation is used interchangeably with the discrimination index, discrimination coefficients are considered superior to the discrimination index [24]. The superiority came from the fact that discrimination coefficients are calculated using all examinees' responses in the item rather than only 54% of the examinees such as in the discrimination index.

The difference between Point-biserial correlation (rBP) and discrimination indexes is that rBP is the correlation between an item in the exam and the overall student score [2, 66]. In cases of highly discriminating items, the examinees who responded to the item correctly also did well on the test. In general, the examinees who responded to the item incorrectly also tended to perform poorly on the overall test. It was suggested that point biserial can express the predictive validity better than Biserial correlation coefficients [61, 67].

14. Interpretation of discrimination index

Discrimination power of items more than 0.15 was reported as evidence of item validity [50, 53]. While any item with less than 0.15 or negative should be reviewed [50] (Table 5).

When interpreting the discrimination power of an item to decide about, especial consideration should be related to its difficulty. Items with a high difficulty index (most of the examinee answer it right) and those with low difficulty index (most of the examinee answer it wrong) commonly have low discrimination power [35, 63]. In both cases, such items will not discriminate examinees as the majority are on one side. Thus items with a moderate difficulty index are more likely to have good discrimination power.

The common causes of poor discrimination power of item include technical or writing flaws, untaught or not well covered content material, ambiguous wording, gray areas of opinion and controversy, and wrong keys [12, 50, 62, 66].

In general, the statistical data obtained from item analysis can help item constructors and exam composers to detect defective items. The decision to revise an item or distractors must be based on the difficulty index, discrimination index, and distractor efficiency. Revision of items can lead to modification in the teaching method or the content material [68].

Author	Discrimination power	Interpretation
Elfaki, Bahamdan et al. [53]	≥ 0.35	Excellent
	0.25–0.34	Good
	0.21–0.24	Acceptable
	≤ 0.20	Poor
Obon and Rey [12]	≥ 0.50	Very Good Item (Definitely Retain)
	0.40–0.49	Good Item (Very Usable)
	0.30–0.39	Fair Quality (Usable Item)
	0.20–0.29	Potentially Poor Item (Consider Revising)
	≤ 0.20	Potentially Very Poor (Possibly Revise Substantially or Discard)
Bhat and Prasad [57]	> 0.35	Excellent
	0.2–0.35	Good
	< 0.2	Poor
Sugianto [6]	> 0.40	Very good
	0.30–0.39	Reasonably good possibly need to improvement
	0.20–0.29	Marginal item usually needing and being to improvement
	< 0.19	Poor item rejected or improved by revision
Aljehani, Pullishery et al. [66] and Sharma [4]	≥ 0.40	Very discriminating, very good item (Keep)
	0.30–0.39	Discriminating item, good item (Keep)
	0.20–0.29	Moderately discriminating, fair item (Keep)
	< 0.20	Not discriminating item, marginal item (Revise/Discard)
	Negative	Worst/ defective item (Definitely Discard)
Ramzan, Imran et al. [63]	> 0.30	Excellent discrimination
	0.20–0.29	Good discrimination
	0–0.19	Poor discrimination
	00	Defective
Uddin et al. [50]	≥ 0.35	Excellent
	0.25–0.34	Good
	0.21–0.24	Acceptable
	< 0.20	Poor

Table 5. Reference values and interpretation of discrimination index (power).

15. Item analysis application

Figure 1:

In this Example 1.

The number of examinees was 21.

The number of test items (Total possible) is 40.

Exam Item Analysis Report				Exams Graded: 21				
Instructor:	0	Total Possible:	40	Average:	30.3 (75.83%)			
Exam Name:	0	Highest Score:	38(95%)	Median:	30(75%)			
Exam Date:	0	Lowest Score:	14(35%)	KR20:	0.8238477			
Correct responses are shown in bold and italics				Pt. Biserial	Disc. Index	Correct	Pct. Incorrect	
Q1	<i>A (18, 85.71%)</i>	B (0, 0.00%)	C (0, 0.00%)	D (3, 14.29%)	0.58	0.6	18, 85.7%	14.3%
Q2	A (0, 0.00%)	B (0, 0.00%)	C (0, 0.00%)	<i>D (21, 100.00%)</i>	0	0	21, 100.0%	0.0%
Q3	A (1, 4.76%)	<i>B (20, 95.24%)</i>	C (0, 0.00%)	D (0, 0.00%)	0.1	0.2	20, 95.2%	4.8%
Q4	A (4, 19.05%)	<i>B (15, 71.43%)</i>	C (0, 0.00%)	D (2, 9.52%)	-0.02	0	15, 71.4%	28.6%
Q5	A (0, 0.00%)	<i>B (15, 71.43%)</i>	C (3, 14.29%)	D (3, 14.29%)	0.29	0.25	15, 71.4%	28.6%
Q6	A (2, 9.52%)	B (2, 9.52%)	C (3, 14.29%)	<i>D (14, 66.67%)</i>	0.43	0.6	14, 66.7%	33.3%
Q7	A (8, 38.10%)	B (4, 19.05%)	<i>C (6, 28.57%)</i>	D (3, 14.29%)	0.39	0.67	6, 28.6%	71.4%
Q8	A (0, 0.00%)	<i>B (16, 76.19%)</i>	C (4, 19.05%)	D (1, 4.76%)	0.04	-0.33	16, 76.2%	23.8%
Q9	A (0, 0.00%)	B (1, 4.76%)	<i>C (20, 95.24%)</i>	D (0, 0.00%)	0.06	0	20, 95.2%	4.8%
Q10	<i>A (19, 90.48%)</i>	B (1, 4.76%)	C (0, 0.00%)	D (1, 4.76%)	0.27	0.2	19, 90.5%	9.5%

Figure 1.

Standard item analysis of mid-course examination. The total number of items is 40, and the total number of the examinee is 21. The KR20 is 0.82. Pt. Biserial: Point biserial correlation, Disc Index: discrimination index, Correct: number and percentage of the correct answer (difficulty index), Pct. Incorrect: percentage of an incorrect answer.

The highest and lowest scores were 38 and 14 respectively.

The class average (mean) (30.3) is more than the class median (30) which represents a positively skewed distribution of examinee scores. Despite this, examinee scores may show normal ball-shape distribution. If the median is larger than the average (mean), the examinee scores will be negatively skewed distribution. Average equals median, the examinees' scores are symmetrically (zero skewed) and normally distributed with ball-shaped.

The KR20 (Cronbach's alpha) is 0.82 which is an acceptable value for most of the authors. Such value of internal consistency of exam allows deciding pass/fail. Lower values put the exam in questionable status.

- Item 1: the difficulty index is 85.7% (easy). Although it has high discrimination power (DE = 0.6, Pbiserial = 0.58), two distractors are non-functional (B, C).

Comment: the item needs reediting. Distractors B and C need to be revised or changed by more plausible ones before being re-used.

- Item 2: the difficulty index is 100% (easy). It has low discrimination power (DE = 00, Pbiserial = 00), all distractors are non-functional.

Comment: the item needs major revision or rewriting. This item is absolutely easy with no difficulty or discrimination index. Such items should be removed from the question bank and removal from the exam is considered valid.

- Item 6: the difficulty index is 66.7% (moderate). It has high discrimination power (DE = 0.6, Pbiserial = 0.43) and all the distractors are functional.

Comment: The item has acceptable indices. Such items can be saved in the question bank for further use. The distractors need to be updated to have more efficiency.

- Item 7: the difficulty index is 28.6% (difficult). Although it has high discrimination power (DE = 0.67, Pbiserial = 0.39), all the distractors are functional.

Comment: The item has acceptable indices. Such items can be saved in the question bank for further use. The distractors need to be updated to have more efficiency.

- Item 8: the difficulty index is 76.2% (moderate). This item has a negative discrimination index (-0.33) and poor Pbiserial (0.04). Only one distractor is functional (C). The negative discrimination index is caused by the increased number of students in the lower account (27%) than those in the upper account (27%).

Comment: although the item has a moderate difficulty index, but is poorly discriminating. Such an item needs major revision.

Exam Item Analysis Report				Exams Graded: 25				
Instructor:	0	Total Possible:	40	Average:	24.6(61.40%)			
Exam Name:	0	Highest Score:	33(82.5%)	Median:	25(62.50%)			
Exam Date:	0	Lowest Score:	13(32.5%)	KR20	0.7409171			
Correct responses are shown in bold and italics				Pt. Biserial	Disc. Index	Correct	Pct. Incorrect	
Q1	A (1, 4.00%)	B (4, 16.00%)	C (1, 4.00%)	D (19, 76.00%)	0.39	0.5	19, 76.0%	24.0%
Q4	A (0, 0.00%)	B (2, 8.00%)	C (1, 4.00%)	D (22, 88.00%)	0.36	0.5	22, 88.0%	12.0%
Q5	A (2, 8.00%)	B (22, 88.00%)	C (0, 0.00%)	D (1, 4.00%)	0.2	0.33	22, 88.0%	12.0%
Q8	A (1, 4.00%)	B (18, 72.00%)	C (1, 4.00%)	D (5, 20.00%)	-0.06	-0.17	1, 4.0%	96.0%
Q9	A (11, 44.00%)	B (7, 28.00%)	C (5, 20.00%)	D (2, 8.00%)	0.09	0.17	5, 20.0%	80.0%
Q10	A (3, 12.00%)	B (6, 24.00%)	C (10, 40.00%)	D (6, 24.00%)	0.32	0.17	3, 12.0%	88.0%
Q11	A (11, 44.00%)	B (1, 4.00%)	C (3, 12.00%)	D (10, 40.00%)	0.01	0	11, 44.0%	56.0%
Q12	A (5, 20.00%)	B (1, 4.00%)	C (13, 52.00%)	D (6, 24.00%)	0.19	0.33	13, 52.0%	48.0%

Figure 2.

Standard item analysis of Mid-course examination. The total number of items is 40, and the total number of examinee is 25. The KR20 is 0.74. Pt.Biserial: Point biserial correlation, Disc Index: discrimination index, Correct: number and percentage of the correct answer (difficulty index), Pct. Incorrect: percentage of an incorrect answer.

Figure 2:

In this Example 2.

The number of examinees was 25.

The number of test items is 40.

The highest and lowest scores were 33 and 13 respectively.

The class average (mean) (24.6) is more than the class median (25), distribution of examinee scores is skewed to the left. Despite this, examinee scores may show normal ball shape distribution.

The KR20 (Cronbach's alpha) is 0.74 which is an acceptable value for most of the authors. Such a value of internal consistency is suitable for class tests.

- Item 8: the difficulty index is 4.0% (difficult). It has negative discrimination power (DE = -0.17, Pbiserial = -0.06), one distractors is non-functional (C).

Comment: the correct answer is (A) while most of the examinees chose (B). According to distractor analysis, this item is miss-keyed rather than an implausible distractor.

- Item 9: the difficulty index is 20% (difficult). It has low discrimination power (DE = 0.17, Pbiserial = 0.09), all distractors are functional.

Comment: distractor analysis show option number (A) and (B) are more selected by examinees. This can be due to implausible. The presence of implausible can affect the item difficulty index. Distractors in this item should be revised or changed with plausible ones.

- Item 11: the difficulty index is 44.0% (moderate). It has low discrimination power (DE = 0.0, Pbiserial = 0.01) and only one the distractors is non-functional.

Comment: The item has an acceptable difficulty index. Distractor (D) is more selected by upper examinee such as the key answer. Such a situation can favor missed key or implausible distractors. The distractors need to be updated to have more efficiency.


Author details

Assad Ali Rezigalla

Department of Basic Medical Sciences (Unit of Anatomy), College of Medicine,
University of Bisha, Saudi Arabia

*Address all correspondence to: assadkafe@yahoo.com

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Stassen, M.L., K. Doherty, and M. Poe, *Program-based review and assessment: Tools and techniques for program improvement*. 2004: Office of Academic Planning & Assessment, University of Massachusetts Amherst.
- [2] Tavakol, M. and R. Dennick, *Post-examination analysis of objective tests*. *Med Teach*, 2011. **33**(6): p. 447-58.
- [3] Benson, J., *A comparison of three types of item analysis in test development using classical and latent trait methods*, in *GRADUATE COUNCIL OF THE UNIVERSITY OF FLORIDA*. 1978, UNIVERSITY OF FLORIDA: FLORIDA, USA. p. 134.
- [4] Sharma, L.R., *Analysis of difficulty index, discrimination index and distractor efficiency of multiple choice questions of speech sounds of english*. *International Research Journal of MMC*, 2021. **2**(1): p. 15-28.
- [5] Thompson, B. and J.E. Levitov, *Using microcomputers to score and evaluate items*. *Collegiate Microcomputer archive*, 1985. **3**.
- [6] Sugianto, A., *Item analysis of english summative test: Efl teacher-made test*. *Indonesian EFL Research*, 2020. **1**(1): p. 35-54.
- [7] Kumar, H. and S.K. Rout, *Major tools and techniques in educational evaluation, in Measurement and evaluation in education*. 2016, Vikas Publishing House Pvt. Ltd.: India. p. 256.
- [8] Case, S.M. and D.B. Swanson, *Constructing written test questions for the basic and clinical sciences*. 3 ed. 1998, United States of America: National Board of Medical Examiners Philadelphia, PA. 129.
- [9] Kruyen, P.M., *Using short tests and questionnaires for making decisions about individuals: When is short too short?* 2012, Netherlands Ridderprint BV, Ridderkerk.
- [10] Akinboboye, J.T. and M.A. Ayanwale, *Bloom taxonomy usage and psychometric analysis of classroom teacher made test*. *AFRICAN MULTIDISCIPLINARY JOURNAL OF DEVELOPMENT*, 2021. **10**(1): p. 10-21.
- [11] Brookhart, S.M. and A.J. Nitko, *Education assessment of students*. New jearsey: Merrill prentice hall. 2018, New Jersey: Pearson; 8th edition.
- [12] Obon, A.M. and K.A.M. Rey, *Analysis of multiple-choice questions (mcqs): Item and test statistics from the 2nd year nursing qualifying exam in a university in cavite, philippines*. *Abstract Proceedings International Scholars Conference*, 2019. **7**(1): p. 499-511.
- [13] Downing, S. and R. Yudkowsky, *Assessment in health professions education*. 2009, New York and London: Routledge and Taylor & Francis
- [14] Silao, C.V.O. and R.G. Luciano, *Development of an automated test item analysis system with optical mark recognition (omr)* *International Journal of Electrical Engineering and Technology (IJEET)*, 2021. **12**(1): p. 67-79.
- [15] Reinhardt, B.M., *Factors affecting coefficient alpha: A mini monte carlo study*, in *The Annual Meeting of the Southwest Educational Research Association (January 26, 1991)*. 1991, University of Texas: San Antonio, Texas, USA. p. 31.
- [16] Tavakol, M. and R. Dennick, *Making sense of cronbach's alpha*. *International journal of medical education*, 2011. **2**: p. 53-55.
- [17] Graham, J.M., *Congeneric and (essentially) tau-equivalent estimates of*

score reliability: What they are and how to use them. Educational and Psychological Measurement, 2006. 66(6): p. 930-944.

[18] Rezigalla, A.A., A.M.E. Eleragi, and M. Ishag, *Comparison between students' perception toward an examination and item analysis, reliability and validity of the examination*. Sudan Journal of Medical Sciences, 2020. 15(2): p. 114-123.

[19] Considine, J., M. Botti, and S. Thomas, *Design, format, validity and reliability of multiple choice questions for use in nursing research and education*. Collegian, 2005. 12(1): p. 19-24.

[20] Cortina, J.M., *What is coefficient alpha? An examination of theory and applications*. Journal of applied psychology, 1993. 78(1): p. 98.

[21] McNeish, D., *Thanks coefficient alpha, we'll take it from here*. Psychol Methods, 2018. 23(3): p. 412-433.

[22] Panayides, P., *Coefficient alpha: Interpret with caution*. Europe's Journal of Psychology, 2013. 9(4): p. 687-696.

[23] Al-Osail, A.M., et al., *Is cronbach's alpha sufficient for assessing the reliability of the osce for an internal medicine course?* BMC research notes, 2015. 8(1): p. 1-6.

[24] McCowan, R.J. and S.C. McCowan, *Item analysis for criterion-referenced tests*. 1999, New York: Center for Development of Human Services.

[25] Salkind, N.J., *Encyclopedia of research design*. Vol. 1. 2010: sage.

[26] Robinson, J.P., P.R. Shaver, and L.S. Wrightsman, *Scale selection and evaluation*, in Measures of political attitudes, J.P. Robinson, P.R. Shaver, and L.S. Wrightsman, Editors. 1999, Academic Press: USA. p. 509.

[27] Cicchetti, D.V., *Guidelines, criteria, and rules of thumb for evaluating*

normed and standardized assessment instruments in psychology. Psychological assessment, 1994. 6(4): p. 284.

[28] Axelson, R.D. and C.D. Kreiter, *Reliability*, in Assessment in health professions education, R. Yudkowsky, Y.S. Park, and S.M. Downing, Editors. 2019, Routledge: London.

[29] Hassan, S. and R. Hod, *Use of item analysis to improve the quality of single best answer multiple choice question in summative assessment of undergraduate medical students in malaysia*. Education in Medicine Journal, 2017. 9(3): p. 33-43.

[30] Green, S.B. and M.S. Thompson, *Structural equation modeling in clinical psychology research*, in *Handbook of research methods in clinical psychology*, M.C. Roberts and S.S. Ilardi, Editors. 2008, Wiley-Blackwell. p. 138-175.

[31] Mahjabeen, W., et al., *Difficulty index, discrimination index and distractor efficiency in multiple choice questions*. Annals of PIMS-Shaheed Zulfiqar Ali Bhutto Medical University, 2017. 13(4): p. 310-315.

[32] Mehta, G. and V. Mokhasi, *Item analysis of multiple choice questions-an assessment of the assessment tool*. Int J Health Sci Res, 2014. 4(7): p. 197-202.

[33] Tarrant, M., J. Ware, and A.M. Mohammed, *An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis*. BMC Medical Education, 2009. 9(1): p. 40.

[34] Puthiaparampil, T. and M. Rahman, *How important is distractor efficiency for grading best answer questions?* BMC medical education, 2021. 21(1): p. 1-6.

[35] Gajjar, S., et al., *Item and test analysis to identify quality multiple choice questions (mcqs) from an assessment of medical students of ahmedabad, gujarat*.

Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine, 2014. **39**(1): p. 17.

[36] Kumar, D., et al., *Item analysis of multiple choice questions: A quality assurance test for an assessment tool*. Medical Journal Armed Forces India, 2021. **77**: p. S85-S89.

[37] Date, A.P., et al., *Item analysis as tool to validate multiple choice question bank in pharmacology*. International Journal of Basic & Clinical Pharmacology, 2019. **8**(9): p. 1999-2003.

[38] Abdalla, M.E., *What does item analysis tell us? Factors affecting the reliability of multiple choice questions (mcqs)*. Gezira Journal of Health Sciences, 2011. **7**(2).

[39] Fozzard, N., et al., *Analysis of mcq and distractor use in a large first year health faculty foundation program: Assessing the effects of changing from five to four options*. BMC Med Educ, 2018. **18**(1): p. 252.

[40] Sajjad, M., S. Iltaf, and R.A. Khan, Nonfunctional distractor analysis: An indicator for quality of multiple choice questions. Pak J Med Sci, 2020. **36**(5): p. 982-986.

[41] Haladyna, T.M., S.M. Downing, and M.C. Rodriguez, *A review of multiple-choice item-writing guidelines for classroom assessment*. Applied measurement in education, 2002. **15**(3): p. 309-333.

[42] Swanson, D.B., K.Z. Holtzman, and K. Allbee, *Measurement characteristics of content-parallel single-best-answer and extended-matching questions in relation to number and source of options*. Academic Medicine, 2008. **83**(10): p. S21-S24.

[43] Frary, R.B., *More multiple-choice item writing do's and don'ts*. Practical Assessment, Research, Evaluation, 1994. **4**(1): p. 11.

[44] Abdulghani, H.M., et al., *The relationship between non-functioning distractors and item difficulty of multiple choice questions: A descriptive analysis*. Journal of Health Specialties, 2014. **2**(4): p. 148.

[45] Alhummayani, F.M., *Evaluation of the multiple-choice question item analysis of the sixth year undergraduate orthodontic tests at the faculty of dentistry, king abdulaziz university, saudi arabia*. Egyptian Orthodontic Journal, 2020. **57**(June 2020): p. 1-18.

[46] Licon-Chávez, A.L. and L.R. Velázquez-Liaño, *Quality assessment of a multiple choice test through psychometric properties*. MedEdPublish, 2020. **9**.

[47] Hassan, S., *Item analysis, reliability statistics and standard error of measurement to improve the quality and impact of multiple choice questions in undergraduate medical education in faculty of medicine at unisza*. Malaysian Journal of Public Health Medicine, 2016. **16**(3): p. 7-15.

[48] Hingorjo, M.R. and F. Jaleel, *Analysis of one-best mcqs: The difficulty index, discrimination index and distractor efficiency*. J Pak Med Assoc, 2012. **62**(2): p. 142-7.

[49] Lord, F.M., *The relation of the reliability of multiple-choice tests to the distribution of item difficulties*. Psychometrika, 1952. **17**(2): p. 181-194.

[50] Uddin, I., et al., *Item analysis of multiple choice questions in pharmacology*. J Saidu Med Coll Swat, 2020. **10**(2): p. 128-131.

[51] Kim, M.-K., et al., *Incorporation of bloom's taxonomy into multiple-choice examination questions for a pharmacotherapeutics course*. American journal of pharmaceutical education, 2012. **76**(6).

- [52] Nevid, J.S. and N. McClelland, *Using action verbs as learning outcomes: Applying bloom's taxonomy in measuring instructional objectives in introductory psychology*. Journal of Education and Training Studies, 2013. 1(2): p. 19-24.
- [53] Elfaki, O., K. Bahamdan, and S. Al-Humayed, *Evaluating the quality of multiple-choice questions used for final exams at the department of internal medicine, college of medicine, king khalid university*. Sudan Med Monit, 2015. 10: p. 123-27.
- [54] Oermann, M.H. and K.B. Gaberson, *Evaluation and testing in nursing education*. 6 ed. 2019, New York: Springer Publishing Company.
- [55] Bukvova, H., K. Figl, and G. Neumann, *Improving the quality of multiple-choice exams by providing feedback from item analysis*.
- [56] Kaur M, Singla S, Mahajan R. Item analysis of in use multiple choice questions in pharmacology. International Journal of Applied and Basic Medical Research. 2016;6(3): 170-173.
- [57] Bhat SK, Prasad KHL. Item analysis and optimizing multiple-choice questions for a viable question bank in ophthalmology: A cross-sectional study. Indian J Ophthalmol. 2021;69(2):343-6.
- [58] Rogausch, A., R. Hofer, and R. Krebs, *Rarely selected distractors in high stakes medical multiple-choice examinations and their recognition by item authors: A simulation and survey*. BMC Medical Education, 2010. 10(1): p. 1-9.
- [59] Wood, D.A. and D.C. Adkins, *Test construction: Development and interpretation of achievement tests*. 1960: CE Merrill Books.
- [60] Wiersma, W. and S.G. Jurs, *Educational measurement and testingallyn & bacon*. 1990, Boston. 415.
- [61] Matlock-Hetzler, S., *Basic concepts in item and test analysis*. 1997.
- [62] Sim, S.-M. and R.I. Rasiah, *Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper*. Annals-Academy of Medicine Singapore, 2006. 35(2): p. 67.
- [63] Ramzan, M., et al., *Item analysis of multiple-choice questions at the department of community medicine, wah medical college, pakistan*. Life and Science, 2020. 1(2): p. 4-4.
- [64] Setiyana, R., *Analysis of summative tests for english*. English Education Journal, 2016. 7(4): p. 433-447.
- [65] Oermann, M.H. and K.B. Gaberson, *Evaluation and testing in nursing education*. 2016: Springer Publishing Company.
- [66] Aljehani, D.K., et al., *Relationship of text length of multiple-choice questions on item psychometric properties—a retrospective study*. Saudi Journal for Health Sciences, 2020. 9(2): p. 84.
- [67] Henrysson, S., *Gathering, analyzing, and using data on test items*, in *Educational measurement*, R.L. Thorndike, Editor. 1971, American Council on Education: Washington DC. p. 141.
- [68] Maulina, N. and R. Novirianthy, *Item analysis and peer-review evaluation of specific health problems and applied research block examination*. Jurnal Pendidikan Kedokteran Indonesia: The Indonesian Journal of Medical Education, 2020. 9(2): p. 131-137.