Evaluation of the Effect of Items' Format and Type on Psychometric Properties of Sixth Year Pharmacy Students Clinical Clerkship Assessment Tools

Suha Al Muhaissen (✉ s.muhaissen@ju.edu.jo)

University of Jordan <https://orcid.org/0000-0002-5177-1222>

Hatim S. AlKhatib

The University of Jordan

Gayle Brazeau

University of New England

Amal Akour

The University of Jordan

Research article

Keywords: Assessment Items, Clinical Clerkships, Difficulty Index, Discriminating Index, Point Biserial

Posted Date: November 27th, 2019

DOI: <https://doi.org/10.21203/rs.2.17768/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background Examination is the traditional assessment tool, and are used to guide the improvement of academic programs. Accreditation committees' emphasis on the implementation of standardized examinations. The aims of the current study are to evaluate the quality of assessment tools of sixth year PharmD students as a function of assessment item format and type/structure and to assess the effect of the number of response choices on the characteristics of MCQs as assessment items.

Methods. A total of 173 assessment items used in the examinations of sixth year rotations of PharmD program were included. Items were classified as case based or non-case based and as MCQs or open-ended. The psychometric characteristics of examinations were studied as a function of the level of the Bloom's levels addressed by an item, item format, and number of choices in MCQs.

Results. Items addressing intellectual and analysis skills were more difficult, while items associated with multiple cognitive levels were more discriminative. No differences were found between case based and noncase based items in terms of their difficulty, with a slightly better discrimination in the letter. Open-ended items were easier, yet more discriminative. MCQs with higher number of options were easier and more discriminative. Open-ended questions were significantly easier and more discriminative in comparison to MCQs as case based items while they were more difficult and more discriminative as non-case based items.

Conclusion. Item formats, structure, and number of options in MCQs, affected students' performance and overall examination quality. The discrimination of items associated with multiple Bloom's levels was significantly higher than those associated with a single level. Noncase based items and open-ended items were easier and more discriminative than case based items and MCQs, respectively.

Introduction

A good quality assessment tool is essential for generating reliable data to guide program improvements. Examinations are the traditional evaluation method of students' performance used by instructors throughout educational history [1]. As a result, this leads to a significant responsibility on whom one of their essential tasks is the creation and selection of appropriate examinations' items [2].

The Accreditation Council for Pharmacy Education (ACPE) standards for the Doctor of Pharmacy (PharmD) programs recommend the implementation of an extensive assessment plan to prepare graduates for practice [3]. A plan should include the use of standardized, systematic, reliable, and valid assessment. It also requires both knowledge and performance evaluation and measurement of the achieved professional competencies. In order to achieve such assessment quality, it is essential to analyze and evaluate the assessment tools after application and use the results to improve them. Such analysis and evaluation are needed to specify the assessment characteristics; whether it is performance-oriented, the thinking order it evaluates, and its real-life context [4].

Assessment can be performed using different items formats and types; however, assessment tools should be selected so as to suit the expected position of a PharmD graduate in healthcare team [5]. The assessment items could be classified according to their format as either case-based or non-case based. An item belonging to either format category could be further classified depending on its type/structure as Multiple Choice Questions (MCQ) or an open ended/constructed response question [6, 9].

Case-based evaluation items format has the distinct advantage over non-case based format as it can simulate realistic decision making scenarios and allow student to attempt to solve problems and find alternative plans using individualized detailed information [1, 2, 7]. MCQs are a popular assessment item type where the examinee has to choose the correct answer to the "stem" question from multiple possible answers. Properly constructed MCQs allow the examiner to serve and cover many learning objectives and thinking levels [2, 6, 10]. On the other hand, an evaluation that depends on answering open ended/constructed response items, allows the exploration of various alternatives rather than concentrating on one correct answer, and encasement of higher thinking orders [2].

National Pharmacy Licensure Examination (NAPLEX) is a health profession examination that measures a candidate's knowledge of the practice of pharmacy [1]. NAPLEX utilizes assessment items of different formats and types, including constructed response/open ended, and MCQs (A-type, K-type, true-false items, etc.) [8] with the case-based type as the most prevalent item format [2].

While the quality of assessment items of different formats and types (case-based learning, MCQs, and open-ended items) has been addressed by many authors [2, 7, 9, 11-15], there are only two studies [2, 16] that attempted to compare the quality of case based and non-case based assessment items.

This study evaluates the quality of sixth year examinations written by a panel of expert and based on the Bloom's levels. The primary objective of the study was to evaluate the quality of the assessment items as a function of item format (case based versus non-case based) and assessment item type/structure (MCQs and open-ended/essay items). Another objective was to assess the effect of the number of response choices on the characteristics of MCQs items.

Methods

Data Collection

Assessment items used in the examinations of six rotations (Cardiology, Critical Care, Respiratory, Endocrinology, Oncology, and Nephrology rotations of the sixth/senior year of the PharmD program offered by the School of Pharmacy at the University of Jordan, (SP-UJ) were mapped to the Bloom's levels. All examinations were final examinations that were 60 to 75 minutes in length. The examinations were offered in the first and second semesters of the academic year 2015-2016. Each examination was constructed and reviewed by an expert panel composed of an academic adviser/rotation coordinator, an academic staff member, and at least two preceptors. Each item covered at least one of the four Bloom's

educational learning objectives levels (Knowledge and understanding, Intellectual analytical and cognitive skills, Transferable-key skills, and Subject-specific skills) [9].

A total of 173 assessment items were included in this study. Neither individual student scores nor item author information were collected or used in the analyses. The student names and University ID numbers were covered to maintain confidentiality. No item was excluded.

Each item was reviewed and categorized as either case-based or non-case based by the authors. Case-based items were those that were asked in a scenario-based format (i.e. patient profiles with accompanying test questions) so that in order to properly analyze and answer, a student must refer to the information provided in the patient profile [2]. While non-case based “stand-alone questions,” had answers that could be drawn solely from the information provided in the question stem [2]. Items were further classified according to their type as MCQs or open-ended (essay) item. MCQs were further classified according to the number of answer options.

The examinations were characterized in terms of their reliability using Cronbach's Alpha [17] and individual items were characterized in terms of their level of difficulty, discrimination index, point biserial, and number of options (for MCQs) [6, 7, 10, 11, 13, 15, 18, 20]. Student's grades for each item or sub-item were transferred to SPSS (IBM, Chicago, IL) and psychometric parameters were estimated. Item performance/psychometric characteristics that were calculated included Difficulty Index (DIF I), Discriminating Index (DI), and point bi-serial (PBS). The values of these psychometric parameters were used to classify the item quality (Table 1) [10, 12, 13, 19, 20]. DIF I was then calculated as the percentage of the correct responses or the average grade of the specified item relative to the total mark assigned for the item. The desired value for this psychometric parameter ranges from 20%-30% at the lower limit, to 75%-80% at the higher limit. DI represents the difference between the average grade of the students in the upper quartile (students with highest totals) relative to the item total grade and the average item grade of the students in the lower quartile (students with lowest totals) relative to the item total grade divided by the number of students in a quartile. PBS is also a measure of the item discriminative power, this correlation is a comparison of performance on an item relative to whole test performance [10, 18, 19, 21, 22], PBS was estimated using SPSS reliability output [10, 19, 22]. DI and PBS values can range from -1 to 1. High values of DI and PBS indicate that an item was correctly answered by high-performing students, and/or incorrectly answered by low-performing students. On the other side, low or negative indices reveal that an item was incorrectly answered by high-performing students, and/or correctly answered by low-performing students; suggesting a poor or flawed item, or low ability to differentiate between students.

Statistical Analysis

The differences in means of item performance characteristics (DIF I, DI, and PBS) as a function of its format and type were assessed using a one-way analysis of variance (MANOVA), and the effect of the 4 Bloom's levels on the correct responses fraction and discrimination indices being tested, and then the post-hoc Bonferroni correction test, analyses of variance (ANOVA) on the dependent variables were conducted as follow-up to the MANOVA. Pearson's correlation coefficient and t-test of independent

samples were used to evaluate correlations and in between group differences, with $p\text{-value} < .05$ indicating statistical significance. All data analysis was performed using SPSS® 23.0 (IBM, Chicago, IL).

Results

A total of a hundred and seventy-three items, each answered by 72-83 students were evaluated. These items were compiled from 6 different final examinations of clinical clerkships during the senior year of PharmD program offered by the SP-UJ. The reliability of each of the studied examinations as measured using by Cronbach's Alpha ranged between .62-.80. The reliability of the studied items differed according to item type; MCQs had an average Cronbach's Alpha of .61, while for open ended items the average Cronbach's Alpha was .27. A significant positive Pearson's correlation was observed between the reliability of each examination measured as the difference between Cronbach's Alpha value of the examination and Cronbach's Alpha value if a specific item was deleted and item psychometric parameters; DIFI ($r=.16, p<.05$), DI ($r=.63, p<.001$), and PBS ($r=.85, p<.001$).

Table 1 shows the characteristics of the studied assessment items. Over three-quarters of the items studied (77.5%) were case-based. More than half of the items were of open-ended structure that measured students' knowledge and understanding skills measuring two levels. The psychometric parameters of the sample items analysis showed that 54% of the questions had excellent difficulty index (DIF I range 20% - 50%) [9, 10, 12], almost one third had excellent DI, and around 60% of the questions were on the higher end of point bi-serial range, while not more than 8.1% items were of questionable quality.

Table 2 shows the mean and standard deviation values of DIF I, DI, and PBS for the questions addressing the four levels of Bloom's level (Knowledge and understanding skills, Intellectual analytical and cognitive skills, transferable key skills, and subject-specific skills) as well as MANOVA and follow-up tests results.

Upon applying one-way MANOVA analysis, significant differences were found in item characteristics (DIF I, DI, and PBS) as a function of the measured Bloom's level.

As follow up tests to MANOVA, we performed analysis of variance on the dependent variables (DIF I, DI and PBS). Using Bonferroni method, the univariate ANOVA on the DIF I was significant. Similar significant differences were found from ANOVA performed on DI, and on PBS.

The post-hoc analysis to ANOVA for DIF I, DI, and PBS included performing pairwise comparisons. DIF I of Knowledge level items was significantly higher ($p=.03$) while DIF I of Intellectual analytical level was significantly lower than those for the other levels. However, DI and PBS of Knowledge skills and Intellectual analytical skills was significantly higher than the same metric for transferable key and subject-specific levels. This suggests that the discriminative power of test items associated with multiple Bloom's levels was significantly higher than test items associated with one level. On the other hand, the level of difficulty was not affected by of items association with multiple Bloom's levels.

Table 3 represents the analysis of item performance characteristics for different item properties. Case based items were not different in their DIF I ($p=.52$) and PBS ($p=.47$), but they showed significantly lesser DI ($p=.037$) in comparison to non-case based items. Open-ended type items demonstrated significantly higher DIF I ($p=.008$), DI ($p=.000$), and PBS ($p=.000$) relative to MCQs. On the other hand, 4-option MCQs showed significantly lower DIF I ($p=.000$) and PBS ($p=.03$), but they were not different with respect to DI ($p=.81$); suggesting similar discrimination power.

When items were compared based on item type/structure; case based item that are open ended type showed significantly higher DIF I ($p=.001$), DI ($p=.000$), and PBS ($p=.000$) when compared to case based item that are of the MCQs type. While noncase based items which are open ended type had significantly lower DIF I ($p=.02$), and higher DI ($p=.001$) and PBS ($p=.001$). The number of choices that a case based item possessed a significant impact on the DIF I ($p=.001$), and no effect on DI (.972) and PBS (.163). However, the number of choices exhibited no effect on noncase based questions. Open ended items when formatted as case based had significantly higher DIF I ($p=.001$), but lower DI ($p=.001$). While MCQs comparison based on item format exhibited higher discrimination measured by DI ($p=.001$) and PBS ($p=.001$).

MCQ items with four answer options showed significant differences when categorized as case based and noncase based items, this showed as higher DIF I ($p=.00$) and lower DI and PBS ($p=.00$). This was not the case dealing with 5 options items as these items demonstrated no effect of it format on DIF I and DI, but significant slightly higher PBS ($p=.00$) of noncase based options.

Discussion

The present study addressed the quality of sixth year PharmD final examinations. It provided three interesting and valuable outcomes that can be of benefit to academic staff and preceptors. (1) The reliability of an examination correlated significantly with items psychometric parameters, (2) multiple Bloom's levels associated with an item affected the discriminative power of the item, and the (3) structure of an item and the number of options possessed by an item significantly affected the psychometric parameters of the specified item.

The predominant item format in the current study was case based, in which the basic level competencies; knowledge and understanding skills, constituted the majority; around two-thirds, of the measured skills. These competencies are the foundation for the higher competencies levels (e.g. analysis and patients-specific skills). The use of case based items in the assessment of students in a health care professional program, such as a Pharm D program, is necessary. It acts to introduce students to clinical scenarios that simulate patient situation, and enables them to practice decision making during realistic challenges.

Building case based items is a time consuming task, it requires a knowledgeable and practice expert examiner [2, 4]. The psychometric parameters of the studied assessment items in our study showed their high quality with less than 8% classified as poor or flawed items [13, 20]. The benefits implied by the use

of case based items and items psychometrics parameters, in addition to the high values of Cronbach's Alpha of each examination evidenced the high quality and reliability of exams under study [17].

Evaluation of the effect of competency levels on the difficulty of an item shows that items addressing intellectual and analysis skills are more difficult, on the other end of the scale are knowledge and understanding skills which were much easier. This is understandable since the students who don't have the basic knowledge cannot derive correct responses. These findings are in agreement with the findings of Kim and colleagues (2012), where they found that analysis and synthesis items are more difficult [21]. However, the mentioned study showed that the association of items with multiple concepts affect item difficulty. In our study, items with a single concept showed significant effect on DIF I. this probably because we are assessing both MCQs and open-end items while in their study MCQs were only included.

The evaluation on discrimination measures (DI and PBS) of assessment items addressing different competency levels indicates that items associated with multiple levels (knowledge and understanding items and Intellectual and analysis items) can differentiate significantly between students in upper and lower grade' quartiles. These results are comparable to Kim and colleagues (2012) findings that items associated with multiple functions; measure more than one thinking order, are more discriminative [21].

Analysis on DIF I, DI, and PBS of item formats demonstrated no differences between case based and noncase based items in terms of DIF I and PBS, and with a slightly better DI of noncase based items, DI difference was relatively small. These two DI values are still within the same "good" category of DI. These results are similar to that of Phipps and Brackbills (2009) findings [2], demonstrating comparable capability of these two item formats.

The type of an item has significant effect on its psychometric characteristics. Open ended type was easier, yet more discriminative, this is similar to Thawabieh (2016) findings [16]. It is understood that the nature of open ended items allow for the incorporation of many details when answered by students, while utilizing multiple and higher thinking orders allowing for better discrimination between high- and low-performance students. On the other hand, the options in MCQs may provide orientation for students to detect item-writer intention [21].

The number of options an item possessed showed significant impact on both difficulty and discrimination levels measured as DIF I and PBS, respectively. The higher the number of options the easier the item and more discriminative. This is in partial agreement with Phipps and Brackbills (2009) findings where they found that 5-options are more difficult and more discriminative. Despite that, they concluded that due to the very small differences between these two groups, it is explainable/justifiable to use a mix of 4 and 5 responses MCQs in exams. Also, in their study they analyzed both A-type and K-type MCQs [2].

Analyzing case based items and noncase based items separately revealed different behaviors. Case based items that are open ended vs. MCQs, are significantly easier and more discriminative, while the same type of noncase based items is more difficult and more discriminative. This can be attributed to the

fact that case based items provide scenarios that may simplify the item and guide the examinee but still need to be seen in context.

The number of answer options (4 or 5), had no effect in discrimination metrics of either case based or noncase based assessment items, and it only affected the difficulty of case based items, as 4-option questions were more difficult. The idea of writing more plausible and effective options other than the key answer when an item is based on a case that's full of details is clearly more challenging and difficult.

Open ended items that are noncase based are more difficult and more discriminative in comparison with open ended that are case based. In addition, MCQs that are noncase based have larger DI and PBS; showing that noncase-based items are more discriminative. Again, case based items were shown to have similar, if not inferior, behavior to noncase items, limiting their benefit to their ability to address intended learning and course aims, but expressing no unique performance assessment characteristics.

One more result of the current study was the different effect of the item format on the characteristics of 4- and 5-option MCQs. Noncase based, 4-option MCQs items were significantly easier and more discriminative than case based 4-option MCQs. However, case based and noncase based 5-option MCQs items had no differences in DIF I and DI, and differed slightly in PBS while remaining within the same recommended PBS value range.

The previous results showed differences between the two MCQs groups yet cannot be conclusive, as it once again a very challenging time-consuming task not only to construct a case item but also to construct strong, reliable, and efficient choices during the creation of MCQs regardless the item format; being based on case or not.

In a study conducted by Sheaffer and Addo (2013), where they measure both second year Pharm D students' performance and confidence in answering selected-response and constructed-response items, it was concluded that students performed better and felt more confident in answering selected-response items. Moreover, the incorporation of constructed-response teaching and testing method in pharmacy learning and education was recommended [8].

One important limitation of the present study is the unequal number of items per groups which may have affected the analysis. On the other hand items classification based on the Bloom's levels might be subjective [5]; for this reason peer reviewers represented by both, preceptors as the personnel in direct connection with "real life" cases, and Academic staff/educators to minimize the controversy.

The findings of our study, uncovered an important issue; specifically, our students are prepared and trained to deal with real cases situations, and whether the instructors and the used teaching and learning methods do initiate and develop their student's skills and abilities to do so. One other question is important to answer, can and do instructors assess the fact that the used teaching methodology/methodologies are tailored with the type and structure of exams. Is it would seem that

instructors/examiners receive quality training in how to construct an examination/item [6]. Many pedagogical aspects and evaluation perspectives should be included in the assessment tools evaluation.

Conclusion

Reliable and effective assessment of students in health care professional programs where decisions related to patients' treatment are to be made is crucial. PharmD students should be helped to deal with real medical cases during their study course especially senior year. Reliability measures and psychometrics parameters are efficient in evaluating clerkships final exams. The necessity to build exams on the expected courses objective is demanding, the study showed that the discriminative power of items associated with multiple Bloom's levels are significantly higher than those associated with single level. Item formats, structure, and number of options in MCQs, as well as the different combinations of these factors will affect students' performance and over all examination quality. However, it should be kept in mind the important value of case based testing and teaching methods.

Abbreviations

ACPE; Accreditation Council for Pharmacy Education, PharmD; Doctor of Pharmacy, MCQ; Multiple Choice Questions, NAPLEX; National Pharmacy Licensure Examination, SP-UJ; Pharmacy at the University of Jordan, DIF I; Difficulty Index, DI; Discriminating Index, PBS, point bi-serial, MANOVA; one-way analysis of variance, ANOVA; analyses of variance.

Declarations

Ethics approval and consent to participate

This study was approved by by the SP-UJ Scientific Research Committee (IRB: 7/2017).

Consent for publications

Not applicable

Availability of data and materials

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Competing of interest

The authors declare that they have no competing interests.

Funding

Not applicable

Authors' contributions

HA participated in designing the study and participated in the writing of the manuscript. GB revision of data analysis and participated in the writing of the manuscript. AK revised and modified the tested examination and the written manuscript. SA participated in designing the study, revised the data collection entry, carried out the data analysis, and participated in the writing of the manuscript. All authors have read and approved the manuscript. In addition, all authors are aware of this submission and agree with it.

Acknowledgement

The authors would like to acknowledge the faculty members and preceptors at the Department of Biopharmaceutics and Clinical Pharmacy at the SP-UJ for their cooperation. Also, the authors would like to acknowledge Miss. Sara Jalouqa for her help in data entry.

References

1. Tofade T, Elsner J, Haines ST. Best Practice Strategies for Effective Use of Questions as a Teaching Tool, *Am J Pharma Educ.* 2013; 77(7): 1-9 Article 155.
2. Phipps SD, Brackbill ML. Relationship between Assessment Item Format and Item Performance Characteristics, *Am J Pharma Educ.* 2009; 73(8) : 1-6 Article 146.
3. Accreditation Council for Pharmacy Education, 'Accreditation Standards and Key Elements for the Professional Program in Pharmacy Leading to the Doctor of Pharmacy Degree "Standards 2016", 2015.
4. Garavalia LS, Marken PA, Sommi RW. Selecting Appropriate Assessment Methods: Asking the Right Questions. *Am J Pharma educ.* 2002; 66: 108-12.
5. Wallman A, Lindblad AW, Hall S, Lundmark A, Ring L. A Categorization Scheme for Assessing Pharmacy Students' Levels of Reflections during Internship. *Am J Pharma Educ.* 2008; 72(1): 1-10 Article 05.
6. Caldwell DJ, Pate AN. Effects of Question Formats on Student and Item Performance. *Am J Pharma Educ.* 2013; 77(4): 1-5 Article 71.
7. Medina MS. Relationship between Case Question Prompt Format and the Quality of Responses. *Am J Pharma Educ.* 2010; 74(2): 1-7 Article 29.
8. Sheaffer EA, Addo RA. Pharmacy Student Performance on Constructed-Response Versus Selected-Response Calculations Questions. *Am J Pharma Educ.* 2013; 77(1), 1-7 Article 6.
9. Palmer EJ, Devitt PG. Assessment of Higher Order Cognitive Skills in Undergraduate Education: Modified Essay or Multiple Choice Questions? Research Paper. *BMC Med Educ.* 2007; 7(49):1-7.

10. Al Muhaissen SA, Ratka A, Akour A, Alkhatib HS. Currents in Pharmacy Teaching and Learning, <https://doi.org/10.1016/j.cptl.2018.12.006>
11. Chauhan PR, Ratrhod SP, Chauhan BR, Chauhan GR, Adhvaryu A, Chauhan AP. Study of Difficulty Level and Discriminating Index of Stem Type Multiple Choice Questions of Anatomy in Rajkot. *B I O M I R R O R*. 2013; 4(6): 37-40.
12. Sabri S. Item Analysis of Student Comprehensive Test for Research in Teaching Beginner String Ensemble Using Model Based Teaching among Music Students in Public Universities. *Int J Educ Res*. 2013; 1(12):1-14.
13. Siri A, Freddano M. The Use of Item Analysis for the Improvement of Objective Examinations. *Procedia - Social and Behavioral Sciences*. 2011; 29:188-97.
14. Tarrant M, Ware J, Mohammed AM. An Assessment of Functioning and Non-Functioning Distractors in Multiple-Choice Questions: A Descriptive Analysis. *BMC Med Educ*. 2009; 9(40):1-8.
15. Trevisan MS, Sax G, Michael WB. The Effects of the Number of Options Per Item and Student Ability on Test Validity and Reliability. *Educational and Psychological Measurement*. 1991; 51(4):829-37.
16. Thawabieh AM. A Comparison between Two Test Item Formats: Multiple-Choice Items and Completion Items. *British J Educ*. 2016; 4(8): 63-74.
17. Norcini JJ, Swanson DB, Grosso LJ, Webster GD. Reliability, Validity and Efficiency of Multiple Choice Question and Patient Management Problem Item Formats in Assessment of Clinical Competence. *Med* 1985; 19: 238-47.
18. Brown JD. Point-Biserial Correlation Coefficients. *JLT Testing & Evaluation SIG* 2001; 5(3): 13-17.
19. Varma S. Preliminary Item Statistics Using Point-Biserial Correlation and P-Value. 2006. Educational Data Systems. Inc. eddata.com Accessed 13 September 2018.
20. Sim S, Rasiah RI. Relationship between Item Difficulty and Discrimination Indices in True/False-Type Multiple Choice Questions of a Para-Clinical Multidisciplinary Paper. *Annals Academy of Medicine* 2006; 35: 67-71.
21. Kim MK, Patel RA, Uchizono JA, Beck L. Incorporation of Bloom's Taxonomy into Multiple-Choice Examination Questions for a Pharmacotherapeutics Course. *Am J Pharma Educ*. 2012; 76(6): 1-8 Article 114.
22. IBM corporation SPSS, 'Using SPSS for Item Analysis'. SPSS Inc. 1998.

Tables

Table 1. Descriptive Statistics for Evaluated Assessment Items (N=173)

Variable	No. (%)
Format	
Case based item	134 (77.5)
Non-case based item	39 (22.5)
Type/Structure	
Open-ended/essay item	98 (56.6)
Multiple choice item	75 (43.4)
Number of Choices*	
2 (True/False)	11 (14.7)
3	3 (4)
4	33 (44)
5	28 (37.3)
Bloom's level ^{\$}	
Knowledge and Understanding Skills	111 (64.2)
Intellectual Analytical and cognitive skills	57 (32.9)
Transferable Key Skills	28 (16.2)
Subject-Specific Skills	28 (16.2)
Difficulty Index (DIF I) Levels ¹	
Difficult (DIF I<20%)	6 (3.5)
Acceptable/Good (20%≤DIF I<50%)	29 (16.7)
Excellent (50≤DIF I<80%)	93 (53.8)
Easy/Poor (DIF I≥80%)	45 (26)
Discriminating Index (DI) Levels ¹	
Poor/Flawed (DI<0)	5 (2.9)
Poor (0≤DI<.2)	36 (20.8)
Acceptable (.2≤DI<.3)	40 (23.1)
Good (.3≤DI<.4)	39 (22.5)
Excellent (DI≥.4)	53 (30.6)
Point-Biserial (PBS) Levels ²	
Poor/Flawed (PBS<0)	14 (8.1)
Poor (0≤PBS<.15)	28 (16.2)
Recommended (.15≤PBS<.25)	28 (16.2)
Good (PBS≥.25)	103 (59.5)

* Percent relative to MCQs count.

^{\$} Don not sum up 100%, as an item can cover more than one level

Table 2. Item Performance Characteristics Based on Item measured ILO Level

Performance Characteristics	Knowledge and Understanding Skills Mean (SD)	Intellectual Analytical and Cognitive Skills Mean (SD)	Transferable Key Skills Mean (SD)	Subject-Specific Skills Mean (SD)
Level of Difficulty	66.8 (18.9) ^a	56.4 (22.1) ^b	66.7 (17.1)	63.0 (22.3)
Discriminating Index	.41 (.19) ^d	.40 (.19) ^d	.22 (.09)	.25 (.11)
Point-biserial	.39 (.12) ^c	.38 (.18) ^c	.21 (.10)	.23 (.12)

ILO: Intended Learning Outcome

^a higher significant difference between single concept comparison.^b lower significant difference between single concept comparison.^c higher significant difference between pairwise comparison over other skills.^d higher significant difference between pairwise comparison over other skills.

Table 3. Item Performance Characteristics Based on Different Item Properties and Layers (N=173).

	Level of Difficulty Mean (SD)	Discriminating Index Mean (SD)	Point-biserial Mean (SD)
Item Format			
Case Based	64.4 (20.7)	.33 (.21)	0.35 (.20)
Non-case Based	62.9 (20.1) ^{NS}	.45 (.21)*	0.23 (.12) ^{NS}
Item Type/Structure			
Open-ended/Essay	66.5 (18.3)	.39 (.23)	.37 (.13)
Multiple Choice	60.8 (22.8)*	.26 (.22)*	.24 (.11)*
Number of Choices/options			
4-option Items	50 (20.1)	.39 (.21)	.17 (.1)
5-option Items	65.3 (23.3)	.35 (.24) ^{NS}	.29 (.1)*
Item Format: Item Type/Structure			
Case Based: Open-ended	67.5 (18.5)	.38 (.11)	.37 (.13)
Case Based: MCQs	58.2 (23.3)*	.25 (.21)*	.14 (.12)*
Non-case Based: Item Type/Structure			
Non-case Based: Open-ended	56.2 (11.4)	.52 (.23)	.41 (.11)
Non-case Based: MCQs	64.8 (21.7)*	.36 (.21)*	.23 (.11)*
Case Based: Item Type/Structure			
Case Based: 4-option	45.8 (20.9)	.31 (.20)	.19 (.11)
Case Based: 5-option	69.1 (20.8)*	.29 (.20) ^{NS}	.18 (.11) ^{NS}
Non-case Based: Number of Choices/options			
Non-case Based: 4-option	59 (14.8)	.35 (.20)	.24 (.11)
Non-case Based: 5-option	61.1 (25.6) ^{NS}	.39 (.20) ^{NS}	.29 (.12) ^{NS}
Open-ended: Item Format			
Open-ended: Case Based	67.5 (18.5)	.36 (.21)	.38 (.12)
Open-ended: Non-case Based	56.2 (11.4)*	.57 (.22)*	.42 (.14) ^{NS}
MCQ: Item Format			
MCQ: Case Based	58.2 (23.3)	.23 (.20)	.14 (.180)
MCQ: Non-case Based	64.8 (21.7) ^{NS}	.35 (.21)*	.26 (.11)*
4-option: Item Format			
4-option: Case Based	45.8 (20.9)	.27 (.19)	.15 (.09)
4-option: Non-case Based	59 (14.8)*	.38 (.20)*	.24 (.11)*
5-option: Item Format			
5-option: Case Based	69.1 (20.8)	.28 (.17)	.13 (.09)
5-option: Non-case Based	61.1 (25.5) ^{NS}	.33 (.19) ^{NS}	.27 (.11)*

NS: Not Significant

* Significant at measurement level