Quality assessment of a multiple choice test through psychometric properties

Ana Livia Licona-Chávez[1], Pierangeli Kay-to-py Montiel Boehringer[1], Lupita Rachel Velázquez-Liaño[1]

Corresponding author: Dr Pierangeli Kay-to-py Montiel Boehringer kmontiel@xochicalco.edu.mx
Institution: 1. Universidad Xochicalco, campus Ensenada
Categories: Assessment, Comparative Medical Education, Learning Outcomes/Competency, Teaching and Learning

Abstract

Introduction: Instruments for Multiple Choice Questions (MCQs) assessment in health sciences must be designed to ensure validity and reliability. The present paper assesses the quality of a MCQs test in Research Methodology at the Faculty of Medicine at Xochicalco University. It establishes the basis to improve the quality of MCQs and is intended as a baseline for other universities.

Methodology: The peer-reviewed test had 20 MCQs with three distractors with and a single correct response and 89 students took the exam. The tests were graded and analyzed to estimate difficulty index (DIF I), discrimination index (DI) and distractor efficiency (DE). Cronbach’s alpha and ANOVA were calculated with SPSS.

Results: The resulting DIF I (0.49) indicates that the test was moderately difficult, mean discrimination index of 0.25 means that the test has a regular quality to discriminate between skilled and unskilled students and needs to be checked, only 20% of the items were considered excellent and 5% were considered good questions. Alpha coefficient was 0.898 considered good for a MCQs assessment. ANOVA results showed no significant differences between groups.

Discussion and Conclusion: This test shows a high percentage of moderately difficult questions, is unbalanced and a reduction of Non Functional Distractors (NFDs) is needed. However the topic learned has the same standard across the different groups of students.

Keywords: MCQs; Difficulty; Discrimination; Distractor efficiency; Cronbach's alpha; Alpha if item deleted.
Introduction

Empirically developed learning assessment tools limit the proposed educational goals and do not objectively reflect the level of achievement through the grades obtained (Ortiz-Romero et al., 2015). Since the incorporation of multiple choice questions MCQs) for medical testing by the National Board of Medical Examiners (Hubbard, 1978), several guidelines have been created in order to properly construct these kind of questions and to evaluate higher cognitive processing (Haladyna, Downing and Rodríguez, 2002). Research groups mainly from areas of educational psychology and educational assessment focused on the study of MCQs test (Hancock, 1994, Martínez, 1999) to validate the theoretical intentions in each learning domain.

MCQs tests have some advantages over other assessing strategies: they allow the educators to cover a wide range of educative material that was given in a short period of time and they are especially useful evaluating large population of students. Moreover, MCQs test allows the construction of multiple test versions to control cheating and the academic feedback returned to the students becomes easy (Simkin and Kuechler, 2005). On the other hand, statistical information can easily be obtained in order to determine the class performance on a particular question and to asses if the question was appropriate to the context that was presented (Carneson et al., 2016).

Instruments for MCQs assesment in health sciences (Durante et al., 2011) must ensure validity and reliability to be scientifical and practical. Validity is the degree to which it measures what it is supposed to measure and reliability is a statistical concept that represents that the scores obtained by the students will be similar if they were assessed again; it means that the construct measured is consistent across time. Reliability can be measured by a statistical method for internal consistency called Cronbach's alpha. It ranges from 0 to 1, where zero is a no-correlation between the items and one represents a significant covariance. Is suggested that a Cronbach's alpha coefficient of 0.8 or more is desired for high-stakes in-house exams and that reliability can be improved by increasing the number of items given in an exam (Ali, Carr and Ruit, 2016).

MCQs are characterized by high validity and reliability if they are apropiately constructed (Ware and Vik, 2009). Even though medical schools must review the quality of MCQs to ensure their validity and reliability, in México, only the Faculty of Medicine of the National Public University (UNAM) has published these kind of studies (Delgado-Maldonado and Sánchez-Mendiola, 2012; Martínez, et al., 2014; Saldaña, Delgadillo and Méndez, 2014; Borrego-Mora and Santana-Borrego, 2015). These were done to gather enough evidence for the validity and reliability of different assesments using different psychometric analyses. Through them, they were able to identify flaws in their MCQs, to feedback faculty programs and improve the length and quality of their high stake exams.

Item analysis show which questions are good and those that need improvement or to be discarded (Mitra et al., 2009), based on difficulty and discrimination indexes and the distractor efficiency (Sahoo and Singh, 2017). Difficulty index (DIF I) describes the percentage of students who correctly answered the item. It ranges from 0-100%. The higher the percentage, the easier the item and viceversa (Sahoo and Singh, 2017). The recommended range for an ideal difficulty balanced exam is 5% for easy items, 5% for difficult ones, 20% for moderately easy, 20 % for the moderately difficult and 50% for the average ones (Backhoff, Larrazolo and Rosas, 2000). Despite the importance of MCQ tests, no recent publications have been found about this topic. Discrimination index (DI) explains the ability of an item to differentiate between high and low scorers. It ranges between -1.00 and +1.00. Items with higher value of DI better discriminate between students of higher and lower abilities. A highly discriminating item indicates that the students who had high tests scores got the item correct whereas students who had low test scores got the item incorrect (Boopathiraj and Chellamani, 2013; Sahoo and Singh, 2017). When a distractor (DE) distracts few or no testee, it means that it is a poor distractor and should be reviewed (Odukoya et al., 2017). An effective distractor is the one chosen by ≥5% of the students (Kaur, Singla and Mahajan., 2016; Sajitha et al., 2015; Ware and Vik, 2009). DE is determined for each item on the basis of the number of nonfunctional distracter (NFD), if the option is selected by <5% of students (Kaur, Singla and Mahajan., 2016).
Another psychometric property measured is alpha if item is deleted (Pell et al., 2010), this measure indicates that since Cronbach’s alpha tends to increase with the number of items in an assessment, the resulting alpha if one item is deleted should be lower than the overall alpha score, if this item has performed well. Where this is not the case, this may be caused by any of the following reasons: the item is measuring a different construct to the rest of the items or the item is poorly designed or the assessors are not assessing to a common standard or there are teaching issues (the topic tested has not been well taught or has been taught to a different standard across different groups of students).

For medical schools it is important to guarantee the quality of education and therefore the quality of MCQs that are applied. The Faculty of Medicine at Xochicalco University in Ensenada, Baja California, México, has a group of Research Methodology teachers that have been working since 2017 in peer-reviewed MCQs. The objective of this research was to assess the quality of a MCQs exam in a course of 2nd semester of Research Methodology subject in 2019-1, which represents 40% of the grade of each partial evaluation. This research will establish the basis to improve the quality of MCQs and the results will be proposed as a base line for the scientific analysis of multiple choice tests within the faculty itself in its different areas, as well as in other medicine faculties.

Methods

To create the test, the academic program content corresponding to the research methodology subject was divided between lecturers of the 2nd semester and each one provided MCQs to ensure that all the topics were covered. In order to have a consensual evaluation of the test, each item was peer-reviewed and the grammatical or syntax errors were corrected. The test had 20 MCQs with three distractors and a single correct response. A total of 89 students took the test the same day. There were two different versions of the test with items in different order, with topics such as problem statement, causality study and protocol design. The different tests were graded, data obtained was entered in Microsoft Excel 2016 and analyzed to calculate DIF I, DI and DE.

The following formulas of Backhoff et al. (2000) were used to find DIF I and DI:

\[
\text{DIF I} = \frac{C_i}{N_i}
\]

where:

DIF I = difficulty index of item

\[
C_i = \text{number of right answer of item}
\]

\[
N_i = \text{number of right answer plus number of wrong answer}
\]

\[
\text{DI} = \frac{H - L}{N}
\]

where:

DI = discrimination index of item

\[
H = \text{number of students in the upper group who responded correctly}
\]

\[
L = \text{number of students in the lower group who responded correctly}
\]

\[
N = \text{Number of students in the largest group}
\]

The criteria used to categorize the difficulty and discrimination indexes are presented in Table 1 and 2.
Table 1. Difficulty index

| Difficulty index | Category       |
|------------------|----------------|
| > 0.8            | Easy           |
| 0.71 a 0.8       | Moderately easy|
| 0.51 a 0.7       | Average        |
| 0.31 a 0.5       | Moderately difficult |
| < 0.31           | Difficult      |

Taken from Backhoff, E., Larrazolo, N. and Rosas, M. (2000) ‘Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA)’, Revista Electrónica de Investigación Educativa, 2(1), pp. 11-29. [http://redie.uabc.mx/vol2no1/contenido-backhoff.html](http://redie.uabc.mx/vol2no1/contenido-backhoff.html) (Accessed: 15 Jul 2019).

Table 2. Discrimination index

| Discrimination index | Quality | Recommendations                  |
|----------------------|---------|----------------------------------|
| > 0.39               | Excellent| Keep                             |
| 0.30 a 0.39          | Good    | Need improvement                 |
| 0.20 a 0.29          | Regular | Check                            |
| 0.00 a 0.20          | Poor    | Discard or need deep evaluation  |
| < 0.01               | Very poor| Discard definitively              |

Taken from Backhoff, E., Larrazolo, N. and Rosas, M. (2000) ‘Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA)’, Revista Electrónica de Investigación Educativa, 2(1), pp. 11-29. [http://redie.uabc.mx/vol2no1/contenido-backhoff.html](http://redie.uabc.mx/vol2no1/contenido-backhoff.html) (Accessed: 15 Jul 2019).

The criteria used to categorize the distractor efficiency is presented in Table 3.

Table 3. Distractor efficiency criteria

| NFD | Distractor efficiency |
|-----|-----------------------|
| 3   | 0%                    |
| 2   | 33.3%                 |
| 1   | 66.6%                 |
| 0   | 100%                  |

Taken from Sahoo, D. P. and Singh, R. (2017) ‘Item and distracter analysis of multiple choice questions (MCQs) from a preliminary examination of undergraduate medical students’, International Journal of Research in Medical Sciences, 5(12), pp. 5351-5355. [http://dx.doi.org/10.18203/2320-6012.ijrms20175453](http://dx.doi.org/10.18203/2320-6012.ijrms20175453)

The Cronbach’s alpha, the resulting alpha if item deleted scores and one-way analysis of variance (ANOVA) were calculated by the Statistical Package for Social Sciences (SPSS) 22. To analyse the differences between mean groups A, B y C, an ANOVA P value of <0.050 was considered statistically significant.

Ethics approval was granted by our medical school Ethical Committee.
Results/Analysis

Results

Difficulty index

Out of 20 items, 1 (5%) was easy (DIF I >0.8), 3 (15%) were moderately easy (DIF I 0.71-0.8), only 2 (10%) were average (DIF I 0.51-0.7), 11 (55%) were moderately difficult and 3 (15%) were difficult (Figure 1). Mean DIF I of the exam was 0.49 (S.D. 0.18).

Figure 1. Difficulty comparison between expected balance and this exam

Discrimination index

Mean discrimination index of the exam was 0.25 (S.D. 0.16). Four (20%) out of the 20 test items had a discrimination index higher than 0.39, one (5%) showed good discrimination power (0.30-0.39) and 7 (35%) showed regular discrimination (0.20-0.39), nevertheless, 8 (40%) showed poor discrimination power (0.0-0.20) (Figure 2).

Figure 2. Quality and recommendations for items according to the discrimination index
Distractor analysis

A total of 20 items had 60 distractors, among these, 18 (30%) were nonfunctional and 42 (70%) of the distractors were considered functional. Five items had a 100% distractor efficiency (DE), 12 had 66% and 3 had 33% (Table 4).

Table 4. Distractor analysis

| Description                | Percentage of Items |
|----------------------------|---------------------|
| To keep                    | 20%                 |
| Need improvement           | 5%                  |
| To check                   | 35%                 |
| To discard or need deep evaluation | 40%              |
| Excellent (~0.39)          | 4                   |
| Good (0.30 - 0.39)         | 1                   |
| Regular (0.20 - 0.29)      | 7                   |
| Poor (0.00 - 0.20)         | 8                   |
| Very poor (< 0.01)         | 0                   |

Reliability Coefficient and alpha de Cronbach if item deleted

The alpha score of the assessment was 0.898. The resulting alpha if item deleted, that were below the overall alpha score, were 75% of the items. Questions 1, 7, 11, 12 and 14, (25% of the exam) obtained the same score or above 0.898. Questions 2, 6, 8, 9, 13, 14, 17 and 18 got a low discrimination index, but their alpha if item deleted were below the alpha score. Only item 14 has a low discrimination index and a score above the Cronbach alpha's, was too
easy and had one NFD. On the other hand, item 20 had a discrimination index that needs improvement, an alpha score if item is deleted below 0.898, is considered difficult and has no NFDs (Table 5).

### Table 5. Item metrics

| Item number | Difficulty | Discrimination | DI Quality | No functional distractors NFD’s | Cronbach’s alpha if item deleted |
|-------------|------------|----------------|------------|---------------------------------|----------------------------------|
| 1           | 0.78       | 0.29           | regular    | 2                               | 0.903                            |
| 2           | 0.34       | 0.04           | poor       | 1                               | 0.896                            |
| 3           | 0.46       | 0.29           | regular    | 2                               | 0.886                            |
| 4           | 0.21       | 0.21           | regular    | 1                               | 0.895                            |
| 5           | 0.75       | 0.54           | excellent  | 1                               | 0.897                            |
| 6           | 0.37       | 0.17           | poor       | 0                               | 0.892                            |
| 7           | 0.56       | 0.29           | regular    | 1                               | 0.898                            |
| 8           | 0.47       | 0.17           | poor       | 1                               | 0.886                            |
| 9           | 0.48       | 0.17           | poor       | 1                               | 0.885                            |
| 10          | 0.21       | 0.21           | regular    | 1                               | 0.895                            |
| 11          | 0.77       | 0.54           | excellent  | 1                               | 0.901                            |
| 12          | 0.64       | 0.50           | excellent  | 1                               | 0.913                            |
| 13          | 0.5        | 0.13           | poor       | 2                               | 0.886                            |
| 14          | 0.88       | 0.08           | poor       | 1                               | 0.903                            |
| 15          | 0.45       | 0.21           | regular    | 0                               | 0.889                            |
| 16          | 0.36       | 0.21           | regular    | 0                               | 0.892                            |
| 17          | 0.46       | 0.13           | poor       | 0                               | 0.885                            |
| 18          | 0.44       | 0.17           | poor       | 1                               | 0.889                            |
| 19          | 0.4        | 0.42           | excellent  | 1                               | 0.889                            |
| 20          | 0.3        | 0.33           | good       | 0                               | 0.884                            |
| Average     | 0.49       | 0.25           | regular    | 0.9                             | 0.89                             |

**ANOVA**

The averages obtained in two of the groups were accredited (upper than 70) and in the third group not accredited, however, the ANOVA did not show significant differences between the averages of the three groups (Table 6).

### Table 6. ANOVA comparison between groups

| Group | N   | Average | p     |
|-------|-----|---------|-------|
| A     | 30  | 69.67   |       |
| B     | 29  | 72.07   |       |
| C     | 30  | 77.33   | 0.60* |
| Total | 89  | 73.03   |       |

*non significant

**Discussion**

According to Backhoff *et al.* (2000) the mean DIF I (0.49, S.D.=0.18) of the exam applied for the present research, means it was moderately difficult. Even more, 55% of the items are classified as moderately difficult. The mean DIF I is lower compared to that obtained for MCQ test by Backhoff *et al.* (2000) in a high stake examination (DIF I=0.56), and also lower compared with Rao *et al.* (2016) (DIF I= 0.75) for a pathology MCQ test.
Results of DIF I differ from the criteria suggested by Backhoff et al. (2000) for a balanced exam, since a higher percentage of moderately difficult questions (55% vs 20% expected) and moderately easy (15% vs 20% expected) was obtained. The items with average difficulty were lower than expected (10% vs 50% expected) and the items easy (5%) were as expected. However, the exam did not have the 5% of difficult items suggested by Backhoff. These results indicate that some questions must be reviewed in order to balance the exam according to the suggested criteria.

Mean discrimination index was 0.25 (S.D. 0.16) which means that the test has a regular quality to discriminate between skilled and unskilled students and needs to be checked. None of the items showed a very poor DI, which is an advantage because a negative DI value indicates the presence of ambiguous questions or questions with the answer wrongly marked, so none of the analyzed items were definitely discarded. According with the criteria used, 20% of the items were considered as excellent and 5% were considered good questions, indicating that they discriminate among students. Nevertheless, 75% of the items could be rejected due to their poor discrimination power (<0.20), so deep evaluation of these items should be conducted. Other studies (Kaur, Singla, and Mahajan, 2016) that use similar criteria to interpret results, discarded all the questions with DI under 0.20 without further evaluation. On the other hand, seven of the eight items with poor DI were considered moderately difficult according to their DIF and six of them had at least one NFD. It has been seen that the number of NFD disturb the discrimination power of the MCQs, items with lower NFDs correlate with good or excellent DI (Rao et al., 2016; Kheyami et al., 2018).

Among the different parameters used in this study, the DI is the most accurate because it takes into account all the questions as well as all the students for the evaluation (Saldaña et al., 2014). In this study, 20% of the items were kept for subsequent use and the other 80% still required further improvement.

Even though this test has only 20 MCQs, its alpha coefficient is 0.898 which is good for a MCQs assessment. The high consistency in this first analysed exam was probably because the questions were peer reviewed. Tavakol and Dennick (2012) suggested in the AMEE guide 66 the use of alpha-if-item-deleted for high-stake examinations. Although the Cronbach's alpha if item deleted is not commonly used to measure MCQs quality, in the present study it was helpful to measure the individual items. The Cronbach's alpha if item deleted was the same or above overall alpha score in 5 items (1, 7, 11, 12, and 14), which means that if they are taken away, the exam will get the same or a better Cronbach's alpha. Item 14 had three metrics that indicate it as a flaw question: a low discrimination index, an above Cronbach's alpha if item deleted and one NFD. There were also items with 1 or 2 NFDs whose Cronbach's alpha if item deleted was below 0.898, although it is known that careful creation and selection of distractors reduces the cueing effect and improves the MCQ tests (Ali et al., 2016).

ANOVA results obtained show no significant differences between groups. This could indicate that although each group has a different teacher, learning level seem to be similar between groups, probably due to common academic planning and schedules between Research Methodology teachers in our medical faculty.

Conclusion

This paper is the first psychometric analysis carried out of a MCQ test in a Medical school in northwest Mexico. This MCQ exam showed a high percentage of moderately difficult questions and lacked of average questions, resulting in an unbalanced test, even though, the items were previously peer reviewed. Further analysis must be done in order to increase the percentage of MCQs with an average DIF I. On the other hand, the test has a low discrimination index, only 20% of the items can be kept, the rest need deeper evaluation. To improve the discrimination power of these MCQs, a reduction of NFDs will be needed. It also shows the advantages of the psychometric analysis of MCQs test examinations which facilitates the faculty feedback to improve their MCQs and
distractors in order to develop a validated question bank. The ANOVA test was useful to show that even though the teaching methods are different on the same topics, the topic learned has the same standard across different groups of students.

**Take Home Messages**

- Quality assessment of a multiple choice test (MCQs) delivers validated questions and allows to form a trustworthy question bank.
- Multiple choice questions used for assessment have a better design when they are peer-reviewed.
- The measurement of psychometric properties of multiple choice questions (MCQs) gives a guide in where their flaws are, in order to improve them.

**Notes On Contributors**

Dr Ana Livia Licona-Chávez coordinates the medical research area since 2008, has been teaching for 12 years.

Dr Pierangeli Kay-to-py Montiel Boehringer has been a general practitioner since 1993, has been teaching for more than 30 years and has done medical education research since 2013.

Prof Lupita Rachel Velázquez-Liaño has been teaching since 2014, now is responsible for all the laboratories in this medical school.

**Acknowledgements**

The authors would like to thank Simitrio Rojas MD, Faculty of Medicine headmaster and the Xochicalco University in Ensenada, the group of Research Methodology teachers for their excellent work and the University authorities for their financial support. They also extend their gratitude to Walter Daessle PhD for his advice on this research manuscript.

The source of Figures 1 and 2 are the authors.

**Bibliography/References**

Ali, S. H., Carr, P. A. and Ruit, K. G. (2016) ‘Validity and reliability of scores obtained on multiple-choice questions: why functioning distractors matter’, *Journal of the Scholarship of Teaching and Learning*, 16(1), pp. 1-14. [https://doi.org/10.14434/josotl.v16i1.19106](https://doi.org/10.14434/josotl.v16i1.19106)

Backhoff, E., Larrazolo, N. and Rosas, M. (2000) ‘Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA)’, *Revista Electrónica de Investigación Educativa*, 2(1), pp. 11-29. [http://redie.uabc.mx/vol2no1/contenido-backhoff.html](http://redie.uabc.mx/vol2no1/contenido-backhoff.html) (Accessed: 15 Jul 2019).

Boopathiraj, C. and Chellamani, K. (2013) ‘Analysis of test items on difficulty level and discrimination index in the test for research in education’, *International Journal of Social Science & Interdisciplinary Research*, 2(2), pp. 189-193. [http://indianresearchjournals.com/pdf/IJSSIR/2013/February/15.pdf](http://indianresearchjournals.com/pdf/IJSSIR/2013/February/15.pdf) (Accessed: 02 Aug 2019).

Borrego-Mora, P. and Santana-Borrego, M. (2015) ‘Modificación de los índices psicométricos de los exámenes departamentales diseñados para la evaluación de los residentes de Medicina Interna del posgrado de la Facultad de medicina, UNAM, en función de cinco o cuatro opciones de respuesta’, *Medicina Interna de México*, 31(3), pp.
259-273. https://www.medigraphic.com/cgi-bin/new/resumen.cgi?IDARTICULO=58588 (Accessed: 02 Aug 2019).

Carneson, J., Delpierre, G. and Masters K. (2016) ‘Designing and Managing Multiple Choice Questions’. 2nd edn. pp. 1-29. https://doi.org/10.13140/RG.2.2.22028.31369

Delgado-Maldonado, L. and Sánchez-Mendiola, M.(2012) ‘Análisis del examen professional de la Facultad de Medicina de la UNAM: Una experiencia de evaluación objetiva del aprendizaje con la teoría de respuesta al ítem’, Investigación en Educación Médica, 1(3), pp. 130-139. http://riem.facmed.unam.mx/node/225 (Accessed: 17 Jul 2019).

Durante, M., Lozano, J., Martínez, A., Morales. S., et al. (2011) Evaluación de Competencias en Ciencias de la Salud. México: Panamericana.

Haladyna, T. M., Downing, S. M. and Rodríguez, M. C. (2002) ‘A review of multiple-choice item- writing guidelines for classroom assessment’, Applied Measurement in Education, 15(3), pp. 309-334. https://doi.org/10.1207/S15324818AME1503_5

Hancock, G. R. (1994) ‘Cognitive complexity and the comparability of multiple-choice and constructed-response test formats’, The Journal of Experimental Education, 62(2), pp. 143-157. https://doi.org/10.1080/00220973.1994.9943836

Hubbard, J. P. (1978) Measuring medical education: The tests and the experience of the National Board of Medical Examiners. 2nd edn. Philadelphia: Lea & Febiger. https://trove.nla.gov.au/work/11499822?q&versionId=13503574 (Accessed: 12 Sep 2019).

Kaur, M., Singla, S. and Mahajan, R. (2016) ‘Item analysis of in use multiple choice questions in pharmacology’, International Journal of Applied Basic Medical Research, 6 (3), pp. 170-173. https://doi.org/10.4103/2229-516x.186965

Kheyami, D., Jaradat, A., Al-Shibani, T. and Ali, F. A. (2018) ‘Items analysis of multiple choice questions at the department of paediatrics, Arabian Gulf University, Manama, Bahrain’, Sultan Qaboos University Medical Journal, 18(1), pp. 68-74. https://doi.org/10.18295/squmj.2018.18.01.011

Martínez, A., Trejo, J., Fortoul, T., Flores, F., et al. (2014) ‘Evaluación diagnóstica de conocimientos y competencias en estudiantes de medicina al término del segundo año de la carrera: el reto de construir el avión mientras vuelca’, Gaceta Médica de México, 150(1), pp. 35-48. https://www.medigraphic.com/cgi-bin/new/resumen.cgi?IDARTICULO=47949 (Accessed: 22 Aug 2019).

Martínez, M. E. (1999) ‘Cognition and the question of test item format’, Educational Psychology, 34(4), pp. 207-218. https://doi.org/10.1080/01443410.1999.11135017-0499-2

Mitra, N. K., Nagaraja, H. S., Ponnudurai, G. and Judson, J. P. (2009) ‘The levels of difficulty and discrimination indices in type a multiple choice questions of pre-clinical semester 1 multidisciplinary summative test’, International e-Journal of Science, Medicine & Education, 3(1), pp. 2-7. https://www.researchgate.net/publication/227858882 (Accessed: 28 Jun 2019).

Odukoya, J. A., Adekeye, O., Igbinoba, A. O. and Afolabi, A. (2017) ‘Item analysis of university-wide multiple choice objective examinations: the experience of a Nigerian private university’, Quality & Quantity, 52(3), pp. 983-997. https://doi.org/10.1007/s11135-017-0499-2

Ortiz-Romero, G., Díaz-Rojas, P., Llanos-Domínguez, O., Pérez-Pérez, S., et al. (2015) ‘Difficulty and discrimination of the items of the exams of Research Methodology and Statistics’, EDUMECENTRO, 7(2), pp. 2077-2874. http://www.revedumecentro.sld.cu/index.php/edumc/article/view/474 (Accessed: 03 Jun 2019).
Pell, G., Fuller, R., Homer, M. and Roberts, T. (2010) ‘How to measure the quality of the OSCE: A review of metrics- AMEE guide no. 49’, *Medical Teacher*, 32(10), pp. 802-811.
https://doi.org/10.3109/0142159X.2010.507716

Rao, C., Kishan-Prasad, H. L., Sajitha, K., Permi, H., *et al.* (2016) ‘Item analysis of multiple choice questions: Assessing an assessment tool in medical students’, *International Journal of Educational and Psychological Researches*, 2(4), pp. 201-204. https://doi.org/10.4103/2395-2296.189670

Saldaña, Y., Delgadillo, H. and Méndez, I. (2014) ‘Evaluación de un examen parcial de Bioquímica’, *Revista Educación Bioquímica*, 33(4), pp. 104-110. http://ref.scielo.org/p9zkhb (Accessed: 17 Oct 2019).

Sahoo, D. P. and Singh, R. (2017) ‘Item and distracter analysis of multiple choice questions (MCQs) from a preliminary examination of undergraduate medical students’, *International Journal of Research in Medical Sciences*, 5(12), pp. 5351-5355. http://dx.doi.org/10.18203/2320-6012.ijrms20175453

Sajitha, K., Permi, H. S., Rao, C. and Prasad, K. (2015) ‘Role of item analysis in post validation of multiple choice questions in formative assessment of medical students’, *Nitte University Journal of Health Science*, 5 (4), pp. 58-61. http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=113856963&site=ehost-live (Accessed: 05 Jun 2019).

Simkin, M. G. and Kuechler, W. L. (2005) ‘Multiple-Choice test and student understanding: what is the connection?’, *Decision Sciences Journal of Innovative Education*, 3(1), pp. 74-97.
https://doi.org/10.1111/j.1540-4609.2005.00053.x

Tavakol, M. and Dennick, R. (2012) ‘Post-examination interpretation of objective test data: monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66’, *Medical Teacher*, 34(3), pp 161-175.
https://doi.org/10.3109/0142159x.2012.651178

Ware, J. and Vik, T. (2009) ‘Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations’, *Medical Teacher*, 31(3), pp. 238-243.
https://doi.org/10.1080/01421590802155597

**Appendices**

None.

**Declarations**

The author has declared that there are no conflicts of interest.

*This has been published under Creative Commons "CC BY 4.0"* (https://creativecommons.org/licenses/by-sa/4.0/)

**Ethics Statement**

Ethics approval was granted by Universidad Xochicalco, campus Ensenada, Ethical Committee. This Committee gave us their approval through a statement with the approval reference number: MD/767/20-1, signed by a member and the chairwoman. We have this approval and its translation in english.
External Funding

This article has not had any External Funding

MedEdPublish: rapid, post-publication, peer-reviewed articles on healthcare professions’ education. For more information please visit www.mededpublish.org or contact mededpublish@dundee.ac.uk.