IBEROAMERICAN JOURNAL OF MEDICINE
iberoamericanjm
Journal homepage: www.iberoamericanjm.tk

Review

How to Use and Apply Assessment Tools in Medical Education?

Said Said Elshama[a,*]

[a]*Department of Forensic Medicine and Clinical Toxicology, College of Medicine, Suez Canal University, Ismailia City, Egypt. College of Medicine, Taif University, Taif, Saudi Arabia*

ARTICLE INFO

Article history:
Received 24 July 2020
Received in revised form 01 August 2020
Accepted 10 August 2020

Keywords:
Assessment
Methods
Medical Education

ABSTRACT

Assessment in medical education usually gives the evidence that learning was carried out and the learning objectives were achieved. The assessment program is a measurement tool to evaluate the progress in knowledge, skills, behaviors, and the attitude of students. So, the planning for an effective assessment program should be based on instructional objectives, instructional activities, and efficient assessment methods. Thus, a well-designed assessment procedure should be characterized by validity and reliability. There are two methods for interpreting the results of students' performance, norm-referenced and criterion-referenced; the first gives a relative ranking of students while the second describes learning tasks that students can and cannot perform. The information that gets from the assessment results should be used effectively to evaluate and revise the instructional course for more improvement. Therefore, the reporting of the assessment results to stakeholders should be clear, comprehensive, and understandable to prevent misinterpretation that may affect students and other stakeholders adversely.

1. INTRODUCTION

Assessment is a tool for determining the extent of instruction intended learning outcomes achievement by students; it is considered an integrated process with the instruction process. Moreover, a well-integrated designed assessment gives a credible impression about the effectiveness of the instruction process. In addition, the student assessment leads to student motivation, student self-evaluation development, retention and transfer of the learning [1].

Therefore, the integration of assessment with an instruction should be depending on essential principles for effective assessment. These principles should include clear intended learning outcomes, using the different assessment procedures, the relevance of procedures to instruction, an adequate sample of the student performance, the fairness of procedures, the judgment of successful performance according to specific criteria, the feedback to the students about the strength and weakness of the performance for the correction, the comprehensive grading, and the reporting system. Thus, the choice of assessment method selection should be depending on using the most efficient and

* *Corresponding author.*
E-mail address: saidelshama@yahoo.com

http://doi.org/10.5281/zenodo.3978444

appropriate method for the intended learning outcomes assessment. Noteworthy, student learning improvement is considered the main objective of the assessment program [2].

In this context, the planning for the student assessment should be based on instructional objectives, instructional activities, and assessment methods. So, the instructional objectives should describe the intended learning outcomes in performance terms wherein this performance is evidence of the student learning at the end of the learning experience. Moreover, the revised bloom's taxonomy of educational objectives is considered the framework for identification of the previous factors via two dimensions; the first includes six cognitive process categories (remember, understand, apply, analyze, evaluate and create) while the second includes four knowledge categories (factual, conceptual, procedural and Metacognitive). This taxonomy prepares the assessment procedures and instruments in alignment with the instructional objectives and activities wherein harmony and alignment between objectives (intended learning outcomes), instructional activities, and assessment are the title of effective planning for the student assessment [3].

Worthwhile, the planning of assessment and instruction are complemented each other. So, the planning for them should be done at the same time to have answers for some necessary questions that help for the success of the assessment program such as what is the extent of the need for pretesting?. What is the type of assessment during and at the end of instruction? Therefore, preparation of achievement test should be based on a set of steps that include instructional objectives specification, test specification, construction of the relevant test items, and arrangement of the test items, clear direction preparation, revision and evaluation of the assembled test, administration of the test, and the test item analysis [4].

In the related context, the assessment types may be classified according to timing into placement assessment that is a given test at the beginning of the course to identify the necessary prerequisite skills of the instruction success; it is a pretest that determines entry assessment and covers the intended learning outcomes of the planned instruction. The formative assessment (process-focused) is used for the learner progress monitoring during the instruction by identification the strength and weak points of the student performance; its design depends on measuring the extent of the learning outcomes mastering by the learners in the limited section of instruction wherein its results are a method of the learning improvement. At the end of instruction, the extent of the learning outcomes achievement and the terminal performance of students should be measured by summative assessment (outcome-focused); it is a comprehensive method for the mastering identification or the grades assigning, it aims to provide the student's feedback and evaluation of the instruction effectiveness [5].

2. MAJOR TYPES OF ASSESSMENT

Initially, the assessment includes testing and performance assessments; it is classified into tests for selected response and other for supply response in addition to performance assessments restricted or extended.

Selected response tests measure understanding and thinking skills wherein the student chooses the correct or the best answer (Multiple-choice questions (MCQ), true-false and matching tests). It is a common use because of the administration of a large number of the selected response items to the students' group in a short time with rapid scoring of its results by the hand or machine. Its scoring is completely objective, but it is low in realism because the student selects the response from a given set of the possible answers and then there is a limited response to the listed alternatives. On the other hand, the student can respond by the word, short phrase, or complete essay in supply response tests; it requires more time for its results scoring, but its scoring is more subjective and then personal bias stands against the judgment. It is more realistic in comparison with the selected type because it has great freedom of the response with a moderate structure [6].

Restricted performance assessments assess the performance of highly structured limited task (writing a brief paragraph for a given topic); it is more realistic in comparison with the selected type because it has great freedom of the response with moderate structure as the supply response tests. On the other hand, extended performance assessments assess the comprehensive and less structured performance task (writing a short story); it is high in realism because it simulates the performance in the real world wherein it is an integration of ideas and skills of different learning sources. Noteworthy, the performance assessments usually are time-consuming and depend on the quality performance criteria. Moreover, it is applied by the rating scale or the set of scoring rubrics based on subjective judgment [7].

3. TYPES OF TESTS

MCQ are the most useful selection type item; it is designed to measure simple and complex intended learning outcomes. It consists of the stem (problem situation) and several options (choices); the stem is a question or an incomplete statement while options are several answers (correct answer and plausible wrong answers which are called distracters). The best answer form is another type of multiple-choice item for more complex achievement wherein all options are partially correct but one option is clearly better than the others [8].

To prepare the effective multiple-choice item, it should be the design of the item for one intended learning outcome measurement. Furthermore, the stem of the item should be a single clear problem formulation with simple and clear

language along with much wording in the stem of the item with avoidance of the repeated same material in the options. Moreover, the item stem should be in a positive form emphasizing the negative wording using underline or capitalization or its near position for the statement end. The intended answer may be correct or clearly best wherein all options are consistent with the item stem grammatically and parallel in the form avoiding the verbal clues to prevent discrimination of the correct or incorrect answer such as similarity of the wording in the stem and correct answer, stereotyped phraseology of the correct answer, great detail of the correct answer, absolute terms in the distracters "always, never, all, none" or there are two inclusive responses or two responses have the same meaning. Moreover, the correct answer length should vary as well as the position of the correct answer should vary randomly, besides avoidance using the phrase "all of the above" as an alternative, but the phrase "none of the above" should be used with extreme caution. In addition, the difficulty of the item is controlled by the complexity of the stem problem or by the homogeneity of alternatives. Each item should be independent for other items in the test along with the application of normal rules of grammar and using the efficient item format [9].

In addition, distracters should be plausible and attractive to the uninformed; it should be stated in the student language with good sound words and similar to the correct answer in the length and complexity of wording. Distracters should represent common misconceptions or errors of students; it should be homogenous and has extraneous clues without overusing. Noteworthy, breaking any one of the above rules may be encouraged if it will improve the item effectiveness according to experiences of the test maker in the item writing [10].

Matching items type is a simple variation of multiple-choice items wherein it should shift to matching items when there are a number of related similar factors. Matching items type is a series of stems (premises) and series of answers (responses) which are arranged in the columns under the guiding directions for the matching. The matching items type should include matching item material homogeneity and a shortlist of items with brief responses on the right. Moreover, the number of responses should be larger or smaller than premises with responses using more than once and placed in alphabetical or numerical order. Directions should be specific and a basis for matching wherein it should indicate that the use of response may be once or more than once, or not at all. Worthwhile, the matching items should be placed on the same page with the responses [11].

The extended matching question (EMQ) is different from the single best answer multiple-choice questions and superior to it for the assessment of the problem-solving and clinical reasoning skills of the students. It consists of a theme (symptom, diagnosis, treatment), options list (answers), lead-in statement (question), and two stems (two clinical problems) [12].

4. HOW TO ASSESS THE PSYCHOMOTOR DOMAIN IN MEDICAL EDUCATION?

Objective Structured Clinical Examination (OSCE) is considered the used tool for assessment of the psychomotor domain; it is an examination for competence assessment (content skills, process skills, and clinical management). It is considered the standardized tool for clinical competencies assessment such as history taking, physical examination, and technical procedures. According to the Millers pyramid, OSCE measures the category which is called shows how; it consists of multiple stations and a wide sampling of clinical and communication skills with a lot of examiners and patients within a limited time by using a checklist or global rating scale. Therefore, it has high reliability because the use of detailed checklists may decrease inter-rater unreliability and then reinforces the reliability. In addition, the test results depend on the direct observation and the repeated measurements that help the examiner to assess many different qualitative aspects such as efficiency and the students' skill performance. Moreover, there is also acceptability for this exam because every student does the same task. It is also a valid exam depending on content (good sampling of matching skills with the learning outcomes), construct validity, and authentic length of the station [13].

To design a good OSCE, it should determine the examined skills types in alignment with the learning objectives of the module and the types of assessment tools (ex. checklist). Moreover, it also should determine the number of stations (10-15 stations), the time of station and the length of examination (10 minutes X 10 stations = 100 minutes) besides the preparation of resources such as examination rooms, manikins, examiners, patients, and volunteers [14].

Furthermore, the marks scheme should be constructed depending on discrimination actions to distinguish between good and poor performance. In addition, the preparation of instructions is also considered essential for the examiner, patient, and student. At first, it should outline the required task exactly at every station for the student along with outlining the marking scheme instructions about the action and performance of the student at every station for the examiner. Secondly, it should outline the dealing approach between the patient and the student. Finally, it should evaluate the exam after finishing it. Noteworthy, the success OSCE depends on the availability of facilities such as manikins and other tools, examiners, real patients, actors, technical and administrative teams, and training [15].

At last but not least, the use of short stations in the OSCE is considered a controversial issue wherein some educators think that it is destructive to the validity of the test. Some educators adopt this view because the use of short stations does not allow to assess other aspects of shows how level such as the ability of students to deal with complicated situations that need the integrated different skills such as decision making, drawing the conclusions based on physical examination and investigation and management

skills of the case. Thus, the use of short stations becomes limited to the technical skills only according to some views. On another hand, other educators prefer the use of long stations as an alternative indicating the limited influence of the station length on the reliability. Therefore, I think that the best is the determination of the assessment task by using a good balance for the content apart from the controversial views to ensure the authenticity and the efficiency of measurement [16].

5. HOW TO ASSESS THE AFFECTIVE DOMAIN IN MEDICAL EDUCATION?

Worthwhile, the performance tasks usually contain knowledge, skill, and the affective components (affective domain) that describes the learning objectives which address feeling, emotion, and the degree of acceptance or rejection. Moreover, the affective domain has many parameters such as attitude that is an important mental parameter of the affective domain; it consists of cognition, affects, behavioral intentions and evaluation. The second parameter of the affective domain is the motivation that means initiation, direction, and human behavior persistence; it includes also engaging reasons in a special behavior such as basic needs, object, goal, and the desirable ideal. Thirdly, another parameter is the self-efficacy that is considered a personal perception for the ability of performance in a particular manner [17].

Thus, the affective domain is difficult to assess because it emphasizes attitude, feeling, emotion, and values. So, it should be stated in specific, measurable, observable objectives to translate into quantitative terms. Therefore, the taxonomy of affective domain classifies the behavioral objectives into observable behaviors in the quantitative terms such as receiving (accept, attend, recognize), responding (discuss, complete, examine), valuing (accept, seek, defend), organization (discriminate, organize, systematize), and characterization (verify, internalize) [18].

In this context, the assessment of affective domain depends on many tools that assess attitudes, interests, motivations, and self-efficacy. These tools include self-report, rating scales, semantic differential scales, Thurstone scale, and checklist. The self-report is written reflections that are done by an individual about his attitude or feeling toward an idea or people or concept while the rating scales are a number of the designed categories to extract the quantitative information such as Likert scale and 1-10 rating scale. Semantic differential scales "SD" assess the personal reaction to specific ideas or concepts in rating terms on bipolar scales while the Thurstone scale assesses the attitude by determination favorability position on the issue [19].

6. HOW TO ASSESS THE COMPREHENSIVE DOMAIN "COMBINED DOMAINS" IN MEDICAL EDUCATION?

Portfolio-based assessment is a live alternative to traditional high stakes testing. So, it is used for summative and formative assessment wherein it has value as a source of self-satisfaction. The portfolio is considered one of the useful and popular assessment tools of the student performance in undergraduate and postgraduate medical education; it aims to link the objectives of instructional course with clinical experience that is recorded in a standardized manner to facilitate the learning, teaching, and assessment [20].

The portfolio is a collection of systematic, selected, purposeful and organized student work (materials) that show the personal ability of every student (evidence of performance) and his professional development via measuring the growth of knowledge, skills, and attitudes. Therefore, the content of the portfolio (evidence of the learning achievement) consists of clinical tutor reports, selected student assignments, a list of attained skills, and evidence of communication skills, assessment results, and the reflective diary [21].

In this context, we can divide the portfolio into two types; developmental and showcase portfolio. The developmental type is usually used throughout the instructional course (formative) and assesses the student learning progress while the showcase type is used at the end of the course (summative) and shows the student's best work samples and the final level of performance [22].

In addition, portfolios have many advantages such as the learning progress assessment over the times, positive effect for the coverage of the best student work, and providing the greater motivation because of comparison between the present and past work. Furthermore, its advantages include an improvement in the self-assessment skills of the student, providing reflective learning, adjustment of the individual differences, providing the connection between theory and practice besides communication with the students and parents for the learning progress, and an increase in collaboration between student and teacher. However and for fair judgment, we should remind that portfolios have some disadvantages such as the time consuming because of the portfolio entries selection, periodic revision, and providing the feedback [23].

To plan the portfolios, there are many steps that should be applied such as determination of the portfolio purpose and the involved entries types with a determination of the guidelines for entries selection and evaluation. In addition, it should also determine the procedures of portfolio maintenance and using, and the criteria of portfolio evaluation. Finally, we should discriminate between portfolio evaluation as a structure and the student evaluation as performance progress. The structural evaluation of the portfolio depends on makeup, organization, and content while overall evaluation of the student performance progress that is shown in the portfolio

is determined via the rating scale based on the learning outcomes assessment. Thus, the holistic rubrics of each involved area in the portfolio determine the final level of student performance [24].

7. HOW TO DEAL WITH THE ASSESSMENT RESULTS?

Firstly, the assessment results should be summarized concisely into informative data such as tallies, percentages, and qualitative data (themes, grouped listings). Secondly, the assessment results should be sharing as a summarization for these results or in a brief report associated with essential information such as identification of the successful student rules, satisfactory evidence for his success, and the determined action for unsatisfactory results. Moreover, the venues of the assessment results sharing should be determined via choosing one venue or more such as web sites, emails, newsletters, presentations, brochures, posters, or banners [25].

In this context, the reporting of assessment results should be fair, honest, balanced, objectively, useful, and documented with providing appropriate attribution. So, it should give the most impact via using the meaningful, attractive, interesting title and headings. Furthermore, the reporting of assessment results should be short, cascade from major points to details with informed commentary. In the related context, grading of results is also considered an essential element because it provides us effective feedback about the learning process and the suggestions for its improvement wherein assigning grades are a valid measure for learner achievement [26].

Noteworthy, the performance assessment has different types such as essay tests, ratings, and multiple-choice questions wherein it translates the student performance to grades that represent the extent or degree of intended learning outcomes achievement. Therefore, every medical school should be having a clear grading policy for valid judgment. Moreover, grading may be divided into two types; the first is an absolute grading while the second is relative grading. Absolute grading is based on a comparison between the student performance and pre-specified standard of performance depending on the mastering of the learning and cutoff points identification while the relative grading depends on a comparison between the student performance and the group members' performance for individual ranking in the group [27].

In addition, the validity of the grading system should be based on the efficacy and fairness of the assigning grades. Therefore, there are some guidelines that should be applied during the designing of the grading system. Initially, the students should be aware of the grading system of the course achievement at the beginning of the course including components of assessment, the weight of every test grade, and the description of every letter grade. Worthwhile, these guidelines should be written in detail in the study guide of every module. Secondly, grades should be based on student achievement only without addition to extraneous factors such as effort or misbehavior. Thirdly, grades should also be based on varieties of valid assessment data and all learning outcomes while the results should be involved in the final grade for more validity of the grade. Fourthly, the weighting method should be used for combining scores of the grading with a selection of a suitable frame for the grading reference. Finally, the revision of the borderline cases should be done by re-examining all achievement evidence [28].

However, the results or test scores interpretation is an important step in dealing with the assessment results wherein it is considered a translation of the quantitative data to equal numerical set; it is a process for score analysis to generate meaningful quality. Noteworthy, there are different types of scores; the first is the raw score that is a number of the received points in the test that have not meaningful interpretation while the second is the scaled score that is a result transformation through a consistent scale. In addition, the test score interpretation should depend on the referencing framework that is a structure for comparison of the student performance to something external to the assessment itself; it is a comparison of the student score to the predetermined standard of performance (standard criteria) [29].

Thus, the referencing framework for the test score interpretation may be a criterion-referenced framework or norm-referenced framework. The criterion-referenced framework is the description of individual performance in the test without referring to the performance of others wherein the criterion is the domain of performance that is a reference of the student assessment results. Worthwhile, this interpretation is meaningful if the test is designed specifically for this purpose. So, the test performance using criterion-referenced assessment can be measured by the speed of performance (task performance within a fixed time), the degree of performance accuracy, the percentage (proportions number of maximum points gained) such as the percentage of the corrected answers or the percentage of the learning objectives achievement, the quality rankings (quality level of performance such as an excellent rating of 4, good rating of 3), the percentage of the correct score (standard for judgment of the performance mastering of the learning objectives), and the expectancy table (it interprets raw score in expected performance terms) [30].

The norm-referenced framework is a comparison of the individual test score with other students' test scores who take the same test. Therefore, it determines the student standing in the reference group wherein the student score is not treated individually but it is related to the group. Moreover, norm-referenced scores depend on the transformation of the raw score mathematically wherein the raw score in the norm-referenced framework is not valid for the student performance interpretation. So, it should be converted into the derived score that is a numerical report of the test performance on the score scale. The percentage of the norm group that is scored below a particular raw score is identified as percentile ranks; it is

different from the percentage of corrected answers items that is criterion-referenced interpretation. Developmental scores or scales are one of the norm-referenced scores that identify the development of students across various grades or age levels wherein the grade equivalent score is matching the particular raw score that equals the obtaining grade level of the student. The standardized scores of norm-referenced scores are transforming scores for the test performance comparison across two or more different measures; it divides into linear standard scores and normalized standard scores wherein the linear standardized scores (Z-scores and T-scores) compare between two distributions of the test performance and maintain the same distribution shape of corresponding raw scores while the normalized standard scores (stanines and deviation IQ scores) depend on the knowledge of normal distribution characters in the interpretation and convert the distribution of the raw scores to normal distribution. Finally, I want to remind that all norm-referenced scores contain errors because there is not test act as a perfect measure [31].

Finally and conclusively, there is not a gold standard-setting in the assessment. According to the above mentioned, there are two types of standard-setting methods; criterion-referenced or absolute method, wherein the standard-setting does not depend on the test results (independent) while norm-referenced or relative method wherein the standard setting is based on the test results. The norm-referenced standard is considered the method of choice to rank examinees while the criterion-referenced standard is considered the most appropriate to fulfill whether examinees' mastering of a specific domain meets the pre-set requirements. Regrettably, two standard-setting approaches have disadvantages that diminish their credibility because it leads to widely divergent results on the same test. The criterion-referenced method with a pre-fixed cut-off score leads to a large variation in failure rates while the norm-referenced method leads to a large variation in cut-off scores. In addition, the procedures of a criterion-referenced standard setting require panels to determine a minimum acceptable level per test item. Moreover, these procedures are considered time-consuming and costly. So, the cut-off scores are established in the form of a pre-fixed percentage of the corrected answers of test questions because of the inability to use regularly the panels for standard-setting procedures. However, merging a pre-fixed cut-off score with a relative point of reference as a compromise method may reduce the disadvantages of conventional criterion and norm-referenced methods besides making the optimal use of their advantages [32].

So, every educational institution should have a vision for the interpretation of the assessment results; this vision should determine benchmarks or standards wherein the interpretation of assessment results should be based on it. Benchmark or standard may be local, external, internal, value-added, historical trends, strengths and weaknesses perspective, and capability or productivity. According to the benchmark or standard choice, we can compare our students with their peers inside or outside the institution at a national or international level and determine what the extent of the improvement achievement for the students or the educational program, the strength and weakness points, capability and productivity of the students. However, some schools adopt standardized achievement tests that depend on the norm-referenced approach to interpret their results. It compares the student performance to the representative sample of students' performance in the norm group at a regional or national level; it is designed to determine the common set of goals achievement by the students. So, there are some guidelines that should be applied when standardized achievement tests are constructed. At first, the test content should be depending on many the used textbooks besides the test items should be constructed by test experts and subject matters. Moreover, the test items should also be selected depending on the test specifications, and then it is revised and analyzed for the difficulty via using the rigid directions for the test. In addition, the test scores should be interpreting according to the norm-referenced framework whereas the test manual should be included the procedures of scoring, interpretation, and the use of results. Finally, we can modify the standardized achievement test and interpret its scores according to the criterion-referenced framework if we can modify multiple-choice items and add open-ended performance task [33].

Noteworthy, the percentage of the correct score is considered one of the best methods of reporting of the criterion-referenced test results wherein it tells us about the percentage of corrected answers in the test. However, the norm-referenced scores have different types that are used with standardized tests such as percentile ranks, grade equivalent scores, and the standard scores. The percentile rank is different from the percentage of the corrected answers (criterion-referenced) because it indicates the relative position in the group as a percentage of students scoring while the grade equivalent scores indicate the relative test performance as a grade level. The standard scores depend on statistics such as mean and standard deviation of the scores set [34].

On the other hand, the assessment feedback is important for the stakeholders such as students, parents, and the educational authority wherein its importance for the students and parents is determining the level of achievement and the position of students among their peers. In addition, it is also important for the governmental educational administrators to evaluate the instruction and the learning process, the extent of learning outcomes achievement, and the success of the educational policy of this medical school. Thus, we should use a detailed reporting system about the performance of the learning outcomes of the course [35].

In the end, the report of results should be comprehensive, well organized in an arranged manner without lengthening and confusion issues, rating the performance, and informative based on the list of specific learning outcomes. However, the report format choice depends on the report

material and audience. So, we can use a full report as a complete assessment activities record or assessment summary as a note, brochure, or flyer to highlight the particular findings or specific issues. Thus, the components of the assessment report should include a description of activities, results interpretation, and suggestions. Moreover, the determination of audience or stakeholders should be known before the determination of content, format and the method of assessment results reporting because every stakeholder needs different content and style of the results report according to his scope such as accrediting organization, higher education commission, medical education committee, students, and the parents. Furthermore, the assessment results may be used as a method for curriculum evaluation and revision or accreditation or employment. Therefore, web reporting is considered one easy access wherein it is used for a wide range of audiences [36].

At last, we would like to mention that communication of the assessment results should be clear, understandable, interesting, explainable, and appropriate for the content. Thus, it may be a chart, table, or graph according to the available data. Effective tables and charts should have a meaningful and self-explanatory title and content with a clear label for every table or chart. Moreover, the results should be classified into groups if it is much, and it should be easy for the readers to detect the differences and trends. At the end of this paragraph, we should refer that the confidentiality of the assessment result reporting is a title of the participant's credibility in the assessment process [37].

8. HOW TO DESIGN A SUCCESSFUL ASSESSMENT PROGRAM?

Continuing with what we started, we can summarize the ingredients for designing a successful assessment program for the medical student. At first, the rules and procedures of assessment should be clear to the students at the beginning of the module; it should also be involved in the study guide of the module. Secondly, using a well-designed assessment procedure that is characterized by validity and reliability; the validity means appropriate and meaningfulness of inferences that extracted from the assessment results for the intended use, it should include the content that means the representative of the learning objectives in the assessment and congruence of the assessment instrument with the purpose (construct validity). Moreover, it should also include the predictive validity that means the ability of the instrument to predict performance in the future besides the reliability of an assessment that is the consistency of the assessment results which can be interpreted by norm-referenced or criterion-referenced, it is a necessary prerequisite of the valid test. Noteworthy, a highly reliable test doesn't mean necessary its validity. In addition, we can divide reliability into many types; the inter-rater reliability means consistency of the performance rating by different examiners (raters) while the inter-case reliability is a measurement of the student performance from one case to another with consistent variables. Furthermore, the test-retest reliability is measured by the correlation of one score with others; it is an indicator of consistency over time. Worthwhile, increasing the testing time and the number of questions are considered methods for improvement of examination reliability. In the related context, the acceptability of the instrument for the users determines its usefulness to measure what it is supposed to measure (face validity) besides the utility of assessment instrument that should be depending on the reliability, validity, educational impact, costs, and the acceptability of method [38].

Thirdly, the choosing of an assessment instrument for any examination should be depending on multiple levels of clinical competence that are suggested by Miller (Millers Pyramid). MCQ, Essay, and Oral exam are suitable instruments to test knowledge (knows) while clinical scenarios based MCQ, Oral exam, and the Extended matching items are suitable assessment instruments to test understanding and concept building (knows how). Moreover, the OSCE and the standardized patient are suitable to test the performance (shows how) while the performance log (logbook), checklist, and portfolio are suitable to test the concerned task performance in a real-life situation (does). Thus, it should choose one or two assessment instruments from each level to reflect the real ability of examinee [39].

Fourthly, it should use the blueprinting for the tested objectives specification and determination of its relative weight in the examination wherein the table of specification is the blueprint of the test; it identifies the types of test items that should be included in the test according to the time spent and the cognitive level of every objective. So, it should align the summative test with the studied subject matter and the used cognitive process during the instruction. Worthwhile, the table of specification improves the validity of the test that is based on the quality of the evidence (test content and response process); the test content is the studied subject matter while the response process is the kind of thinking that is required in the test. In addition, there are many approaches to develop and use the table of specification; one approach of them depends on a selection of the tested learning outcomes wherein we can select and put the learning objectives according to the terms of Bloom's taxonomy in the cognitive domain [40].

Fifthly, a referencing framework should be applied to get accurate and useful results interpretations. Norm-referenced interpretation is a survey testing to measure the individual differences in the achievement wherein it depends on the other student's performance for determination the passing and fail grade of the given student. On another hand, the criterion-referenced interpretation is a mastery testing to describe the tasks that the student can perform with comparison his performance to a specific achievement domain wherein it depends on the certain determining level of knowledge or skill for

passing the exam. Noteworthy, the criterion-referenced framework does not depend on other performances of examinees but it is based on the particular examinee performance [41].

In addition, the standard sitting may be used that is a special boundary one score to determine who performs well and who does not wherein the credibility of the standard is different according to who sets the standard, characters of the used methods, and the outcome. In the end, the assessment should have feasibility that depends on the availability of resources such as availability of the time for test development, test administration, analysis of papers, availability of training for examiners and the costs [42].

9. CONCLUSIONS

Assessment in medical education is a tool to evaluate the learning process through the student assessment. The assessment program evaluates the medical student in different domains such as cognitive, psychomotor, and affective via using tests for the selected response and other for the supply response in addition to the performance assessments restricted or extended. So, the planning for a well-designed assessment program should be based on effective ingredients for the success wherein it should be characterized by validity and reliability. Moreover, interpretation and reporting of the assessment results to stakeholders should be clear, comprehensive, and understandable to enable different stakeholders to evaluate and revise the instructional course effectively for more improvement.

10. REFERENCES

1. Wolming S, Wikstrom C. The concept of validity in theory and practice. Assess Educ Princ Pol Pract. 2010;17(2):117-32. doi: 10.1080/09695941003693856.

2. Gronlund NE. Assessment of Student Achievement. 8th ed. Pearson USA; 2006.

3. Amin Z, Seng CY, Eng KH. Practical Guide to Medical Student Assessment. World Scientific Publishing Co. Pte. Ltd. Singapore; 2006.

4. Elshama SS. How to Develop Medical Education (Implementation View). 1st ed. Scholars' Press Germany; 2016.

5. Begum N, Hossain S, Talukder MH. Influence of formative assessment on summative assessment in undergraduate medical students. Bangladesh J Med Educ. 2013;4(1):16-9. doi: 10.3329/bjme.v4i1.32191.

6. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ. 2004;38(9):974-979. doi: 10.1111/j.1365-2929.2004.01916.x.

7. Nair BR, Parsons K. Performance-based assessment: Innovation in medical education. Arch Med Health Sci. 2014;2:123-5.

8. Schuwirth LW, van der Vleuten CP. ABC of learning and teaching in medicine: Written assessment. BMJ. 2003;326(7390):643-5. doi: 10.1136/bmj.326.7390.643.

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. doi: 10.1186/1472-6920-7-49.

10. Al-Wardy NM. Assessment methods in undergraduate medical education. Sultan Qaboos Univ Med J. 2010;10(2):203-9.

11. Gibbs T, Brigden D, Hellenberg D. Assessment and evaluation in medical education, S Afr Fam Pract. 2006;48(1):5-7. doi: 10.1080/20786204.2006.10873311.

12. Wood EJ. What are Extended Matching Sets Questions? Biosci Educ J. 2003;1(1):1-8. doi: 10.3108/beej.2003.01010002.

13. Carraccio C, Englander R. The objective structured clinical examination: a step in the direction of competency-based evaluation. Arch Pediatr Adolesc Med. 2000;154(7):736-41. doi: 10.1001/archpedi.154.7.736.

14. Zayyan M. Objective structured clinical examination: the assessment of choice. Oman Med J. 2011;26(4):219-22. doi: 10.5001/omj.2011.55.

15. Khan A, Ayub M, Shah Z. An audit of the medical students' perceptions regarding objective structured clinical examination. Educ Res Int. 2016. doi: 10.1155/2016/4806398.

16. Elshama SS. How to Use Simulation in Medical Education. 1st ed. Scholars' Press Germany; 2016.

17. Yanofsky SD, Nyquist JG. Using the Affective Domain to Enhance Teaching of the ACGME Competencies in Anesthesiology Training. J Educ Perioper Med. 2014;12(1):E055.

18. Lurie SJ, Mooney CJ, Lyness JM. Measurement of the general competencies of the accreditation council for graduate medical education: a systematic review. Acad Med. 2009;84(3):301-9. doi: 10.1097/ACM.0b013e3181971f08.

19. Boud D, Falchikov N. Aligning assessment with long-term learning. Assess Eval High Educ. 2006;31(4):399-413. doi: 10.1080/02602930600679050.

20. Thistlethwaite J. How to keep a portfolio. Clin Teach. 2006;3(2):118-23. doi: 10.1111/j.1743-498X.2006.00078.x.

21. Haldane T. "Portfolios" as a method of assessment in medical education. Gastroenterol Hepatol Bed Bench. 2014;7(2):89-93.

22. Roberts C, Newble DI, O'Rourke AJ. Portfolio-based assessments in medical education: are they valid and reliable for summative purposes? Med Educ. 2002;36(10):899-900. doi: 10.1046/j.1365-2923.2002.01288.x.

23. Davis MH, Friedman Ben David M, Harden RM, Howie P, Ker J, McGhee C, et al. Portfolio assessment in medical students' final examinations. Med Teach. 2001;23(4):357-66. doi: 10.1080/01421590120063349.

24. Jenkins L, Mash B, Derese A. Reliability testing of a portfolio assessment tool for postgraduate family medicine training in South Africa. Afr J Prim Health Care Fam Med. 2013;5(1):577. doi: 10.4102/phcfm.v5i1.577.

25. Epstein RM. Assessment in medical education. N Engl J Med. 2007;356(4):387-96. doi: 10.1056/NEJMra054784.

26. McLachlan JC. The relationship between assessment and learning. Med Educ. 2006;40(8):716-7. doi: 10.1111/j.1365-2929.2006.02518.x.

27. Hays R, Gupta TS, Veitch J. The practical value of the standard error of measurement in borderline pass/fail decisions. Med Educ. 2008;42(8):810-5. doi: 10.1111/j.1365-2923.2008.03103.x.

28. Downing SM, Tekian A, Yudkowsky R. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. Teach Learn Med. 2006;18(1):50-7. doi: 10.1207/s15328015tlm1801_11.

29. Muijtjens AM, Schuwirth LW, Cohen-Schotanus J, Thoben AJ, van der Vleuten CP. Benchmarking by cross-institutional comparison of student achievement in a progress test. Med Educ. 2008;42(1):82-88. doi: 10.1111/j.1365-2923.2007.02896.x.

30. Lok B, McNaught C, Young K. Criterion-referenced and norm-referenced assessments: compatibility and complementarity. Assess Eval High Educ. 2016;41(3):450-65. doi: 10.1080/02602938.2015.1022136.

31. McKinley DW, Norcini JJ. How to set standards on performance-based examinations: AMEE Guide No. 85. Med Teach. 2014;36(2):97-110. doi: 10.3109/0142159X.2013.853119.

32. Cohen-Schotanus J, van der Vleuten CP. A standard setting method with the best performing students as point of reference: practical and affordable. Med Teach. 2010;32(2):154-160. doi: 10.3109/01421590903196979.

33. Allen D, Tanner K. Rubrics: tools for making learning goals and evaluation criteria explicit for both teachers and learners. CBE Life Sci Educ. 2006;5(3):197-203. doi: 10.1187/cbe.06-06-0168.

34. Becker DF, Pomplun MR. Technical reporting and documentation. In: Downing SM, Haladyna TM, editors. Handbook of test development. New York: Routledge; 2006:711-24.

35. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. 2004;38(9):1006-1012. doi: 10.1111/j.1365-2929.2004.01932.x.

36. Wong J, Cheung E. Ethics assessment in medical students. Med Teach. 2003;25(1):5-8. doi: 10.1080/0142159021000061341.

37. Kibble JD. Best practices in summative assessment. Adv Physiol Educ. 2017;41(1):110-9. doi: 10.1152/advan.00116.2016.

38. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003;37(9):830-7. doi: 10.1046/j.1365-2923.2003.01594.x.

39. Shumway JM, Harden RM; Association for Medical Education in Europe. AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. Med Teach. 2003;25(6):569-84. doi: 10.1080/0142159032000151907.

40. Schuwirth LW, van der Vleuten CP. General overview of the theories used in assessment: AMEE Guide No. 57. Med Teach. 2011;33(10):783-97. doi: 10.3109/0142159X.2011.611022.

41. Elfaki OA, Salih KMA. Comparison of Two Standard Setting Methods in a Medical Students MCQs Exam in Internal Medicine. American Journal of Medicine and Medical Sciences. 2015;5(4):164-7.

42. Ben-David MF. AMEE Guide No. 18: Standard setting in student assessment, Med Teach. 2000;22(2):120-30. doi: 10.1080/01421590078526.