# Item analysis of examinations in the Faculty of Medicine of Tunis

## Analyse docimologique des épreuves écrites à la Faculté de Médecine de Tunis

Amene Hermi[1], Wafa Achour[2]

*1-Doha / Faculté de Médecine de Tunis*
*2-Service des laboratoires, Centre national de greffe de moelle osseuse, Tunis / Faculté de Médecine de Tunis,*

**RÉSUMÉ**

**Prérequis :** Parmi les diverses techniques docimologiques permettant l'étude objective des épreuves, l'analyse des items (questions) utilise différents indices et coefficients obtenus à partir des questions et de l'examen pour en juger la valeur. Notre travail avait pour objectif principal de déterminer les indices métriques principaux des questions d'examen.

**Méthodes :** Notre base de données a été l'ensemble des notes de tous les étudiants, à tous les niveaux, à toutes les questions, des épreuves écrites passées à la session principale de l'année universitaire 2012-2013 à la Faculté de Médecine de Tunis, soit un total de 2515 étudiants, de 66 épreuves (thèmes / certificats), de 187 disciplines et de 3138 questions. Nous avons utilisé pour l'analyse des questions le fichier « AnItem.xls ».

**Résultats :** L'indice de difficulté moyen des 66 épreuves a été optimal optimal, soit 0.59. Les questions classées à difficulté acceptable ou faciles ont été prédominantes, représentant 89.17% du total des questions. Le pouvoir discriminant de toutes épreuves a été moyen, soit un indice moyen de discrimination de 0.28. Les questions à mauvaise discrimination ont représenté 23,62% du total des questions. La meilleure discrimination a été retrouvée pour les disciplines à difficulté variant de 0,4 à 0,6. Les questions idéales ont représenté 27,02% des questions avec de larges variations au sein des disciplines. L'homogénéité interne, toutes épreuves confondues, a été acceptable (alpha de Cronbach de 0,79) mais la majorité des disciplines avait une homogénéité non acceptable. Ces dernières avaient au maximum 33 questions (chacune) et la corrélation entre leur alpha et le nombre de leurs questions a été positive. Les courbes de distribution des scores ont été majoritairement platykurtiques (72.73%) et à asymétrie négative (89,39%). Le Premier Cycle des Etudes Médicales 1 (PCEM1), soit la première année, a été le niveau avec les meilleurs indicateurs métriques.

**Conclusion :** L'analyse d'items a montré que nos épreuves étaient de consistance interne acceptable et de bonne qualité en termes de difficulté et de discrimination. Elles tendaient à la facilité et discriminaient surtout entre les étudiants moyens. Une analyse continue permettra d'améliorer encore plus leur qualité.

**Mots-clés**

Indice de difficulté, indice de discrimination, consistance interne, courbe de distribution des scores.

**SUMMARY**

**Background:** Item analysis is the process of collecting, summarizing and using information from students' responses to assess test items' quality. This study used this approach to evaluate the quality of items and examinations given in the Faculty of Medicine of Tunis (FMT).

**Methods:** This study concerned the examinations of 2012-2013 (principal session). It analyzed 3138 items from 66 examinations, of which, 46 were multidisciplinary (187 disciplines). A total of 2515 students took the examinations. "AnItem.xls" file was used for the analysis that focused on difficulty, discrimination and internal consistency.

**Results:** Mean difficulty for all examinations was optimum (mean difficulty index: 0.59). Majority of items (89.17%) were either easy or of acceptable difficulty. Mean discrimination for all examinations was moderate (mean item discrimination coefficient: 0.28) with poor discrimination in 23.62% of items. Maximal discrimination occurred with disciplines of difficulty index between 0.4-0.6. « Ideal » items represented 27.02%. Mean internal consistency for all examinations was acceptable (Cronbach's alpha: 0.79). Disciplines with nonacceptable internal consistency (68.45%) contained a maximum of 33 items (each one) and a positive correlation between their alpha and the number of their questions. Distributions were mostly (72.73%) platykurtic and negatively asymmetric (89.39%). First year of studies had the best parameters.

**Conclusion:** Our examinations had an acceptable internal consistency, and a good level of difficulty and discrimination. They tended to facility and discriminated basically students of medium level. Item analysis is useful as a guide to item writers to improve the overall quality of questions in the future.

**Key-words**

Difficulty index, discrimination index, internal consistency, score distribution

Item analysis is a process which examines student responses to individual test items (questions) in order to assess the quality of those items and of the test as a whole. The data generated is used by the examiner to revise and then modify or remove specific items from subsequent exams. In addition, item analysis is valuable for increasing instructors' skills in test construction, and identifying specific areas of course content which need greater emphasis or clarity [1, 2]. Many statistic and metric tools can be used in this type of analysis, such as the difficulty and discrimination indices, the reliability of the test and the test score distribution. The difficulty index, also called ease index, determines the item's difficulty and ranges from 0 to +1 [2, 3]. The higher is the value, the easier is the item [2].The discrimination index, determines whether those who did well on the entire test did well on a particular item. It ranges between -1 and +1 [1, 4]. A positive value indicates that students with high scores on the overall test are also getting the item right and that students with low scores on the overall test are getting the item wrong (which we would expect). A negative value implies that students who get the item correct tend to do poorly on the overall test and that students who get the item wrong tend to do well on the test (which would indicate an anomaly) [1, 4]. Reliability of the test refers to the extent to which the test is likely to produce consistent scores. It reflects whether the obtained score is a stable indication of the student's performance on a particular test [2]. It is measured by Cronbach's alpha, an index varying between 0 (no reliability) and 1 (perfect reliability). The higher the score, the more reliable the generated examination is [5, 6]. The test score distribution complements the item analysis data by providing a description of how the class as a whole performed on the test. The purpose of this study was to lead, for the first time in our faculty, the Faculty of Medicine of Tunis (FMT), an item analysis of the administered summative tests based on the different statistic and metric tools cited above.

## METHODS

Our study, conducted at the FMT, interested the examinations of the main session of the academic year 2012-2013. We included in this study, 3138 items taken from 66 multidisciplinary examinations (from the total of 69 examinations of the session) carried out in January and June for the 5 first levels of studies. A total of 2515 students sat for those examinations. In the three last levels (3rd, 4th and 5th years), each cohort was divided into 2 groups A and B. They passed the certificates in a crossed way. As an example, for the 3rd level, in January, while group A was passing the cardiology, group B passed the neurology and vice-versa in June. The items were either multiple choice questions (MCQs), short-answer questions (SAQ) or essay questions (EQ). The MCQs did not have neither the same number of options nor the same type or scoring.

We used for the analysis, an Excel file, "AnItem.xls", used at the assessment office of the Faculty of Medicine of Montreal University since 2000 and downloaded easily online. For each examination, the Faculty kept an Excel file containing the marks obtained by every student on each of the question. From those files we created an "AnItem.xls" for each test.

We analyzed the items for their level of difficulty and their discrimination power.

Difficulty index was calculated as follows: the mean mark obtained by all candidates attempting the item divided by the maximum mark available on the item [3]. We used these intervals to classify the index values: "_> 0.7 – Easy, 0.3- 0.7 – Acceptable, _ < 0.3 – Difficult, _ 0.5-0.6 – Optimum [7, 8]. Discrimination index was calculated using the corrected item-total correlation. It is a Pearson correlation between the sum of marks of the examinees on the item and the sum of their scores on all the other items [9]. We used this classification: "_≥ 0.4 – Excellent, _0.3-0.39 – Good, _ 02-0.29 – Marginal (needs improvement), _ ≤ 0.19 – Poor (reject or revise) [10]. Discrimination index of 0.2 or higher was acceptable and the test item would be able to differentiate between the weak and good students [11]. An index equal to zero indicates that there was no discrimination. When it was negative, this showed incoherence: the best students fail the item; the weakest ones answer correctly [12]. Items with acceptable difficulty and good/excellent discrimination were considered as 'ideal' [1].

For each examination, the percentage of those questions was concluded.

The relationship between the item difficulty and discrimination index for each test paper was also determined by Pearson correlation (r) [13].

In terms of reliability, it is important first to perceive the significance of a high and low test reliability. In fact, a high test reliability means that all items tend to «pull together.» Students who answered a given item correctly were more likely to answer other items correctly. If a parallel test were developed by using similar items, the relative scores of students would show little change. In the other hand, a low test reliability means that the questions tend to be unrelated to each other in terms of who answered them correctly. The resulting test scores reflect peculiarities of the items or the testing situation more than students' knowledge of the subject matter [2]. Reliability of each exam was assessed by Cronbach's alpha. An index of 0.7 and above was acceptable [14]. We used the criteria of George and Mallery (2003) providing the following rules of thumb: "_≥ 0.9 – Excellent (at the level of best standardized tests), _ ≥ 0.8 – Good (very good for a classroom test), _ ≥ 0.7 – Acceptable (Good for a classroom test; in the range of most. There are probably a few items which could be improved)" [15]. The alpha of

the examination if item deleted was also determined. When the alpha value was higher than the current alpha with the item included, one should consider deleting this item to improve the overall reliability of the exam. For each examination, we determined the percentage of items altering or improving alpha.

The test score distribution provides a visual representation of the score distribution which indicates the range of test scores obtained by the students and the number of students obtaining each score value. The abscissa axis presents the score values from the lowest obtained score to the highest. The ordinate axis indicates the number of students who received each particular score [2, 10]. The score distribution assumes the properties of a normal distribution. Two characteristics should be analyzed. The skewness of the distribution which is an indicator of the overall difficulty of the test, and its shape, which tells us about the group concerned more with discrimination. A positively skewed distribution (positive coefficient of skewness) means that most of the test scores are low with only a few students at the high end. The test tends to be difficult. A negatively skewed distribution (negative coefficient of skewness) means that most of the scores are high with only a few students at the low end. The test tends to be easy [10]. The coefficient of kurtosis measures the peakness or flatness of a distribution. In other words, kurtosis refers to the degree of dispersion among the scores. Depending on its value, three forms are described and referred to as mesokurtic, platykurtic and leptokurtic. Mesokurtic distributions (kurtosis=0) have peaks of medium height, and are moderate in breadth. Leptokurtic distributions (kurtosis > 0) are tall and thin, with only a few scores in the middle having high frequency. Discrimination is mostly within higher and lower students.

Platykurtic distributions (kurtosis < 0) are short and more dispersed (broader). There are many scores around the middle score that all have relatively high frequency. In this case, the discrimination is more between the medium students [16, 17].

In our study, the score distribution for each exam was determined and corresponding coefficients of skewness and kurtosis were calculated.

## RESULTS

The analysis indicates that the majority of items analyzed had an acceptable difficulty (52.33%). Easy items represented 36.84% whereas difficult ones were a minority (10.83%). In all levels, except for the 4th (B) and 5th (A) levels, items with acceptable range of difficulty overpassed 50%. No difficulty index=0 was found. Controversy, items with an index =1 were found in 4 examinations. The mean difficulty index of all the exams was 0.59 (optimum). All the levels had an index in the acceptable range (optimum for 1st, 2nd and 3rd levels). The majority of tests (72.73% or 61/66) had an acceptable difficulty and 27.27% (5/66) were easy. No examinations were classed as difficult. Examinations with an optimum difficulty represented 34.84% (23/66) of the total.

The items with good discrimination were the majority (29.09% total items). Those with excellent, poor and marginal discrimination represented respectively 19.76%, 23.62% and 27.53%. Some items with a negative index were found in 31 tests. The mean discrimination index for all levels was 0.28 (marginal discrimination). In fact, 56.06% (37/66) of total tests had marginal discrimination. This predominance was found in all levels except for the 1st one (majority of tests with good discrimination), the unique level that had tests with excellent mean discrimination index. The two tests with poor discrimination (3.03%) were found in the 4th level (A) and the 5th level (B). Exams with good discrimination were 37.88% of the total (25/66).

The coefficient of correlation between the mean indices of difficulty and discrimination of all tests was (-0.09). Exams with negative index were a minority in all levels (28.79%: 19/66) except for the 2nd one (4/7 of tests). Those with a positive index were 93.61% (low positive correlation: r<0.5).

Ideal items were 27.02% of total items with a maximal percentage in the 1st level (42.81%).

The mean alpha of all exams was 0.79 (acceptable). Reliability was good in all levels (with a best coefficient in the 1st level) except for the 4th (A) and 5th (B) (acceptable). Majority of exams had a good reliability (74.24% or

**Table 1:** Characteristics of items per level

| | 1st | 2nd | 3rd | | 4th | | 5th | |
|---|---|---|---|---|---|---|---|---|
| | | | A | B | A | B | A | B |
| **Number of items** | 459 | 422 | 389 | 391 | 273 | 284 | 458 | 462 |
| **Difficult** | 12.38 | 13.12 | 16.08 | 12.50 | 8.36 | 9,39 | 6.85 | 7.97 |
| **Acceptable difficulty** | 57.27 | 53.73 | 54.37 | 55.71 | 55.48 | 46,04 | 45.10 | 50.92 |
| **Easy** | 30.35 | 33.15 | 29.55 | 31.79 | 36.16 | 44,57 | 48.05 | 41.11 |
| **Bad discrimination** | 9.04 | 24.52 | 24.84 | 27.91 | 29.52 | 22,37 | 23.33 | 27.47 |
| **Marginal discrimination** | 20.63 | 32.12 | 26.89 | 27.67 | 28.56 | 27.24 | 29.45 | 27.68 |
| **Good discrimination** | 30.91 | 28.09 | 29.56 | 30.02 | 28.50 | 31.44 | 27.03 | 27.17 |
| **Excellent discrimination** | 39.42 | 15.27 | 18.71 | 14.40 | 13.42 | 18.95 | 20.19 | 17.68 |
| **Ideal** | 42.81 | 25.72 | 30.61 | 25.63 | 22.00 | 22.83 | 22.09 | 24.53 |
| **Alterating alpha** | 7.01 | 0.47 | 10.89 | 1.21 | 3.51 | 0.50 | 2.74 | 2.97 |

49/66). This predominance was found in all levels, except for the 5th (B). Exams with an alpha under 0.7 (4.55% or 3/66) were found in the 2nd, 4th (A) and 5th (B) levels. Only 3.66% of total items alterated the reliability of exams (with extremes of 0.47% in the 2nd and of 10.89% in the 3rd (A). We found that 39.91% of items improved it.
The majority of exams (89.39% or 59/66) had negatively skewed distributions. This predominance was found in all levels. The 5 exams with positively skewed distributions were present in all levels except for the 3rd, 4th (B) and 5th (A) ones. Perfect symmetry was found in 2 exams. The majority of scores distributions were platykurtic (72.73% or 48/66). This predominance was found in all levels except for the 4th one. Only the 3rd (A) level did not have any exam with leptokurtic score distribution.
The 1st level was the only level with all best indices.
The characteristics of items and exams per level are presented in tables I and II. Mean indices per level are presented in table III.

## DISCUSSION

In our study, the mean difficulty of all tests was optimum (mean index = 0.59). Items with an acceptable difficulty and those easy were 89.17% of total items. This percentage was almost close to that found in an Iranian item analysis of 1496 MCQs realized in a medical school (92.38%): easy items and those with acceptable difficulty were respectively 44.58% and 47.8% [8]. Although our target is always to write questions with acceptable difficulty, the test should include easy questions (so not remove systematically), that should be placed at the start of the test as 'warm-up' questions to motivate the student and give him self-confidence to distinguish the weak students and difficult questions (to be preferably at the end of test) to differentiate students of higher level. Thus, all levels of difficulty must be found and the location issues should not be arbitrary. Inclusion of very difficult items in the test depends upon the target of the teacher, who may want to include them in order to identify top scorers [1]. The difficult items should be reviewed for possible confusing language, ambiguity, areas of controversy, or even a wrong key [1, 13]. The item content may be a part of a course content that students did not learn well. It may also not be adequately taught in this particular academic session for certain reasons (absenteeism, insufficient time, etc.) [13]. On the other hand, some items turn out to be easier than expected because the students may have learned the content particularly well. The incorrect choices of an MCQ may be obviously incorrect. Some fault in the wording may also provide a clue to the correct answer. It is important also to underline that item difficulty values are extremely dependent on the group for which they are computed. The students of the FMT are the top graduates of the country. This could perhaps explain the optimum mean difficulty index of tests found.
The mean discrimination for all tests was marginal (mean index= 0.28). A total of 80.24% of items were at the level of marginal or better discrimination. Ware and Vik recommend that at least 60% of items should have marginal or better discrimination [18]. A Malaysian study analyzing 120 MCQs of 12 multidisciplinary exams, from 2003 to 2006, administered in a medical school shows that 67 % of items have marginal discrimination or better and that 37.5% have excellent discrimination [1].

**Table 2 :** Characteristics of examinations per level

| Examinations number | | 1st | 2nd | 3rd | | 4th | | 5th | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | A | B | A | B | A | B |
| **Total** | | 11 | 7 | 7 | 7 | 6 | 6 | 11 | 11 |
| **Difficult** | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Acceptable difficulty (optimum)** | | 9 (2) | 7 (3) | 7 (6) | 7 (5) | 6 (2) | 5 (2) | 10 (2) | 10 (1) |
| **Easy** | | 2 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| **Discrimination** | Bad | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| | Marginal | 1 | 5 | 5 | 5 | 4 | 3 | 6 | 8 |
| | Good | 8 | 2 | 2 | 2 | 1 | 3 | 5 | 2 |
| | Excellent | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Correlation between difficulty and discrimination indices** | | 2 | 4 | 1 | 2 | 2 | 1 | 3 | 4 |
| **Reliability** | Good | 11 | 6 | 6 | 4 | 4 | 5 | 8 | 5 |
| | Acceptable | 0 | 0 | 1 | 3 | 1 | 1 | 3 | 5 |
| | Non acceptable | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| **Scores distribution** | (+) skewness | 1 | 1 | 0 | 0 | 2 | 0 | 0 | 1 |
| | (-) skewness | 8 | 6 | 7 | 7 | 4 | 6 | 11 | 10 |
| | Symmetric | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Platykurtic | 10 | 5 | 7 | 5 | 3 | 3 | 7 | 8 |
| | Leptokurtic | 1 | 2 | 0 | 2 | 3 | 3 | 4 | 3 |

This percentage is 50.7 in another study conducted in Oman in 2009, about 150 MCQ [18]. Few common causes for the poor discrimination are ambiguous wording, grey areas of opinion, wrong keys and areas of controversy. Items showing poor discrimination should be referred back to the content experts for revision to improve the standard of these test items [19]. However, there may be other factors that need to be taken into account, especially when dealing with a multidisciplinary paper. Students' performance in Pharmacology's items may not accurately predict their performance in those of anatomy, neither their overall performance in the total test [19]. Our study found that there were items with a negative index in 31/66 exams. In fact, it is possible that a "good" student might not risk attempting a "difficult" item for fear of losing hard-earned marks on the other items of the same question. However, a "weak" student might take the risk to guess as he knows so little on the topic that he has nothing much to lose, and the least he can obtain for the whole question is zero marks. This could then result in a negative discrimination index [19]. Also, it should be noted that an item discrimination value is unique to a group of examinees. An item with satisfactory discrimination for one group may be unsatisfactory for another.

Pearson correlation between mean difficulty and discrimination indices showed that discrimination index correlate negatively with difficulty index (r = -0.09). Negative correlation signifies that with increasing difficulty index values, there is decrease in discrimination index. As the test items get easier, the discrimination index decreases, thus it fails to differentiate weak and good students [13]. Several authors believe that the items of an index difficulty ranging between 0.40 and 0.60 maximize discrimination [12]. In fact, the relationship between the difficulty and discrimination indices values is not linear, but more dome-shaped. Initially, the discrimination power increased with the index difficulty of the items, until it reached a plateau with moderately easy/difficult items and then began to decline with further increase in difficulty index [19]. But, as we explained above, although difficult and easy items do not discriminate among students, they may be useful if the intent of the test is to determine whether the students have all mastered the material, but they contribute little to the test if the intent is to determine which students know the most and which know the least.

In our study, ideal items represented 27.02% of total items with a maximum percentage in the 1st level (the only level with both optimum difficulty and good discrimination). The percentage is 64% in a Pakistani study [1]. This percentage over passing the quarter of items administered in the session should serve as a core of items bank that have to be improved throughout years. In our study, mean alpha of all tests was 0.79 indicating an acceptable reliability. It was 0.91 in a Malaysian study [11]. Mean alpha was under 0.7 in 3/66 exams. Many factors can influence the reliability of test items. For example: length of the test (reduces the chance of guessing, so improves reliability), time limit for the test (increases test anxiety and affects students' performance), difficulty of test item (difficult and easy items induce error so cause low reliability), student's awareness of how they will be assessed (results in better performance of students), the test taker ( perhaps the subject is having a bad day which causes poor performance), the test itself (the questions on the instrument may be unclear, induce error and cause low reliability), testing conditions (there may be distractions during the testing) and test scoring (scores may be applying different standards when evaluating the subjects' responses) [6].

The score distribution of a test is another powerful tool for examining test scores results. The majority of exams (89.39%) were negatively skewed. There can be several reasons for this result, the most desirable one, being that the teaching is effective and that the students are highly motivated. It can also mean that the test is too easy or that a copy of the test or the answer key is circulating among the students. Whatever the case, further investigation is warranted [10]. In the other hand, the positively skewed distributions found in exams are a red flag that indicates that those tests need further analysis. It might mean that the test is too difficult, that the items are poorly written or confusing, that the teaching/learning activities are inadequate, that students' motivation are low, or that the objectives are unrealistic. This type of distribution alerts to investigate the cause of the problem and take corrective action.

We found that 48 of the exams had platykurtic distributions. That is to say that discrimination was better among medium scores. The few leptokurtic distributions indicated that, in the corresponding exams, discrimination was rather among higher and lower scores. The shape desired should always respond to the constructor's

**Table 3 :** Mean indices of examinations per level

| | 1st | 2nd | 3rd | | 4th | | 5th | |
|---|---|---|---|---|---|---|---|---|
| | | | A | B | A | B | A | B |
| **Mean difficutly index** | 056 | 057 | 055 | 057 | 061 | 063 | 065 | 062 |
| **Mean discrimination index** | 036 | 027 | 028 | 026 | 025 | 028 | 028 | 027 |
| **Mean Cronbach's alpha** | 084 | 081 | 083 | 080 | 072 | 081 | 080 | 076 |

purpose: to discriminate between strong, weak or medium level students: either to produce a spread of scores, reflecting differences in students' achievement, or to make all examinees score as high as possible.

The results of this study should initiate a change in the way items are selected for any examination and there should be a proper assessment strategy as part of the curriculum development. Much more of this kind of analysis should be carried out after each examination to identify the areas of potential weakness in our test items and improve the standard of assessment. Furthermore, we believe that every department has a responsibility to provide an item analysis report to teachers following every exam administered. With such a report in hand, teachers will have the information that they need to allow them to work toward improving the quality of their test items. Active steps should be taken to ensure high quality examinations throughout years.

## CONCLUSION

Our examinations had an acceptable internal consistency, as well as a good level of difficulty and discrimination. They tended to facility and their discrimination power concerned basically the students of medium level. Item analysis results can be used as an objective and practical approach for our medical school to evaluate the quality of the examinations and provide instructors with insights on how to improve the quality of the exams and potentially clarify students' misunderstanding of concepts in the future.

**ACKNOWLEDGMENTS**

The authors acknowledge Professor Ahmed Maherzi, the Dean, Professor Rym Goucha Louzir, the Vice Dean, all the members of the assessment office especially its president Professor Kalthoum Kallel, as well as the administration of the FMT for their support in our research.

## References

1. Hingorjo MR, Jaleel F. Analysis of One-Best MCQs: the difficulty index, discrimination index and distractor efficiency. J Pass Med Assoc 2012; 62:142-147.
2. University of Washington. Office of educational assessment. Understanding item analysis. (http://www.washington.edu/oea/services/scanning_scoring/scoring/item_analysis.html). Accessed February 21, 2015.
3. MacAlpine M. A summary of methods of item analysis. CAA Centre Bluepaper 2, University of Luton 2002.
4. Educational Data Systems. Preliminary item statistics using point-biserial correlation and p-values. (http://www.eddata.com/r esources/ publications/EDS _Point_Biserial.pdf). Accessed February 21, 2015.
5. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ 2011; 2:53-55.
6. Meshkani Z, Hossein Abadie F. Multivariate analysis of factors influencing reliability of teacher made tests. IJME 2005; 6:149-152.
7. Kartik A, Neeraj R. Itemized analysis of questions of multiple choice question (MCQ) exam. IJSR 2013; 2:279-280.
8. Tabatabaee M, Bahreyni Toosi M, Derakhshan A, Dalloee M, Gholami H. Analytic assessment of multiple-choice tests. IJME 2003; 2:87-91.
9. Henrysson S. Correction of item-total correlations in item analysis. Psychometrika 1963; 28:211-218.
10. McDonald M. Systematic assessment of learning outcomes: Developing multiple-choice exams. Mississauga: Jones and Bartlett Publishers; 2002.
11. Barman A, Ja'afar R, Rahim F, Noor A. Psychometric characteristics of MCQs used in assessing pahase-II undergraduate medical students of university Sains Malaysia. The Open Medical Education Journal 2010; 3:1-4.
12. Bouzidi L, Jaillet A. Can online Peer assessment be trusted?. Educational Technology and society 2009; 12:257-268.
13. Mitra NK, Nagaraja HS, Ponnudurai G, Judson JP. The levels of difficulty and discrimination indices in type A multiple choice questions of pre-clinical semester 1 multidisciplinary summative tests. IeJSME 2009; 3: 2-7.
14. Yu CH. Talk presented at: SAS Users Group International 26; April 23, 2001; California, CA. (http://www2.sas.com/proceedings/sugi26/p246-26.pdf). Accessed February 21, 2015.
15. Gliem J, Gliem R. Talk presented at: Midwest Research to Practice Conference in Adult, Continuing, and Community Education. October 9, 2003; Ohio, OH. (https://scholarworks .iupui.edu/bitstream/handle/1805/344/Gliem%20&%20..?sequence=1). Accessed February 20, 2015.
16. Jackson S. Research methods: a modular approach. Belmont, CA: Cengage Learning; 2008.
17. Lord F. Applications of item response theory to practical testing problems. New York, NY: Routledge 2008.
18. Theodorsson T, El Shafie K, Al Wardy A, Khan A, Al Mahrezi A, Al Shafaee M. Assessment of family doctors in Oman: getting the questions right preliminary findings of a performance analysis of multiple choice questions. The Internet Journal of Medical Education 2009; 1.
19. Si-Mui S, Rashiah R. Relationship between item difficulty and discrimination indices in true/false type multiple choice questions of a para-clinical multidisciplinary paper. Ann Acad Med Singapore 2006; 35:67-71.