Advances in Medical Education and Practice

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Effect of blueprinting methods on test difficulty, discrimination, and reliability indices: cross-sectional study in an integrated learning program

This article was published in the following Dove Medical Press journal:
Advances in Medical Education and Practice

Hussein Abdellatif[1,2]
Abdullah M Al-Shahrani[3]

[1]Anatomy and Embryology Department, Faculty of Medicine, University of Mansoura, Mansoura, Egypt; [2]Department of Anatomy, College of Medicine, University of Bisha, Bisha, Saudi Arabia; [3]Department of Family Medicine, College of Medicine, University of Bisha, Bisha, Saudi Arabia

Purpose: Exam blueprinting achieves valid assessment of students by defining exactly what is intended to be measured in which learning domain and defines what level of competence is required. We aimed to detect the impact of newly applied method for blueprinting that depends on total course credit hours and relate the results with item analysis reports for students' performance.
Participants and methods: A new method for blueprint construction was created. This method utilizes course credit hours for blueprint creation. Survey analysis was conducted for two groups of students (n=80); one utilized our new method (credit hours based) for blueprinting and the other used traditional method depending on exam duration and time for individual test items.
Results: Results of both methods were related to item analysis of students' achievements. No significant difference was found between both groups in terms related to test difficulty, discrimination, or reliability indices. Both achieved close degrees of test validity and reliability measures.
Conclusion: We concluded that our method using credit hours system for blueprinting could be considered easy and feasible and may eventually be utilized for blueprint construction and implementation.

Keywords: learning, students, reproducibility, competency, curriculum, blueprint

Introduction

Assessment of skills or knowledge is of equal importance to teaching/learning of the skill or knowledge.[1] Assessment or knowledge evaluation is not a new concept and we have all at some point taken preinstructional assessment tests, interim mastery test, and mastery tests. It has been frequently asked by academics "what is the purpose of these tests?"

Herman (1992) pointed out that "People [students] perform better when they know the goal, see models, know how their performance compares to the standard".[2] The basic purpose of all tests is discrimination (to distinguish the level of aptitude, abilities, and skills among the test takers) regardless of the way how the test was constructed or conducted. For professional and academic interest, the objective would include such discriminators as proficiency, analytical and reasoning skills, technical aptitude, and behavioral traits, among many others.[3] There are two commonly used approaches to achieve such discrimination between test holders, the first is norm-referenced approach wherein the relative performance of test takers is considered and the second one is criterion-referenced approach wherein the performance of test takers is referenced against a predetermined benchmark.[4]

Correspondence: Hussein Abdellatif
Anatomy and Embryology Department, Faculty of Medicine, University of Mansoura, Mansoura 67611, Egypt
Tel +20 114 591 9955
Email Hussein.abdullatif@hotmail.com

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/AMEP.S190827

There have been many characters for psychometrically sound tests. Of these, validity, reliability, integrity, and achievement of variance among scores are the most considered when constructing a test for evaluation.[5]

Validity of the test is a measure of how the test evaluates what is supposed to measure, in other words, it evaluates and measures test usefulness. Validity also ensures that students' are satisfying minimum performance level showing intended level of competence set out at the intended learning outcomes.[6]

Exam blueprinting achieves valid assessment of students by defining exactly what is intended to be measured in which learning domain and defines what level of competence is required. One of the major tasks to get a valid test is to ensure the concept of content validity which means that each test item must at least represent one learning outcome.[7] Careful combination of highly representative items is the matter, which results in better test validity rather than constructing high-quality representative items alone.[8] The tool of choice to achieve the best combination and representativeness of issues in exam is test blueprint.[9] Careful blueprinting helps to reduce to major validity threats, first is the construct with under representation (biased sampling of course content) and the other is to construct with irrelevant variance (usage of inappropriate tools for assessment).[10] There are several methods for blueprint construction, of these the curriculum design and the learning approach are the major players.[11]

The initial step in blueprinting is to construct a table of specification (TOS), which shows what will be tested in relation to what has been taught. This ensures that the assessment has content validity and that the same emphasis on content during instruction is represented in the assessment. Besides, it ensures test item alignment with objectives. This helps to minimize the possible bias in test construction. The objectives are aligned with the learning domains, which are either cognitive domains (Bloom's cognitive skills) or clinical skills. This creates what is known as a two-dimensional matrix for blueprint design.[12] Cognitive skills can be divided into six levels according to Bloom's Taxonomy,[13] these are knowledge, understanding, application, analysis, synthesis, and evaluation. Application of all these domains in exam construction is relatively difficult, so many institutes use the simplified approach of Ward's (1983) taxonomy who divided cognitive domain into three levels which are recall, application, and problem solving.[7,14]

Assessment of test construction and individual exam items is done by item analysis, which is a process that examines student response to individual test items.[15] Exam item analysis is valuable in improving items that will be used later in exams, and it is used to eliminate misleading and ambiguous items in single test discrimination. In addition, it is used to increase instructors' skills in test construction and highlights course contents that require more emphasis and clarity.[16] Of these, item difficulty, discrimination, and test reliability indices are commonly used to assess the performance of individual test items on the basis that overall quality of test is derived from quality of its individual items.

Several medical schools have currently shifted to the credit hours policy in their learning approach; this eventually helps to overcome some learning difficulties such as overcrowding of the curriculum.[17] The validity of this learning system in achieving the required objectives has been discussed by many methods; however, the impact of implementing such learning strategy (credit hours) directly upon students' assessment results has not been yet detected.

In the current work, we aimed to detect the impact of a newly applied method for creating a test blueprint that depends on total course credit hours on item analysis results including difficulty, discrimination, and reliability indices. To ensure the validity of this newly applied method, we compared item analysis results with previously utilized one that creates a test blueprint using the overall assessment time and time allocated for examinee to answer each type of questions.

Participants and methods

Study design

The study was a cross-sectional observational study.

Study site

The study was conducted in College of Medicine, University of Bisha, Saudi Arabia.

Study participants

The whole students (males, n=80) enrolled in the course of basic structure and function during the academic year 2017–2018 were included in the study. The course of basic structure and function applied for second year medical students is part of our integrated problem-based learning (PBL) curriculum applied at College of Medicine, University of Bisha. The course of basic structure and function is part of phase I curriculum, which includes in addition seven other modules that are introduced one after the other. The course of basic structure and function included a number of learning objectives that were covered in the whole teaching process. By the end of this course, students were able to describe the basic features of structure and function of the human body

at level of cells, tissues, and organs; discuss the different phases of cell growth and division; express basic skills in dealing with cadavers and shows accepted degree of respect to human dignity; and recognize the basics of preserving and maintaining the human body after death (different methods of fixation). We used interactive lectures, practical classes for teaching the learning objectives in addition to sessions of self-directed learning, PBL, and seminars. The learning environment was the same for all participants in the study. Students were categorized into two groups:

Group 1 (n=40): Received the course in January 2017 and exam was performed upon course completion. Test was created using the blueprint method that utilizes overall assessment time and time allocated for each examinee to answer each type of test items.

Group 2 (n=40): Received the course in January 2018 and performed their exam upon course completion. Test was created using the novel blueprint method that utilizes the adopted credit hour policy and its synchrony with the academic load of curriculum.

All factors that may possibly affect students' performance or their academic achievements were relatively constant except for methods used in blueprint creation in both groups. Both have same learning tools including study guide (revised and approved by curriculum committee), learning outcomes, and instruction team. Test questions were randomly selected from question bank that possess satisfactory difficulty, discrimination, and reliability indices to ensure equal assessment environment for both groups.

Study tools

We created a novel blueprint method that utilizes the adopted credit hours system used in our integrated teaching curriculum. It correlates the course contents (themes) to the learning cognitive skills that are divided into three domains: knowledge, understanding, and application. This two-dimensional matrix of blueprint creation started by constructing a TOS that shows what will be examined in relation to the learning objectives. The curriculum was divided into themes (topics). The themes are weighted according to the number of learning objectives and the total number of contact hours assigned for the learning objectives.

The total number of contact hours is calculated and synchronized with the adopted credit hour system used in our curriculum. The whole length of the learning block (basic structure and function) is calculated. Each block consists of interactive lectures, practical sessions including skill laboratories, and PBL tutorials. For each theoretical activity (interactive lectures and PBL sessions), a ratio of 1:1 of contact to credit hours is utilized (60 minutes for 18 weeks equivalent to 1 credit hour of theory instruction or the equivalent per term). For each practical session (skill laboratory), a ratio of 2:1 of contact to credit hours is utilized (120 minutes for 18 weeks equivalent to 1 credit hour of practical instruction or equivalent per term) (Table 1).[17] This reduces subjectivity in weighting process as contact and credit hours are already defined by raters. For a 3 credit hour course (2 theories and 1 practical), we adopted 10 questions for each credit hour with a total number of 30 questions. Questions are distributed in relation to learning domains according to the number of leaning outcomes in each domain (knowledge 25%, understanding 35%, and application 40%) (Table 1). The overall time of assessment is calculated according to number and type of questions assigned for testing outcomes (1 minute for multiple choice question [MCQ] type A, 2 minutes for complex type of MCQs) with an overall time of 42 minutes (18 type A and 12 complex type).[18]

Table 1 Example of created blueprint table using the credit hours system

Input and output table for created blueprint using credit hours						
Input data				**Output data (no. of questions)**		
Themes (no. of SLOs =150)	**Theoretical (contact hours)**	**Practical (contact hours)**	**PBL (contact hours)**	**Learning domains**		
				Knowledge 25%	**Understanding 35%**	**Application 40%**
T1 (35, 23%)	4	7	4	2	3	2
T2 (30, 20%)	4	7	4	1	2	3
T3 (40, 27%)	6	8	4	2	3	3
T4 (45, 30%)	5	14	4	2	3	4
Credit hours	1.2	1	0.8	Total no. of credit hours: 3 Total no. of questions: 30 MCQs (A type and complex type)		

Notes: For each 18 hours of theoretical and PBL teaching =1 credit hour, each 36 hours of practical teaching =1 credit hour. Percentage of knowledge, understanding, and application learning domains is calculated upon their representation in the learning outcomes.

Abbreviations: MCQs, multiple choice questions; PBL, problem-based learning; SLOs, specific learning outcomes; T, theme.

In group 1, test blueprint was created by utilizing the overall time allocated for the exam (42 minutes) and time assigned for each examinee to answer each type of questions. Tabulation of course into themes, weight of each theme, and proportionate weight of domains to be assessed were calculated as described before. The total number of test items was calculated by using the following formula: $N_i = T\ (Wi)/\sum^{1}{}_{i=1}\ (t_i \times W_i)$, where N is the number of items needed for knowledge (i=1), understanding (i=2), and application (i=3); T is the total time (in minutes) available for running the assessment; W is the assigned weight for knowledge, understanding, and application; t_i is the time allocated for examinee to answer the item that tests knowledge, understanding, and application. This method allows calculation of test items in relation to exam duration to achieve better reliability and discrimination indices (Table 2). The validity of utilized test blueprint methods was evaluated by expert in the field of medical education and by testing its efficacy in other courses. Sample of constructed MCQs in relation to specific learning domains and learning objectives is presented in Table 3.

Assessment and evaluation tools

We aimed to detect the impact of a newly applied method for test blueprint creation on intended outcomes. Evaluation of intervention and effectiveness of an interventional procedure were studied by Kaufman and Keller (1994) and further modified by Freeth et al (2002).[19,20] They described four levels of outcome evaluation including reaction, learning, behavior, and results. In this work, we used this model for evaluating our interventional procedure (blueprint design). Students' reactions toward the learning outcomes, their degree of satisfaction, behavior in the learning process, results, and impaction of performance were all measured during the formative assessment conducted throughout the whole course (in PBL, seminars, team based learning, and self-directed learning). Thus, the remaining level of evaluation, which is learning was measured during their summative assessment and reflected by students' scores and test item analysis results. The learning level of students and their test performance were evaluated by using the exam item analysis reports including difficulty (percentage of students who answer an item correctly), discrimination (Pearson product moment correlation between student responses to a particular item and total test scores on all other items of the test), and reliability (extent to which the test is likely to produce consistent scores) indices (using Apperson DataLink 3000, Serial No. B04524, CA, USA). Comparison between both groups' results was conducted with appropriate statistical test.

Ethical approval

Item analysis reports and students performance results were retrieved from administrative records after approval from the program director and dean of faculty. Confidentiality of the study participants was maintained throughout the study. The study protocol was approved by the local ethical review board (University of Bisha, College of Medicine, R/15.03.135).

Statistical analysis

Quantitative data were presented in terms of mean and SD. Item analysis differences between both groups were evaluated by Student's t-test for statistical significance. $P<0.05$ was considered significant. The effect size was calculated using Cohen's equation: d = Mean1 (gp1) – Mean2 (gp2)/Avg SD, where Avg SD is the average of both SDs. Cohen's d value of 0–0.2 SDs means small effect, 0.2–0.5 medium effects, and >0.5 large effects.[21] Measure for test reliability used was Cronbach's alpha (general form for commonly reported KR20). When coefficient alpha is applied to test with single correct answer (all correct answers worth the same number

Table 2 Example of created exam blueprint using time allocated for running assessment

Themes (no. of SLOs =150)	Learning domains			Totals
	Knowledge 25%	Understanding 35%	Application 40%	
T1 (35, 23%)	1	2	2	5
T2 (30, 20%)	1	1	1	3
T3 (40, 27%)	1	2	3	6
T4 (45, 30%)	1	3	3	7
Totals	4	8	9	21
Assessment tool	MCQ (A type and complex type)			
Time allocated for MCQ	A type: 1 minutes Complex type: 2 minutes			
Total time of assessment	42 minutes			

Abbreviations: MCQ, multiple choice question; SLOs, specific learning outcomes; T, theme.

Table 3 MCQs in relation to specific learning outcomes and domains

Activity: lecture	Educational outcome	Learning domains		
		Recall	Cognitive	Psychomotor
Cell structure and cell cycle regulation	1. Define the cell and describe its components Which of the following is a cellular nonmembranous organelle? A. Golgi bodies B. ER C. Mitochondria D. Ribosomes	MCQ		
	2. Describe the structure and functions of cell organelles Which of the following organelles is responsible for packaging and transport of proteins across the cytoplasm? A. Ribosomes B. Smooth ER C. Rough ER D. Gogli apparatus		MCQ	
	3. Correlate the structure of the cells to the function of the constituting organs Which of the following cells have contractile proteins and produces movement? A. Osteocytes B. Chondrocytes C. Myocytes D. Granulocytes		MCQ	
	4. Explain the roles of checkpoints, cyclin, Cdks, and MPF in cell cycle control The passage of a cell through the stages of the cell cycle is controlled by protein kinases that phosphorylate many different proteins at appropriate times. What are these protein kinases called? A. Cdk-activating kinases B. Cyclin-dependent kinases C. Cyclins D. Tyrosine kinases		MCQ	

Abbreviations: ER, endoplasmic reticulum; MCQ, multiple choice question; MPF, maturation promoting factor.

of points), the resulting coefficient is equal to KR20.[22] Data processing was carried out with SPSS, 17.0, Statistical package for windows.

Results

A total of 80 students were included in the study. All of them performed the final exam of the basic structure and function course. No absentee was reported among participants. Test construction based on blueprint creation considered the total number of credit hours adopted for the whole course in group 2 in comparison with considering the overall time assigned for running the assessment and time allocated for each examinee to answer different item types in group 1. Table 1 shows that learning objectives, weight of each theme in relation to importance and total number of SLOs (specific learning outcomes), actual contact and credit hours, and utilization of three distinctive learning domains (knowledge, understanding, and application) were utilized in blueprint creation. The overall number of test items was 30 questions (related to 3 actual credit hours), and the overall time assigned for assessment was 42 minutes (1 minute for type A MCQ, 2 minutes for complex type questions). Table 2 shows that the overall duration of the exam and the allocated duration for each item type were utilized in blueprint creation. A total of 21 test items were considered with an overall duration of 42 minutes for running the assessment. Questions were distributed among the learning domains, and weight of each theme was also considered.

Table 4 depicts the overall difficulty, discrimination, and reliability indices of both groups. There was no significant difference in difficulty (0.6±0.18, 0.57±0.18) or discrimination (0.38±0.16, 0.36±0.2) indices between both groups (Figure 1, $P>0.05$). Internal consistency reliability as indicated by KR20 value showed no significant deviation

Table 4 Students' performance and exam item analysis reports' in both groups

Parameters	Group 1 (credit hours)	Group 2 (exam time)	*P*-value	95% CI	Cohen's *d*
Exam overall difficulty index	0.6±0.18	0.57±0.18	0.301	0.029–0.112	0.16
Exam overall discrimination index	0.38±0.16	0.36±0.2	0.357	0.02–0.11	0.11
KR20 of both tests	0.72	0.75			

Notes: Data are expressed as mean ± SD. $P<0.05$ is considered significant. Cohen's *d* value of 0–0.2 means small effect, 0.2–0.5 medium effect, and >0.5 large effect.

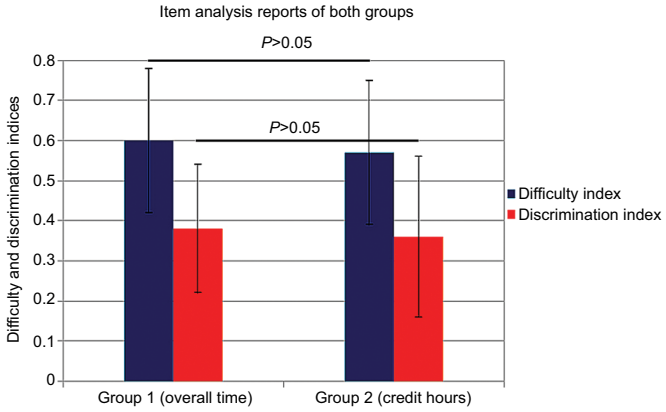

Figure 1 Item analysis reports including difficulty and discrimination indices of group 1 (overall time) and group 2 (credit hours).
Notes: Results show no significant difference between both groups as indicated by $P>0.05$. Data are expressed as mean ± SD.

between both groups (0.72, 0.75, respectively) with an overall good reliability index. Both exams evaluated the same construct, thus achieving construct validity of both study groups. Nonmeaningful difference was found between both groups with regard to difficulty or discrimination indices as indicated by Cohen's *d* value of 0.16 and 0.11, respectively, denoting that different blueprinting methods (applied in our test) have small or minimal effect on difficulty and discrimination indices of exam.

Discussion

Creation of an exam blueprint begins with defining the specific curricular objectives that will be assessed.[9,12] We started with the predefined educational milestones and outcomes stated in our study guide that was approved and revised by our curriculum committee. This provided a framework for developing a competency-based blueprint. Blue print is not an assessment, but it is a guide for it that ensures each learning activity and domain are well represented in the final exam according to their actual importance and weightage in conduction.

Our blueprint creation considered the following steps in its construction: 1) relation of institutional objectives and milestones to the test items; 2) utilization of three learning domains (knowledge, understanding, and application); 3) relative weight of each learning theme (topic) and domain considering the number of theme objective relative to all course objectives and the actual contact hours that reflect its importance in curriculum; 4) adopted credit hours for the whole course; 5) overall time for running the assessment and time allocated for each examinee to answer the test items. Considering all the items for creating a test blueprint ensures a considerable degree of test validity and reliability as each test item is concordant with institutional objectives and milestones and even contains similar text and descriptors as stated in student' study guide. Time effect was nullified by relating the number of test items to the overall duration of exam, and this achieves considerable degree of test reliability (KR20 values are 0.72 and 0.75 for both tests, respectively). Thus, blueprinting is an essential tool in educational curriculum design, which ensures true evaluation of intended learning outcomes.[12]

In the current work, we utilized two methods for creating a test blueprint. One depends on the total number of calculated test credit hours (based on actual theoretical and practical contact hours, ten questions for each credit hour) and the other (old one) considers the overall exam duration assigned for running the assessment and time allocated for answering each item. For a total of 3 credit hours course, we got 30 questions for group 1 (18 type A MCQs and 12 complex time questions, Table 1) with an overall time of 42 minutes.

Thus, in this group, calculation of total test items precedes the assessment time calculation. However, in the second group, we utilized the overall time that was assigned for running the assessment at first (42 minutes) and then we calculated the total number of test items (21 questions, Table 2) considering the time assigned for answering each item type (1 minute for type A MCQs and 2 minutes for complex type MCQs).[18] In this work, we utilized two types of MCQs as sample of assessment tools since our objectives are not clinical; MCQs can achieve good degree of reliability and validity.[23] Our MCQs used in the exam varied in their degrees of difficulty from measuring only simple knowledge domain to other forms, which are more complex (case based or scenarios) that measured higher degrees of thinking rather than memorization. We performed pilot studies before to determine the effective method for students' assessment, and all concluded that this type of assessment best fits with our learning domains and reflects a good degree of reliability. There has been much debate regarding the effectiveness of MCQs in measuring the higher cognitive functions of students and the need for other forms of questions (eg, MEQs) in students' final assessment. Many literatures mentioned that construction of MEQs, which will assess higher order cognitive skills, cannot be assumed to be a simple task. Well-constructed MCQs should be considered a satisfactory replacement for MEQs if the MEQs cannot be designed to adequately test higher order skills. They achieve an accepted level of satisfaction in statistical and intellectual scrutiny. However, others assume that MCQs can be used as a component of assessment of clinical competence; they have part to play in assessment particularly in knowledge domain but other higher cognitive assessment tools are required.[23,24] However, MCQs are not the only type that can be used in our newly developed blueprint method, any other form of questions can be used to match with any learning subject but the time assigned for each examinee to answer a single type of questions will eventually change and hence the time allocated for the overall assessment will subsequently change according to the type of utilized question.[25] We evaluated the usability and effectiveness of our new method applied for blueprinting by comparing results of item analysis reports of both tests. There was no significant difference between both groups regarding the overall difficulty index of exam (0.6±0.18 and 0.57±0.18) in both groups, respectively (P>0.05) (Table 4, Figure 1). This ensures the validity and accuracy of our created blueprint program as there was no decrease in test scores for students in group 1 (credit hours). This passes in concordance with Bridge et al (2003), who stated that lack of congruence between test items and intended learning outcomes can lead to decrease in students' scores. Overall discrimination index of both tests shows no significant deviation (0.38±0.16, 0.36±0.2, respectively) (P>0.05). This shows that our test created with credit hours system achieves a good degree of test homogeneity and high internal consistency with its individual items positively related with total test scores. Effect size of both methods used for blueprint construction was minimal on the difficulty (d=0.16) and discrimination indices (d=0.11) of exam; this confirmed the nonmeaningful difference between both groups and indicated that appropriate combination of items assigned by the blueprint was achieved in both groups.

Thus, it may be concluded that tests that were developed by the newly utilized blueprint method (credit hours system) showed equal degree of reliability, validity, and representativeness when compared with formerly used method dependent on overall exam time and assigned duration for each item type.

Study limitations

The study has limitations; first, the inherent measurement error in any assessment limits the effect of any intervention. Second, our method for blueprint creation should be applied to other courses and the results should be registered. Perception of students toward tests created by our newly developed method should be done and their percentage of satisfaction toward individual test items should be detected and compared with their actual acquisition results registered from their tests' item analysis reports. Finally, though our results are satisfactory and interesting, we have limited number of study participants (guided by total number of students running the course); thus, our results are still tentative and need to be repeated on a larger scale of participants.

Conclusion

Blueprint ensures representative sampling of the curriculum content. Sound blueprinting ensures a good degree of test validity and reliability. Different methods have been incorporated in blueprint construction; many of them are subjected to modification upon institutional educational milestones. However, incorporating time factor allocated for running the assessment, duration for each item type, weight of topics in relation to actual contact hours and number of learning objectives, distinctive domain levels, and relating their content to learning objectives, all are essential factors for achieving test validity and reliability. Linking the results of item analysis with the constructed blueprint provides a method for evaluating and revising the learning outcomes and efficacy of teaching process. Here, in this work, we defined a new method applied in our integrated teaching approach that

depends on total number for course credit hours for creating a test blueprint and compared its validity (both construct and content validity) and reliability (reported by item analysis results) with an already applied method that utilizes overall assessment time and duration assigned for each item type in constructing a test blueprint. No significant difference was found between both methods and both achieved similar degrees of test validity and reliability. Thus, our method could be considered easy and feasible and may eventually be utilized for blueprint construction and implementation.

Acknowledgments

The authors want to give their deepest thanks to the staff of exam office, director of medical education, and vice dean for academic affairs at College of Medicine, University of Bisha, Saudi Arabia, for their great support and efforts in providing the requested data for running this work. Finally, deepest thanks to all members of quality, students' affairs, and assessment committees who spared no effort in helping authors to conduct this work.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Barr RB, Tagg J. From teaching to learning – a new paradigm for undergraduate education. *Change*. 1995;27(6):12–26.
2. Herman JL. *A Practical Guide to Alternative Assessment. Association for Supervision and Curriculum Development*. Alexandria, VA; 1992.
3. Helms JE. Why is there no study of cultural equivalence in standardized cognitive ability testing? *Am Psychol*. 1992;47(9):1083–1101.
4. Guerra-López I. *Evaluating Impact: Evaluation and Continual Improvement for Performance Improvement Practitioners.* Human Resource Development. Amherst, MA: HRD Press: 2007.
5. Spooren P, Brockx B, Mortelmans D. On the validity of student evaluation of teaching: the state of the art. *Rev Educ Res*. 2013;83(4):598–642.
6. McLaughlin K, Coderre S, Woloschuk W, Mandin H. Does blueprint publication affect students' perception of validity of the evaluation process? *Adv Health Sci Educ Theory Pract*. 2005;10(1):15–22.
7. Bridge PD, Musial J, Frank R, Roe T, Sawilowsky S. Measurement practices: methods for developing content-valid student examinations. *Med Teach*. 2003;25(4):414–421.
8. Schuwirth LW, van der Vleuten CP. Programmatic assessment and Kane's validity perspective. *Med Educ*. 2012;46(1):38–48.
9. Sales D, Sturrock A, Boursicot K, Dacre J. Blueprinting for clinical performance deficiencies – lessons and principles from the general medical council's fitness to practise procedures. *Med Teach*. 2010;32(3):e111–e114.
10. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ*. 2004;38(3):327–333.
11. McLaughlin K, Lemaire J, Coderre S. Creating a reliable and valid blueprint for the internal medicine clerkship evaluation. *Med Teach*. 2005;27(6):544–547.
12. Coderre S, Woloschuk W, McLaughlin K. Twelve tips for blueprinting. *Med Teach*. 2009;31(4):322–324.
13. Bloom BS, Krathwohl DR, Masia BB. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. New York, NY: David McKay Company; 1956.
14. Ward Educational Consulting. *Handbook for Test Development*. FL: Ward Educational Consulting Inc; 1983.
15. Oermann MH, Gaberson KB. *Evaluation and Testing in Nursing Education.* NY: Springer Publishing Company; 2016.
16. Bai X, Ola A. A tool for performing item analysis to enhance teaching and learning experiences. *IIS*. 2017;18(1).
17. Morsy MH, Al-Qahtani JM, Al-Ayed MS, Alsareii SA, Alshiek MH, Abdullah M. Credit hours policy – is it working for hybrid problem-based learning curriculum: an experience of Najran School of Medicine KSA. *J Res Med Educ Ethics.* 2015;5(2):129–133.
18. Nitko AJ. *Educational Assessment of Students*. Des Moines, IA: Prentice-Hall; 1996.
19. Kaufman R, Keller JM. Levels of evaluation: beyond Kirkpatrick. *Hum Resource Dev Q*. 1994;5(4):371–380.
20. Freeth D, Hammick M, Koppel I, Reeves S, Barr H. *A Critical Review of Evaluations of Interprofessional Education*. London: LSTN for Health Sciences and Practice, KCL; 2002.
21. Soper DS. Effect Size (Cohen's d) Calculator for a Student t-test [Software]; 2013. Available from: http://www.danielsoper.com/statcalc. Accessed February 6, 2014.
22. Nunnally JC. Psychometric theory – 25 years ago and now. *Educ Res*. 1975;4(10):7–21.
23. 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(1):49.
24. Moss E. Multiple choice questions: their value as an assessment tool. *Curr Opin Anaesthesiol*. 2001;14(6):661–666.
25. Brookhart SM. *How to Assess Higher-Order Thinking Skills in Your Classroom.* Alexandria, VA: ASCD; 2010.