Item difficulty & item discrimination as quality indicators of physiology MCQ examinations at the Faculty of Medicine Khartoum University

*Afraa Musa, Samir Shaheen 2, Abdelmoniem Elmardi 3, Ammar Ahmed1

1 Department of Physiology, 2 Department of Orthopedic Surgery,
3 Department of Anatomy, Faculty of Medicine, University of Khartoum, Sudan

Abstract:

Background: Item analysis is an essential tool used in the evaluation of the quality of MCQ examinations.

Objectives: This study aimed at assessing the quality of items of physiology MCQ examinations at the Faculty of Medicine, Khartoum University, using post-examination item analysis.

Methods: A descriptive, cross-sectional study was carried-out on test statistics reports using Remark-software which provides analysis of students’ responses in the form of Excel file formats. Ten physiology MCQ exams held from September 2015 - September 2016 were included in the study. Each exam paper consisted of 60-80 (five-option) items. Item Difficulty index [DIF I] and Discrimination Index (DI) (using point biserial correlation coefficient/r_{pbis}) were measured as quality indicators.

Results: The mean DIF I of 645 items was (56.01±19.97) & the mean r_{pbis} was (0.37±0.13). The majority of items (63.3%) were of acceptable difficulty; (25.5%) were easy & (11.1%) were difficult. The highest proportion (43.5%) of items were excellent; followed by (30%) good; (16.6%) acceptable and (9.6%) poor discriminating items. There were only 5 negatively discriminating items that constituted 0.8% out of total items. The relationship between difficulty and discrimination indices was dome-shaped with maximal discrimination at acceptable difficulty level.

Conclusion: Generation of high quality examinations with average difficulty and good discrimination power is the outcome of careful construction and meticulous review of exams’ items by our expert staff members. Maximizing exam discrimination is achieved by reducing the number of the very difficult and very easy items.

*Corresponding author: Department of Physiology, Faculty of Medicine, University of Khartoum, Sudan
e-mail: aframusa@gmail.com

Introduction:

Medical educators involved in curriculum planning and development recognize the interplay between assessment and learning, and to a large extent, assessment drives learning. Therefore, developing an appropriate assessment strategy is a key part of effective sustainable curriculum development[6, 21]. Multiple choice question (MCQ) testing is the most efficient form of written assessment, being both reliable and valid by broad coverage of content[3, 4]. It is the preferable objective tool for assessment of knowledge domain used to evaluate under- and post-graduate medical students in almost all medical disciplines[5]. Item analysis is a method of measuring the quality of an examination by looking at its constituent parts (items)[6, 7]. It is a valuable integral component of test development and course assessment that uses statistical methods to provide indicators which test the quality of exams [i.e. item difficulty, item discrimination & distracter analysis] [8, 9].

Item analysis is used to identify potential item problems and what should be done to correct them.
It is extremely essential in creating a pool of valid items to improve the quality of future exams and develop appropriate question bank. Item analysis enables identifying the low achievers; their learning difficulties; and misconceptions which can be corrected by counseling or by modifying the learning methods. On the other hand, it provides a feedback for the teachers on the efficacy of their teaching in order to improve their teaching skills and enhance the students’ learning outcomes. Item analysis is beneficial for both the student and the teacher\textsuperscript{(10, 11)}.

Item Difficulty index (DIF I) is a measure of the proportion of the total examinees who answered an item correctly and is most commonly referred to as the p-value. It reflects easiness of item. DIF I ranges from 0-100% or 0.0-1.0. It is often interpreted as follows: (<30% = too difficult); (30%-70%= recommended, good or acceptable) and (>70%= too easy). DIF I ranges from 50-60% is considered as ideal and recommended by assessment experts \textsuperscript{(6, 8, 12-15)}.

Many methods have been developed to calculate the discriminatory power of individual items; e.g., discrimination index, biserial correlation coefficient, point biserial correlation coefficient, and phi coefficient \textsuperscript{(16)} Item Discrimination Index (DI, Dor d value) is simply the difference between the percentage of high achieving and the low achieving examinees that got the item right. It reflects the ability of the item to distinguish between examinees who are knowledgeable and those who are not. The higher the value of DI, the more able is the item to discriminate between students of higher and lower abilities.

While the discrimination index is a popular and valid measure of item quality, this index is not included in the Remark Software reported item statistics. Instead, Remark Software provides the point biserial correlation \textsuperscript{(6, 12, 13)} Point-biserial correlation coefficient \((r_{pbis})\) is one of several ways to compute item discrimination. It quantifies the relationship between an examinee’s performance on the given item (a categorical variable correct/incorrect answer) and the examinee’s overall assessment score (a continuous variable % score on the examination). A high \(r_{pbis}\) indicates that the students who had high exam scores got the item correct, whereas students who had low exam scores got the item incorrect. The advantage of using discrimination coefficients over the discrimination index (DI) is that every person taking the test is used to compute the discrimination coefficients whereas only 54% (27% upper+27% lower) are used to compute the discrimination index \textsuperscript{(17-19)}.

Point biserial coefficient ranges from –1 to +1. Positive values indicate that high scorers have answered this item correctly. It should be positive for correct answers and negative for distracters. The higher the value, the more discriminating is the item. DI of 1 is ideal as it refers to an item which perfectly discriminates between students of low and high abilities. According to Brown \textsuperscript{(20)} and Crocker & Algina \textsuperscript{(21)}: DI > 0.2 is acceptable and able to discriminate between good and weak students\textsuperscript{(22)}. Items with discrimination values near or less than zero should be removed from the exam. This indicates that low scorers in the exam did better on that item than high scorers students\textsuperscript{(6, 12)}. As a rule of thumb and recommended by Ebel & Frisbie (1986), in terms of discrimination index, (0.40 and greater) are considered very good or excellent items and should be retained; (0.30 to 0.39) are reasonably good but possibly subject to improvement; (0.20 to 0.29) are marginal but acceptable items and need some revision; (<0.2) are considered poor items and need major revision or should be eliminated; and <0 or negative discrimination) are the worst type of items and should definitely be discarded\textsuperscript{(12, 13, 16)}. Items should be carefully reviewed for the presence of common causes of poor discrimination such as ambiguous wording, grey areas of opinion, wrong keys and areas of controversy \textsuperscript{(1)}.

In our institution’s curriculum, the weight of MCQ as an assessment tool in the physiology subject is substantial and accounts for more than half of the students’ course grades. The current study was
conducted with the objective to test the standard of this principal determinant of the physiology exams’ scores using post-examination analysis of the test items. This was accomplished by calculating the two most common statistical parameters reported in item analysis (the difficulty index & the point-biserial correlation coefficient as the discriminative index) as well as the relationship between the two indices.

**Methods:**

The study was descriptive, cross-sectional conducted at the Physiology Department, Faculty of Medicine, University of Khartoum. The study recruited item analysis reports of the physiology MCQ exams of 2nd, 3rd & 4th semesters of 2 batches of medical students. The study included mid-semester & end-of-semester exams held during the period from September 2015 - September 2016. The questions covered all physiology systems except the central nervous system. Each exam paper consisted of 60-80 (five-option) items that comprised of a stem, one correct answer and four distractors. The total number of items was 650 out of which 5 items were excluded before marking the answer sheets because they were wrongly constructed with 2 correct options. The remaining 645 items included: 3225 options (645 correct responses & 2580 distracters). The number of the examinees for each paper ranged from 332–359.

The Remark software which was used for marking the students answer sheets provided detailed statistical analysis of students’ scores, reliability of the exam, item difficulty index, point-biserial correlation as a measure of item discrimination and a detailed distracter analysis (14). The software analyzed the data using Microsoft Excel program that provides the results of performance analysis in different excel sheets for each individual exam. The machine was adjusted so that a correct answer was awarded 1 mark and there were no negative marks for the incorrect answers. Data analysis was carried-out using the software Statistical Package for the Social Sciences (IBM SPSS statistics). P value <0.05 was considered as significant.

The research study was approved by the Research Ethics Committee of the Faculty of Medicine, Khartoum University. As the software provides anonymous data [i.e. deals with students IDs rather than examinees’ names], there was no need for full board review and no need for informed consent from the examinees.

**Results:**

Regarding the performance of the individual ten exams: the number of the examinees ranges from 332–359; the number of test items ranges from 60 in mid-semester exams to 80 in end-of-semester exams. The mean DIF I ranges: 48.62-65.67 and the mean \( r_{\text{phis}} \) ranges 0.32-0.47. The proportions of items with different levels of difficulty and different discriminating abilities are summarized in Table 1.

The mean DIF I of 645 items was 56.01 (SD=19.97) and mean \( r_{\text{phis}} \) was 0.37 (SD=0.13). Analysis of the level of difficulty of 645 items revealed that the majority of our items (63.3%) were in the acceptable difficulty level (DIF I=30-70%); (25.5%) were easy (DIF I>70%) and (11.1%) were difficult items (DIF I<30%) as shown in Figure 1.

With regard to discriminating ability categories: analysis of point biserial correlation coefficient showed the highest proportion (43.5%) of items had excellent discrimination \( (r_{\text{phis}} \geq 0.4) \) followed by (30%) having good \( (r_{\text{phis}} =0.3-0.39) \) and (16.6%) having acceptable discrimination \( (r_{\text{phis}} =0.2-0.29) \) while (9.6%) had poor discriminating ability \( (r_{\text{phis}} <0.2) \) as shown in Figure 2.

Analyzing the two indices together revealed that: 384 items (60%) were considered ideal i.e. having acceptable DIF I (30-70%), as well as \( r_{\text{phis}} (>0.2). In addition, the majority of items with acceptable difficulty (93.89%) had discrimination \( (r_{\text{phis}} =0.2-0.29) \). It was noted that a high proportion of easy (55.8%) and acceptable difficulty (45.7%) items had excellent discrimination. In contradistinction, most of the difficult items (44.4%) were poorly discriminating, and barely contributed in the category of excellent discriminating items (2.8%) as shown in Table 2.

The scattered diagram shown in Figure 3, describes
the relationship between the DIF I and \( r_{pbis} \) of 645 items. It was not linear, but rather dome-shaped. Correlation statistics revealed that \( r_{pbis} \) significantly correlated positively with the DIF I (\( r=0.461, P<0.001 \)). The discrimination was noted to be maximal at the upper range of acceptable difficulty (50-70%); minimally reduced at the level of very easy items but marked reduction was apparent at the level of very difficult items. It was observed that out of 646 items, only five (0.8%) had negative discrimination indices values ranging from 0.21 to 0.03 with corresponding range of DIF I(2.30% -16.47%), indicating their high difficulty level.

### Table 1. Item difficulty and item discrimination of the individual exams

<table>
<thead>
<tr>
<th>Exam</th>
<th>Examinee No</th>
<th>Items No</th>
<th>Mean DIF I</th>
<th>Items Difficulty Level</th>
<th>Discriminating Ability Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mean difficult items</td>
</tr>
<tr>
<td>1. Mid S2</td>
<td>342</td>
<td>60</td>
<td>53.63</td>
<td>8.3</td>
<td>78.3</td>
</tr>
<tr>
<td>September</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Mid S3</td>
<td>347</td>
<td>60</td>
<td>54.99</td>
<td>8.3</td>
<td>66.7</td>
</tr>
<tr>
<td>September</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. End S2</td>
<td>342</td>
<td>70</td>
<td>59.35</td>
<td>7.1</td>
<td>58.6</td>
</tr>
<tr>
<td>November</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. End S3</td>
<td>346</td>
<td>80</td>
<td>50.83</td>
<td>20.3</td>
<td>64.6</td>
</tr>
<tr>
<td>November</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Mid S3</td>
<td>339</td>
<td>60</td>
<td>48.62</td>
<td>17.2</td>
<td>67.2</td>
</tr>
<tr>
<td>February</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Mid S4</td>
<td>350</td>
<td>60</td>
<td>65.67</td>
<td>6.8</td>
<td>40.7</td>
</tr>
<tr>
<td>February</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. End S3</td>
<td>336</td>
<td>60</td>
<td>53.63</td>
<td>10.0</td>
<td>76.7</td>
</tr>
<tr>
<td>April 2016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. End S4</td>
<td>353</td>
<td>80</td>
<td>62.15</td>
<td>6.3</td>
<td>60.0</td>
</tr>
<tr>
<td>April 2016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Mid S2</td>
<td>332</td>
<td>60</td>
<td>58.44</td>
<td>11.7</td>
<td>60.0</td>
</tr>
<tr>
<td>September</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Mid S4</td>
<td>359</td>
<td>60</td>
<td>53.80</td>
<td>15.0</td>
<td>61.7</td>
</tr>
<tr>
<td>September</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
This table summarizes the performance of the ten exams showing the number of the examinees, the number of items, the mean DIFI and the mean $r_{pbis}$ in addition to the proportions of items with different levels of difficulty and discriminating abilities in each individual exam.

Table 2. Discrimination ability of items based on the level of their difficulty*

<table>
<thead>
<tr>
<th>Discriminating Ability based on Difficulty Level</th>
<th>Frequency (Percent) of Difficult items (DIFI&lt;30%)</th>
<th>Frequency (Percent) of Acceptable items (DIFI 30-70%)</th>
<th>Frequency (Percent) of Easy items (DIFI &gt;70%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor discrimination (&lt;0.2)</td>
<td>32 (44.4%)</td>
<td>25 (6.13%)</td>
<td>5 (3.0%)</td>
</tr>
<tr>
<td>Acceptable discrimination (0.2-0.29)</td>
<td>24 (33.3%)</td>
<td>57 (13.97%)</td>
<td>28 (17.0%)</td>
</tr>
<tr>
<td>Good discrimination (0.3-0.39)</td>
<td>14 (19.4%)</td>
<td>140 (34.31%)</td>
<td>40 (24.2%)</td>
</tr>
<tr>
<td>Excellent discrimination (&gt;=0.4)</td>
<td>2 (2.8%)</td>
<td>186 (45.58%)</td>
<td>92 (55.8%)</td>
</tr>
<tr>
<td>Total</td>
<td>72 (100%)</td>
<td>408 (100%)</td>
<td>165 (100%)</td>
</tr>
</tbody>
</table>

*(Total number of items=645)

Discussion:

In this part of the paper the significance of the various parameters of item analysis in exam evaluation will be discussed as follows: mean DIF I, mean discrimination coefficient of total items and of individual exams, proportion of items with; recommended difficulty (DIFI=30-70%), recommended discrimination (>0.2) and

**Figure 1. Proportions of items with various levels of difficulty** *(Total number of items=645)*

**Figure 2. Proportions of items with different discrimination abilities.** *(Total number of items=645)*

**Figure 3. The non-linear relationship between difficulty index & discrimination coefficient (point biserial correlation) of total items (645)**

*note: discrimination ability was maximally corresponding with level of acceptable difficulty items, slightly lowered by the presence of easy items and maximum lowering occurred at the level of very difficult items.
ideal items (fulfilling recommended difficulty and discrimination).

**Difficulty Indicators:**

Our finding of mean DIF I (56.01 ±19.97) of total items corresponded with some previous studies that reported similar figures; (0.54±0.26)\(^{(22)}\), (54.14±17.48)\(^{(23)}\), (52.53±20.59)\(^{(24)}\), (0.55±0.2) \(^{(25)}\), (57.92±19.58) \(^{(26)}\) and (55.9±15.7) \(^{(10)}\). The closer the facility value of a question to 0.5 (DIF I 50%), the more it is contributing to the measurement of the candidates\(^{(6)}\). This mean value is ideal as it reflects the abundance of items with average difficulty which is recommended by medical educators.

On the other hand, some studies revealed high mean DIF I; (61.75±28.20) \(^{(27)}\), (63.06±18.95) \(^{(7)}\), (65.92±22.2) \(^{(28)}\) and (0.82±0.18) \(^{(29)}\) indicating easiers. In contradistinction, few researches on community medicine MCQ exams reported low mean DIF I; as (39.4±21.4%) \(^{(30)}\) and (38.34±22.49) \(^{(31)}\) indicating difficult exams\(^{(31)}\).

The finding that the majority of our items (63.3%) were of acceptable difficulty, 25.5% were easy and 11.1% were difficult items (Figure 1) agreed with two studies. The first study conducted by Pandey et al on four physiology exams, revealed (62%) of 240 items were within acceptable range; (23%) were easy and (15%) were difficult \(^{(24)}\). The second by Kolte on 40 items of physiology examination showed 65% of items in the acceptable difficulty level; 25% were very easy and 10% were very difficult \(^{(28)}\). Many other studies revealed higher proportions of acceptable difficulty items ranging from 74-76% (2,5,10,23,25,32). These results were expected because most of them were performed on small number of items. On the other hand, few studies revealed lower proportions (56%\(^{(33)}\) and 45%\(^{(22)}\)) of acceptable difficulty items and, therefore, higher proportions of very difficult and very easy items.

Questions which are too easy or too difficult for a student are not recommended as they contribute little information regarding the student’s ability \(^{(17,19)}\). Although too difficult items result in frustration and deflated scores, they can serve as an effective feedback to the departments with regard to the effectiveness of teaching. It should be considered as a warning sign of inappropriate topic at this stage of students’ training, misconception, or improper teaching. On the other hand, too easy items result in inflated scores and decline in motivation \(^{(1,9)}\). They should be investigated for the cause of easiness like the presence of non-functioning distracters which should be replaced by more plausible distracters.

Regarding studies that analyzed multiple exams, our ideal range of mean DIF I (48.62- 65.67) of the ten exams (Table 1), were very similar to that of a study conducted by Karelia and colleagues and which included ten pharmacology tests (200 items). They reported a range of mean DIF I of individual tests between 47.17- 58.08% \(^{(34)}\). Another study by Mitra et al on twelve tests (120 items) revealed easier exams, indicated by higher range of mean DIF I(64-89%) of individual tests \(^{(11)}\). Evaluation of two paediatric MCQ exams revealed easy exams indicated from their high mean DIF I(0.67-0.79), which was considered by the authors at optimal level of difficulty \(^{(38)}\).

**Discrimination indicators:**

Our study finding of high mean \(r_{pbi} \) (0.37±0.13) of 645 items, was similar to a mean DI of 0.37±0.15\(^{(5)}\), but slightly higher than other studies’ mean DI; (0.33±0.18)\(^{(7)}\) & (0.30±0.18)\(^{(24)}\). Moreover, some studies reported a much lower mean DI (0.27± 0.28) \(^{(31)}\), (0.25±0.14) \(^{(36)}\), (0.21±0.14) \(^{(22)}\), (DI=0.24±0.2) \(^{(28)}\) and (0.14±0.19) \(^{(30)}\). Our high mean discrimination coefficient reflects abundance of items with high discriminating abilities. Our very high proportion of discriminating items (90.4%) using 0.2 as cut off values for \(r_{pbi} \) represented one of the highest proportions compared to the available literature that ranges from 86% down to 44% (5, 7, 11, 22, 24, 25, 27, 31, 34, 36, 37), though few of these researchers had used lower cut off values (0.1 \(^{(37)}\) & 0.15 \(^{(38,31)}\) for DI or \(r_{pbi} \).

Despite studying a very large sample size (645 items), we have only 5 negative discriminating items (0.8%) compared to high proportions ranging from 4-23% reported by other studies\(^{(22,24,25,27,30,31,37)}\). Very few studies reported that none (zero %)
of their items were negatively discriminating \(^{(5,7,26)}\), each of them consisted of a small number of items (range 40-50). Negative discrimination index is a sign of poorly constructed, misleading or mis-keyed item \(^{(13)}\). In addition to being useless, negatively discriminating items decrease the validity of the test \(^{(12)}\).

Using different cut-off values for dividing discrimination abilities into: excellent, very good, good, acceptable, poor...etc.), wide variation were reported in the literature \(^{(5, 7, 10, 22, 24-29, 31-33, 36-38)}\). The distribution of discrimination abilities of our 645 items was appropriate (Figure 2), showing the highest proportion (43.5%) of excellent discriminating followed by (30%) good discriminating, (16.6%) acceptable discriminating and (9.6%) poor discriminating items. The high discriminating ability of our items was better than the results of a large Canadian study by Di Battista & Kurzawa. They analyzed sixteen classroom tests which revealed that about 15% out of 1198 items had excellent discrimination coefficients \((r_{pbis} > 0.40)\); more than 30% had poor discriminating coefficients \(< 0.20\) and 4% had negative coefficients. Furthermore, the mean \(r_{pbis}\) across the 16 tests ranged from 0.33 down to 0.20, was lower than our ten exams range of mean \(r_{pbis}\) (0.32-0.47) \(^{(36)}\). All of our exams (Table 1), have mean \(r_{phi} (0.3)\), five exams (50%) have mean \(r_{phi} (0.35)\) out of which 2 (20%) have mean \(r_{phi} (0.4)\). These findings are generally similar to Mitra et al study on twelve tests which reported that 7 out of 12 tests had mean DI \((0.4)\), and only one test showed poor mean DI, indicating that these MCQ items were excellent for differentiating between poor and good performers \(^{(11)}\). In addition, our exam items were more discriminating than Karelia et al who reported only three out of ten (30%) tests that had mean DI \((0.35)\) \(^{(34)}\).

**Relation between Difficulty Index and Discrimination Index: (Table 2 & Figure 3)**

Most of the researches that studied the relation between difficulty index and discrimination index, supported our finding that the relation is not linear but rather dome or pyramidal shaped \(^{(1, 24, 29, 36)}\). Discrimination was noted to be maximum in the upper range of acceptable difficulty (50-70); slightly lower in the very easy items and maximal lowering occurs at the level of the very difficult items. Our findings corresponded with the Canadian study findings which showed strong curvilinear relation with both linear and quadratic components and very low mean discrimination coefficient for the most difficult items; higher for items with DIFI \((0.30-0.89)\), and somewhat lower for the easiest items \(^{(36)}\).

Again the positive correlation between discrimination and difficulty indices \((r=0.19, P=0.003<0.01)\) obtained by Pande et al \(^{(24)}\); \((r=0.3076, P=0.05)\) by Shete et al \(^{(22)}\) and \((r=0.11, P>0.10)\) by Karelia et al \(^{(34)}\) was similar to our highly significant positive correlation results \((r=0.461, P=0.000<0.001)\). Mehta et al confirmed the statistically significant association using chi-square test \((p-value=0.0001<0.05)\) \(^{(2)}\). In contrast with our results, was the negative correlation \((r=-0.325)\) reported by Mitra et al\(^{(11)}\).

In this study, 60% out of 645 items were classified as ideal items [i.e. having acceptable DIF I (30-70%), as well as \(r_{pbis} (>0.2)\)]. This proportion was similar to the 64% observed by Hingorjo & Jaleel using similar cut off values \(^{(23)}\). On the other hand, it was higher than the 54% reported by Mehta et al \(^{(2)}\) who used higher cut off values for DI \((>0.25)\). Again, it was much higher than (24%) ideal items reported by Mehta & Mokhasi, who used narrower ranges of DIF I (50-60%) and much higher cut off value for DI \((>0.35)\) \(^{(7)}\). Again Di Battista & Kurzawa showed more than 80% of their items with a DIF I \((0.50-0.80)\) had a discrimination coefficient of at least 0.20 \(^{(36)}\). This confirms one of the basic principles of item response theory, which postulates that “the questions which are at the same level of average candidate’s ability are the most effective to assess and discriminate them” \(^{(2)}\).

**Conclusion:**

This study is considered as one of the few large studies that addressed the issue of item analysis of MCQ exams in the region and the first in the Sudan. It emphasized the high quality of the physiology
MCQ exams in the Faculty of Medicine, University of Khartoum. The examined items satisfied the criteria of recommended difficulty and high discrimination ability. Improvement of overall test discrimination could be achieved by reducing the number of the extremely difficult items and, to a lesser extent, the easy ones.

**Recommendations:**
Although generalization of the findings from this study may be limited, the study was meant to throw light on the importance of item analysis in evaluation, quality assurance and continuous renewal of our assessment. It is expected to raise the awareness of other universities, faculties and departments to the importance of including item analysis results in the routine exams’ evaluation in the academic boards. To achieve this, faculties through their educational development units should provide training workshops to the staff members on how to interpret effectively reports of item analysis.

**Acknowledgements:**
The researchers acknowledge the support provided by the academic and administrative staff of the Physiology Department, the Educational Development Centre and the Dean’s Office at the Faculty of Medicine of Khartoum University for their great help and support in data acquisition.

**Conflict of Interest:**
The authors declare that they have no conflict of interest.

**References:**
15. Ananthakrishnan N. Item analysis-validation


25.Elfaki OA, Bahamdan KA, Al-Humayed S. Evaluating the quality of multiple-choice questions used for final exams at the Department of Internal Medicine, College of Medicine, King Khalid University. Sudan Medical Monitor. 2015;10:123-7.


33.Natekar PE, DeSouza F, Karapurkar N. Errors in


