Educación Médica 25 (2024) 100906

Educación Médica

www.elsevier.es/edumed

REVIEW ARTICLE

Exam blueprinting as a tool to overcome principal validity threats: A scoping review

Hussein Abdellatif[a,b,*], **Amira Ebrahim Alsemeh**[c], **Tarek Khamis**[d], **Mohamed-Rachid Boulassel**[e,f]

[a] *Department of Human and Clinical Anatomy, College of Medicine and Health Sciences, Sultan Qaboos University, Muscat, Sultanate of Oman*
[b] *Anatomy and Embryology Department, Faculty of Medicine, Mansoura University, Mansoura, Egypt*
[c] *Human Anatomy and Embryology Department, Faculty of Medicine, Zagazig University, Zagazig, Egypt*
[d] *Department of Pharmacology and Laboratory of Biotechnology, Faculty of Veterinary Medicine, Zagazig University, Zagazig 44519, Egypt*
[e] *Department of Hematology, Sultan Qaboos University Hospital, College of Medicine and Health Sciences, Sultan Qaboos University, Muscat, Sultanate of Oman*
[f] *Department of Biomedical Science, College of Medicine and Health Sciences, Sultan Qaboos University, Muscat, Sultanate of Oman*

Received 30 November 2023; accepted 16 February 2024
Available online xxxx

KEYWORDS
Blueprinting;
Assessment;
Validity;
Reliability;
Education measurement

Abstract Medical education's teaching and learning process evolves. This requires a change in instruction and assessment. An assessment's validity is threatened by poorly aligned tests or those with irrelevant variance. A well-designed test blueprint bridges gaps and integrates various education pillars, ensuring an accurate representation of learning outcomes and domains. This study reviewed the test blueprint approach in the relevant literature and highlighted its advantages in improving test measures, particularly those for validity and reliability. We searched MEDLINE/PubMed and Scopus for relevant articles; duplicates were eliminated, and those that met our eligibility criteria were chosen. For data extraction, a charting framework was created. Quantitative and qualitative data were reported and analyzed. Thematic analysis was performed for selected studies. To verify the selected studies, experts were consulted.

Out of 487 selected studies, 22 were included in the review, 18 of which focused on blueprinting design and implementation, and 4 of which used blueprinting to improve medical curriculum design and evaluation practices. The review comprised a qualitative study, 2 cohort studies, 8 cross-sectional studies, 6 quasi-experimental studies, 3 review articles, one of which was a practical guide for test blueprint design, and 1 case report. Finally, as evidenced by their findings, the majority of studies addressed the qualities of a good test, the primary threats to

* Corresponding author at: Sultan Qaboos University, College of Medicine and Health Sciences, Department of Human and Clinical Anatomy, P.O. Box: 50, Al-Khodh, Muscat 123, Sultanate of Oman.
E-mail address: h.abdellatif@squ.edu.om (H. Abdellatif).

https://doi.org/10.1016/j.edumed.2024.100906

validity and reliability, the benefits, and the methodologies for designing test blueprints. Nonetheless, a lack of experimental studies was observed. Electronic blueprinting and a blueprint for programmatic assessment approaches were lacking and warrant additional research.

PALABRAS CLAVE
Blueprinting;
evaluación;
validez;
fiabilidad;
medición de la educación

Examen Blueprinting como una herramienta para superar las principales amenazas de validez: una revisión de alcance

Resumen El proceso de enseñanza y aprendizaje de la educación médica evoluciona. Esto requiere un cambio en la instrucción y la evaluación. La validez de una evaluación se ve amenazada por pruebas mal alineadas o por las que tienen una variación irrelevante. Un blueprint de prueba bien diseñado cierra las brechas e integra varios pilares de la educación, asegurando una representación precisa de los resultados y dominios de aprendizaje. En este estudio se examinó el enfoque del blueprint de pruebas en la literatura pertinente y se destacaron sus ventajas en la mejora de las medidas de la evaluación, en particular las de validez y fiabilidad. Se realizaron búsquedas en MEDLINE/PubMed y Scopus por artículos pertinentes; se eliminaron los artículos duplicados y se seleccionaron las que cumplían nuestros criterios de elegibilidad. Para la extracción de datos, se creó un marco de gráficos. Se informaron y analizaron datos cuantitativos y cualitativos. Se realizó un análisis temático de estudios seleccionados. Para verificar los estudios seleccionados, se consultaron expertos. De los 487 estudios seleccionados, 22 se incluyeron en la revisión, 18 de los cuales se centraron en el diseño y la aplicación de los planos de estudios, y cuatro de ellos utilizaron el blueprint de estudios para mejorar las prácticas de diseño y evaluación de los currículos médicos. La revisión consistió en un estudio cualitativo, dos estudios de cohorte, ocho estudios transversales, seis estudios cuasi-experimentales, tres artículos de revisión, uno de los cuales era una guía práctica para el diseño del blueprint de ensayo, y un informe de caso. Por último, como demuestran sus hallazgos, la mayoría de los estudios abordaron las cualidades de una buena prueba, las principales amenazas a la validez y fiabilidad, los beneficios y las metodologías para diseñar planos de pruebas. No obstante, se observó una falta de estudios experimentales. Faltaban planos electrónicos y un blueprint para los enfoques de evaluación programática, lo que justificaba una investigación adicional.

Introduction

Medical education has advanced considerably. Teaching and learning are more scientific and rigorous, the curriculum is based on strong pedagogical principles, and problem-based learning and other active and self-directed learning are the norms. From problem identification to solution provision, teachers' roles have changed. Over the past 30 years, medical schools have faced several issues from society, patients, professionals, and students. They have developed new curricula, learning scenarios, and evaluation tools and understood the importance of staff development. Successful and fascinating innovations have evolved. Knowledge, technical, analytical, and communication skills, interdisciplinary care, counseling, and evidence- and system-based care are needed for effective and efficient healthcare delivery. To adapt to this transition, our assessment methods should be comprehensive, reliable, and robust enough to evaluate the necessary attributes besides knowledge and abilities.[1] In addition, assessment has become an integral component throughout every phase of professional development.

While developing a test, it is imperative to consider various attributes. Of these, Validity (content, criterion-related, and construct validity); Reliability; Standardization; Practicality; Fairness; Comprehensiveness; Optimal Difficulty Level; Discriminative; Scorability and finally ensures the constructive alignment (test items must follow the educational program's learning objectives and outcomes). The characters mentioned all refer to the concept of the "psychometric properties of the test." This term delineates the qualities and practicality of a high-quality exam in assessing psychological attributes such as traits, abilities, and knowledge.[2,3]

Of the aforementioned characteristics, validity is the first key attribute of a good test. Validity is frequently determined by comparing test scores to specific criteria, such as the same behavior, personal achievement, or characteristic

that reflects the property that the test was intended to measure. Only valid tests can provide relevant information about individuals.[3]

Construct underrepresentation and concept-irrelevant variation are the 2 primary sources of validity threats when examining major threats to validity measures.[4,5] Construct underrepresentation is the undersampling or biased evaluation of curriculum or course content. Construct-irrelevant variation, on the other hand, results from improper item formats, too difficult or too easy questions, or inaccurate test modalities.[2]

A test blueprint is a tool that can be used to overcome these major validity threats. A blueprint is an extension of the learning outcomes and course objectives that most educators already possess. Blueprint allows for aligning various skills with the course material and the most appropriate mode of assessment.[6] In addition, a well-designed test blueprint is an integral part of what we refer to as the constructive alignment of a course. There are 3 pillars of education: structured learning objectives, teaching and learning activities, and assessment tasks. These 3 pillars of education can be aligned with the help of a blueprint.[7,8]

Despite the apparent benefits of test blueprint in the process of evaluation, particularly its role in reducing major test validity and reliability concerns, ensuring a proper curriculum design, and creating fair tests with an acceptable degree of score variance, the proper operationalization of test blueprint in medical education still need to be explored further. This includes concerns like the method of blueprint design is not uniform in many medical schools; awareness of test developers regarding the importance and effectiveness of test blueprints and how to adopt them properly is not complete; a fit-for-purpose blueprint with keeping the essential steps of its design to ensure its quality; methods of selecting the appropriate assessment format in alignment with the learning outcomes through the blueprint; and the terms of electronic blueprinting in adaptive testing.[9–11]

We conducted this scoping review to examine the test blueprint approach as described in the pertinent literature. Our objective was to emphasize the benefits of this approach in enhancing test measures, with a specific focus on validity and reliability concerns. The study types, data collection tools, major challenges faced during test blueprint design, and recommendations for a suitable guide for test blueprint design were determined. Potential gaps and paradigms were also identified, and recommendations for domains where further research can be conducted were highlighted.

Methods

The present study employed the framework proposed by Arksey and O'Malley (2005), which is characterized by its descriptive nature and aligns with the review's objective of adopting a descriptive approach.[12]

Stage 1: Identify the research question

As the initial step of this review, the research question was clearly identified. Our primary question was, "What are the most significant threats to validity in assessment, and what is the potential role of test blueprint in reducing these threats?"

Stage 2: Identify the relevant literature

1. The research question guided the identification of the relevant studies. The authors, with the help of a librarian, began an iterative search of the pertinent databases. The keywords and Boolean operators were constructed with care and revised. Consistent with the all-encompassing nature of the research, our search strategy was designed to include the most relevant publications that fall within the scope of the research question. The search terms "Blueprinting," "Validity and Reliability, "Educational measurement(s)," "Blueprinting in Assessment(s)," and "Psychometrics" were used to conduct research in various databases (Appendix A). In conducting research, helpful information sources such as electronic databases (MEDLINE/PubMed), article reference lists, and Scopus (Elsevier) were utilized. Only English-language articles were included. Studies employing qualitative, quantitative, and blended research methodologies were included. Due to cost, time, the necessity for language translation services, and the lack of proficiency in a language other than English, studies written in languages other than English were excluded from the review. Furthermore, our scoping review prioritizes thoroughness, and including non-English studies might be perceived as a potential compromise to quality due to translation issues. Commentaries and conference abstracts were excluded as well. The online management application Covidence (Veritas Health Information, Melbourne, Australia) was used to store and retrieve pertinent literature. This makes data review and extraction simple and efficient.

Stage 3: Select appropriate studies

The inclusion and exclusion of studies were assessed by 2 members (HA and MRB) following predetermined criteria. The process of selecting studies was iterative and challenging, with a focus on identifying articles that aligned with the study's objectives and research questions. Each member revised the designated articles per the criteria for inclusion and exclusion. The disagreement over article selection was discussed and resolved. The measure of inter-rater reliability and agreement was employed. The inter-rater reliability and level of agreement were determined using Cohen's kappa value. The obtained kappa value of 0.83 was deemed as an acceptable threshold for the level of concurrence, as per the criteria established by Cohen (1960).[13]

Stage 4: Data extracting, mapping, and charting

A charting framework was developed after synthesizing and interpreting qualitative data according to themes. Prior to data extraction, the variables of interest (blueprinting in assessment, reliability, validity types, threats) were defined and linked to the research question. Using Covidence and Microsoft Excel, a data charting framework was developed.

Before conducting the final review, the data extraction and charting process was piloted by the authors and another subject matter expert (a member of the examination committee). Additionally, a 3-tiered process was conducted during the data charting procedure. This process included: (1) a graphical representation of the data; (2) summarization of the data using descriptive statistics (e.g., mean, median, and range); and (3) statistical analysis of the data utilizing the appropriate analytical tests. Each article incorporated into the review was assigned a unique identification number.

Stage 5: Summarizing and reporting the results

Data were extracted from Covidence and imported into a Microsoft Excel spreadsheet (Appendix B). Quantitative and qualitative subtypes of data were distinguished. For the analysis of the quantitative data, descriptive statistics were applied. The qualitative data were analyzed employing the reflexive approach to thematic analysis, which was originally devised by Braun and Clarke.[14] A hybrid approach was employed to carry out thematic analysis, incorporating both deductive and inductive coding.

Stage 6: Expert consultation

With a solid background in test blueprint development and evaluation, the principal investigator requested the assistance of 2 experts. In light of their feedback provided, the review considered additional literature to be added.

Results

Study selection

The method for selecting studies is depicted in Fig. 1 using the PRISMA diagram. The identified records included 487 studies. Following removing duplicates, 367 were eligible for the title and abstract screening. 111 studies were subjected to an initial screening. Of these studies, 40 were eliminated because they failed to address the major test validity and reliability threats or did not explicitly address the test blueprint design. Then, the eligibility of 54 studies was determined. Two raters evaluated the records included in this phase, with a 96% inter-rater agreement rate. Finally, 22 (n=22) full-text records were submitted for grid extraction and analysis.

Study population and design

A summary of the included studies is provided in Appendix B. The 22 included studies provided details on blueprinting methods and their role in mitigating the most important validity threats. 18 of the studies focused on blueprinting design and implementation as their primary objective, while 4 focused on blueprinting as a tool to enhance the design of the medical curriculum and the practice of assessment as a whole. 4 studies highlighted the role of blueprints in mitigating the most significant threats to the validity and reliability of assessments. 16 studies included medical programs (undergraduate, residency, or progress testing), 2 studies included nursing schools, 1 study was conducted among English teachers in undergraduate schools, 1 study was conducted in a college of dentistry, and 2 studies provided a practical guide and tips for developing a test blueprint.

One study was qualitative, 2 were cohort studies, 8 were cross-sectional, 6 were quasi-experimental, 3 were review articles, one of which was a practical guide for developing a test blueprint, and 1 was a case report.

Tools for data collection

Eight of the 22 studies selected for the review relied primarily on surveys to evaluate the responses of participants to question papers designed in accordance with a test blueprint. There was no clear evidence of the survey's validity in these studies, and many of them employed instructor-created surveys. In 6 studies, blueprinting-dependent post-test psychometric analysis was used to compare test results with and without blueprinting. One study utilized the item response theory, incorporating both item difficulty and student ability into its analysis, utilizing the Rasch model for one parameter analysis. While others (n=5) employed the classical test theory (assessing the test results with reference to the overall test measures and not the response to the individual test items) in the blueprinting post-intervention psychometric analysis. A number of evaluators who evaluated an intervention (Electronic blueprint design) were used in 1 study to create a 3-dimensional content analysis using the multidimensional scaling tool. One study evaluated the quality of the developed test blueprint using a matrix to determine the degree of congruence between the test and its blueprint. As their primary purpose was to provide a practical guide for test blueprinting, the data collection methods of 4 studies were not identified.

Characters of a good test

Most studies included in this scoping review (n=17) focused on gathering qualitative data on how to design a test blueprint; 3 of these studies analyzed the perspectives of test participants on the design and dissemination of the test blueprint among students. In a number of the studies incorporated into the review, the attributes of "psychometrically sound assessments" have been elaborated. When reporting a test, the attainment of score variance, validity, and reliability are considered to be of the uttermost importance.[15] In the current review, this is delineated as the primary objective of 5 studies. Of these, a study by Kibble (2017) described in depth the attributes that define a good test.[16] The study presented a thorough description of van der Vleuten's notion of assessment utility, wherein he delineated it as an outcome of a number of variables, including, validity, reliability, acceptability, educational impact feasibility, and cost-effectiveness.[17] The study additionally referenced the enhancement of this framework by Norcini et al. (2010), wherein they incorporated the concepts of equivalence (i.e., the likelihood of obtaining comparable outcomes through repeated assessment cycles) and catalytic effect (in which assessment feedback stimulates subsequent learning).[18] Another study detailed a

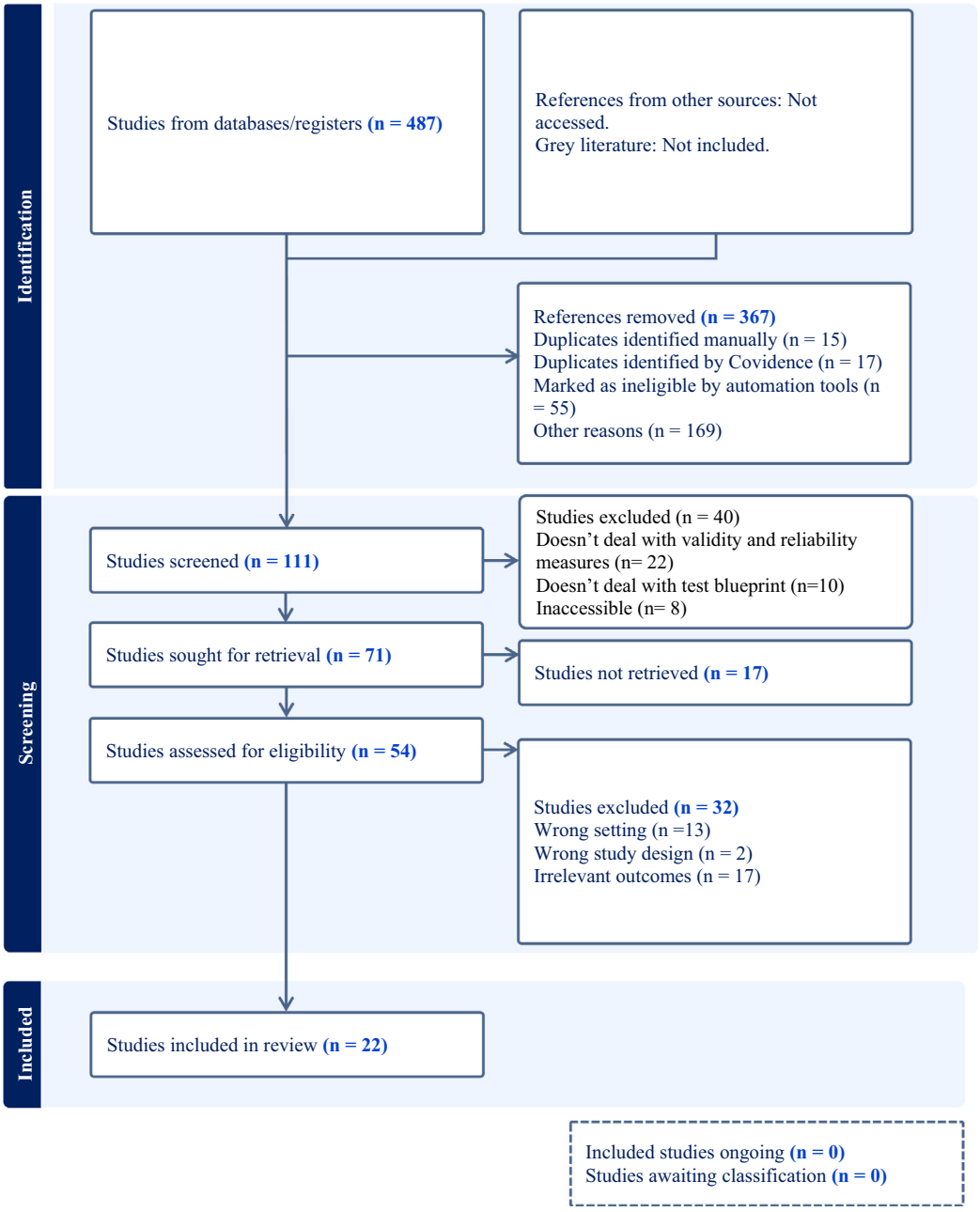

Fig. 1 Flowchart for a review of exam blueprinting as a tool for overcoming principal threats to test validity.*
***** The diagram was created with Covidence (Veritas Health Information, Melbourne, Australia) systematic review software.

qualitative, standardized MCQ construction approach and suggested ways to improve high-stakes exams. Additionally, one of the studies found that a test blueprint improves reliability, item, and person separation indices, and test unidimensionality, and reduces measurement errors to enhance test score variance.

Major validity and reliability threats for test design

The primary concerns on validity and reliability in test design were explained in the majority of the studies (*n* = 17) incorporated in this review. Kibble's (2017) study detailed the main validity types and their subtypes with examples.[16] The study examined how validity expands beyond the test to include scores created at the time, in a given context, by a specific group and how they were used to make decisions. This study, along with 2 other studies (*n*=3), extensively examined the most substantial threats to construct validity. Moreover, 8 studies provided comprehensive discussions on the concept of content validity. 2 of these examined the role of the test blueprint in attaining content validity as its primary objective and indicated that adherence to the test blueprint, learning outcomes, and instructional methods enhances the content validity of a test.

One study included in the review explained the shift towards a unified construct validity concept, which

incorporates what was previously regarded as distinct categories of validity, including content validity and criterion validity. A detailed explanation of "Kane's argument approach to validation" was provided in the study, which examined the transition from multiple to a single validity type. The method introduces a unitary concept of construct validity predicated on the plausibility and coherence of the interpretive argument. The study investigated the impact that Kane's argument framework on validation has had on the domain of educational assessment.[19]

Advantages of having a test blueprint

Blueprinting aims to reduce validity concerns (underrepresentation and irrelevant variation). This assists educational institutions in determining which assessment techniques are relevant to the framework and subject matter of the curriculum. In 18 studies included in the review, this was explicitly explained. The other 4 studies did not directly address the advantages of the test blueprint. They were mainly directed to assess the perspectives of participants towards the process of blueprinting, the impact of its dissemination among participants (teachers and students), and how to construct a valid assessment tool and ensure its educational effectiveness. In 16 of these 18 studies, the role of a blueprint to support the content validation of assessment (ensuring that scores on a specific test generalize to a larger domain of interest) was described in detail.

Of these, 6 studies described additional advantages of the test blueprint as follows: (1) it provides a framework for evaluating the response process validity; (2) the alignment of content categories and competency domains in the blueprint serves as a basis for providing students with feedback; (3) it plays a crucial role in organizing departmental item writing; (4) it provides metadata for managing test materials and contributes to the development of educational quality; and (5) it reduces the bias and affinity of the paper's setter for certain topics.

Two studies specifically demonstrated the need to release the test blueprint to increase fairness and mass acceptance of the assessment plan among teachers and students, which improved students' view of the evaluation process's validity.

Another study showed that blueprinting improved topic distribution and ensured synchronization between question formats on a single pathology exam. Post-examination item analysis was examined in 2 studies. They confirmed that good blueprinting improves test psychometric measures. One of them used classical test theory (CTT) and overall test measures. In which, blueprinting improved test reliability (defined by Cronbach's alpha), item difficulty, and item discrimination indices. However, the other one adopted item response theory (IRT) in test psychometric analysis. The test blueprint improved person separation estimates, ensured broader test score variance, local item independence, and test unidimensionality, and significantly reduced test measurement errors.

Methods for designing a test blueprint

In 21 of the 22 studies included in this review, the methods and processes for developing a test blueprint were described. One study, conducted by Banerjee et al. (2019), focused on the educational potential of social media platforms among College of Medicine students as a means to improve learning outcomes and the assessment format, rather than exploring the methods of blueprinting.[20] The majority of these studies characterized a test blueprint design as a matrix that maintains a balance between computational complexity and purpose-specificity. In almost all studies, it was also observed that there is no specific pattern for designing and implementing a test blueprint and that it is constructed according to a pattern approach that is tailored to the specific needs of the process. Of those 21 studies, 4 (n=4) postgraduate courses adopted the blueprinting process, while 15 undergraduate courses adopted the blueprint design as part of their content validity assurance method. The remaining 2 studies presented a practical guide and general tips for blueprinting that can be applied to numerous assessment plans.

There was an agreement that the process of blueprinting comprises the following main stages: (1) define the purpose and scope of assessment properly; (2) determine the primary domains of knowledge and skills to be evaluated; (3) outline the objectives or learning outcomes to be assessed for each domain in each topic; (4) determine the format for assessment; (5) specify the weight to be assigned to each content category (knowledge, skills domain, etc.).

In a number of studies, the steps involved were described in greater detail, but it was observed that the above-mentioned stages were consistent across all studies that described the methods of test blueprint design. The study by Raymond and Grande (2019) described this process as a 2-dimensional matrix: (1) content-oriented (explain the test by themes or subject matter taught).[21] (2) Process-oriented (describe items in terms of procedural skills students have to acquire, suited for clinical training that emphasizes procedural skills and affective domain). (3) Content by process matrix (designed test items using both methods and better for modern integrated curriculum). This study also described how to determine category weights by incorporating both the impact (relative importance of the outcome or topic in the field) and the frequency (how frequently it is used in clinical practice) measures.

In a study that was primarily focused on establishing an examination for one of the evidence-based medicine (EBM) courses, the process of designing a task analysis and item specification forms were clearly outlined as routine steps for test blueprint design.[22]

Three studies adopted an electronic blueprinting methodology, one of which utilized the ExamSoft question labeling system to map exams to course objectives in a Doctor of Dental Surgery program at a US-accredited dental school.

The revised Bloom's taxonomy [recognize and recall (level 1), understand and interpret (level 2), and apply, analyze, and evaluate (level 3)] and Ward's taxonomy (recall, application, and problem-solving) were used in most studies to categorize learning domain questions.[23,24]

The method of calculating the total number of test items in the question paper was a concern when designing a test blueprint. Among the included studies, it was observed that this attribute varied. 4 of the 21 studies utilized the overall exam duration and the time allotted to each examinee to

respond to each test item. This will determine the number of test items in relation to the exam length, and according to the studies, this will increase the overall test reliability and the expected discrimination index. 2 additional studies employed the credit hours system to calculate the total number of test items. Each course credit hour was designated a predetermined number of questions (e.g., 30 questions/credit) was assigned. The specifics of this approach are described in one of the studies included in this review.[25] The remaining studies determined the overall number of items on the test with a predetermined number that produces an acceptable level of reliability and a good degree of score variance among test-takers based on experience and data retrieved from previous comparable exams.

Multiple choice questions (MCQs), short answer questions (SAQs), and long essay questions (LEQs) were used as question formats in the majority of the studies included in this review, with MCQs being more frequently used in the assessment. In one study, Miller's pyramid framework was used to measure clinical competence utilizing written examinations and well-constructed MCQs to assess cognition and knowledge domains (knows and knows how). At the 'shows' level, OSCEs with simulated patients or model procedures were used. DOCEE (direct observation clinical encounter examination) and Mini-CEX (mini-clinical evaluation exercise) were suggested for the 'does' level.[26]

In one study by Eweda et al. (2020), quality assurance of test blueprinting was evaluated. The primary objective of the study was to determine the extent to which subject matter experts perceived alignment between the developed test items and the predesigned test blueprint. As a percentage of the total number of test items, the degree of alignment was expressed as the proportion of items that exhibited congruence with the predesigned test blueprint.[27]

Discussion

The purpose of this scoping review was to establish a repository of evidence and evaluate the extent of the literature on blueprinting, its design, validation, and implementation in the assessment process.

It provides evidence demonstrating the significance of the test blueprint in attaining an acceptable level of assessment utility. As mentioned by van der Vleuten (1996), the utility index of an assessment depends on its validity (coherent evidence that assessment is used for the stated purpose), reliability (reproducibility of test results), feasibility, and cost-effectiveness (fitting of the test within a specific context), acceptance by the stakeholders, and whether or not the assessment has an educational impact of improving the learning outcomes.[17] Norcini et al. (2010) extended this concept to include equivalence (test results are reproducible if applied in other learning institutions) and the catalytic effect of assessment that guides future learning.[18] These characteristics of assessment utility were discussed to varying degrees in a number of the studies included in this review. Exploring the significance of blueprinting in attaining an accepted level of assessment utility across all subcategories is strongly encouraged in future studies. The broad adoption of faculty development, accreditation and evaluation programs, and initiatives aimed at personal and institutional improvement has led to an intensified recognition of the significance of test blueprint as a crucial component of the assessment process. Therefore, a test blueprint is expected to be regularly incorporated into all courses of the curriculum, with a design that is appropriate for the intended purpose.

It was observed that a considerable number of studies (n = 19) employed a qualitative (survey-based) or quasi-experimental design. However, it is recommended that future studies utilize controlled and experimental research designs, along with validated analytical tools, to obtain more objective and validated structured data. This would enhance the applicability of the findings across a wider range of contexts. Furthermore, the application of blueprinting in domains beyond health professions education, as well as the extended evaluation of its efficacy in enhancing learning outcomes and curriculum implementation, were not thoroughly addressed.

Existing literature provides useful insights into the process of test blueprinting and its role in enhancing the assessment process. However, within this cohort of studies that met our eligibility criteria and were included in this review, the characteristics of a good test were clearly identified in many of them, with test validity, reliability, psychometric characteristics, and educational effectiveness being the most commonly used. Validity is given the most emphasis for tests that are used for decision-making. Kibble explicitly stated the aforementioned and further delineated 5 broad categories of validity evidence, namely: content-related evidence, response process-related evidence, internal structure-related evidence, relation to other variables, and consequential evidence of assessment.[16,19]

In the majority of studies included in the review, the 2 most common categories of validity threats cited were construct with underrepresentation and construct with an irrelevant variation. This is consistent with the findings of Downing (2003) regarding the interpretation of meaningful assessment data.[28] For the construct with underrepresentation resulting from the use of irrelevant, too few items in the sample domain, trivial items with maldistribution across the learning domains, and the use of items with poor reliability, the studies in the review recommended a well-constructed test blueprint that broadly samples the learning domain of interest and possesses a rigorous peer review process for the test items to achieve a high degree of reliability. In the other context, namely constructing a test with irrelevant variation (excess errors and noise in measurements), it was mentioned that poor item quality, lack of review, items with trivial details, and those that are off target with regard to the learning outcomes are the primary causes of this type of validity error. To avoid this form of error, the peer review process is highly recommended. This is consistent with a number of other studies that have emphasized the significance of peer review of test items for reducing errors and enhancing the overall reliability of tests.[29–31] Kibble (2017) also added a crucial determinant for this form of threat (validity with irrelevant variation): instances of cheating and loss of test security.[16] In the context of the good test characteristics described in this review, it is also recommended that a relevant and good standard-setting approach for determining the pass/fail scores is another factor that

affects the test validity (type of irrelevant variation) and that the entire point of validity is to be on firm ground when inferring decisions regarding exam results and outcomes.[28]

Most of the studies (n=18) in the review discussed test blueprint advantages beyond the scope of validity. Based on participant self-reported satisfaction ratings and test scores, the results of our study show that the utilization of a test blueprint is linked with improved performance outcomes. The blueprint helps extrapolate exam scores to a wider range of relevance. Furthermore, a blueprint provides a basis for validating a test response process. Hence, having a well-designed test blueprint improved assessment-related materials, and item design for the entire review process, and reduced paper setters' affinity for some topics. This is consistent with the findings of Raymond and Grande (2019) who elaborated in great detail in their guide on each of the advantages of employing a test blueprint.[21] In which the primary advantages of the test blueprint were as follows: it ensures that the assessment achieves learning objectives, promotes optimal question distribution, increases consistency in difficulty, and contributes to fair, comprehensive examinations. It facilitates modifying of electronic question banks and identifies the content gaps.

This part of the review highlighting the advantages of a test blueprint is consistent with the findings of other studies that described these benefits. Among these studies, a study by Cantrell (2012) listed the benefits of a blueprint in a middle school science classroom as follows: a student taking the same unit test with a blueprint will experience the same cognitive load in relation to the grade level to answer the item correctly. Additionally, the uniformity in item specification enables the comparison of item responses across different tests, irrespective of the content topic, which enhances the validity of score interpretation.[32] Moreover, utilizing a blueprint facilitated the breakdown of data (differential item functioning), wherein students with distinct demographic variables exhibit varying performance levels in diverse content and cognitive domains. Cantrell (2012) illustrated this effect, wherein male and female students' reactions to distinct test items were gauged.[32] The results indicated that female students performed better than their male counterparts in questions about higher cognitive levels, despite males obtaining higher scores in the overall test.

Out of those studies that addressed the advantages of blueprinting, 3 of them reported the benefits of blueprinting in the form of enhancements in the psychometric properties of tests. 2 of which employed classical test theory to analyze post-test results, utilizing measures such as item difficulty, discrimination, and reliability indices (as expressed by Cronbach's alpha). The other one, on the other hand, utilized the one-parameter Rasch model to analyze post-exam performance, focusing on item response patterns and expressing their findings in terms of person and item separation measures, local item independence, unidimensionality, and item fitting indices.[33] Within this context, Gamilli (2018) conducted a study that examined the congruence between a scoring method and a test blueprint in relation to its content allocation. The study has demonstrated that the allocation of optimal weight to test items in the blueprint has a significant impact on both item information and scoring. This study employed both the 2-parameter logistic (2PL) IRT model, which considers item difficulty, item discrimination, and student ability in assessing item response, and the 3-parameter logistic (3PL) IRT model, which incorporates an additional factor known as pseudo-guessing.[34]

Upon perusing the literature, the studies incorporated in this review provided a valuable reference on the methods employed in developing a test blueprint. The review encompassed 21 studies that clearly explained the different approaches implemented in test blueprint design. A study by Banerjee et al. (2019) that was incorporated into the review discussed medical students' significance and attitudes toward various social media platforms as a potential educational resource that might improve the learning and evaluation process.[20] Despite deviating from the review's main goal, this study was included due to its innovative and valuable findings and detailed survey analysis (using constructs, principal component analysis of residuals, and factor analysis). This study used social media to improve student evaluation and update learning outcomes and the exam blueprint to reflect the current updates in the learning processes.

Most studies divided blueprinting into main steps. First, the assessment's purpose and target are set, then learning outcomes are aligned with the primary cognitive domains. Next, the assessment format and category weights are determined. These essential steps were followed in many studies included in the review. Studies agreed that there is no standardized test blueprint design and that a customized approach should be adopted based on the exam's objective. A majority of studies aligned test questions with learning domains using modified Bloom's taxonomy and Ward's frameworks.[23,24]

The study by Raymond and Grande (2019) detailed the process for determining the content category weight utilizing the impact (relative relevance of the outcome in the field) and frequency (how frequently used in clinical practice) approach for determination.[21]

The method that was employed in the literature reviews to ascertain the total count of test items featured in the blueprint, and subsequently, the allocation of items for each category based on its proportional weight in the curriculum, exhibited a lack of uniformity. The number of overall test items in 4 studies was determined by considering both the total duration of the exam and the allotted time for every single item to be answered. However, in 2 other studies, a method was employed that depended on the credit hours assigned to each course. The specifics of this approach have been explained in one of the studies.[25] The remaining studies employed a method that involved utilizing a predetermined number of test items to attain a satisfactory level of reliability (degree of variance in the true score when compared to the observed score). Numerous studies have examined the relationship between the number of test items and the reliability coefficient measure.[35,36]

The utilization of computerized adaptive testing (CAT) has gained progress due to advances in the evaluation process and the implementation of the Item Response Theory (IRT) in the assessment procedure, as noted by Van der Linden and Glas (2010).[37] The likelihood of obtaining precise scores on the same (z) scale with any number of test items is attributed to 2 significant characteristics of the IRT,

namely: the alignment of people and test items on the same scale and the ability to obtain accurate scores on the scale with any number of items.[38] By utilizing CATs, it is possible for each student to be presented with a distinct selection and/or number of items while still achieving an accurate score. Assuming an initial ability level of 0.00 logits, which represents average ability, tests are administered and subsequently assigned based on the examinee's response to the first item of the test. The examination is conducted until a termination criterion is attained, which may be a pre-established measurement accuracy level or a predetermined measurement standard error. The restricted utilization of this testing methodology can be attributed to the requisite resources and insufficient experience with implementing and analyzing IRT among test developers. The review revealed a lack of implementation of this testing method in the studies examined. Electronic blueprinting and linking objectives to test items through a software-based question labeling system could be beneficial in this context. Further investigation into the development and application of a blueprint applied in this type of testing could prove to be a valuable area of exploration for future research.

The literature reviewed in this study pertained to the design of blueprints in conventional assessment methods. However, it is worth noting that in numerous medical education and health professional organizations, there is a shift towards utilizing assessment for learning and programmatic assessment (PA) strategies. PA considers assessment as interconnected with the learning experience and contributes to *in vivo* educational practice, broadening the context of learning.[39] The reviews included in our study don't explain PA's fit-for-purpose blueprint design, which is needed for domain-specific and domain-independent skills evaluation adopted in PA.

When reporting this study, some limitations need to be addressed. First, most of our review's studies were qualitative or quasi-experimental, and controlled experimental studies on the test blueprint's significance are needed. Second, only English-language studies were considered. This aspect could potentially introduce "English-language bias" and compromise the general relevance of our work. Third, our study only examined PubMed and Scopus, not additional databases, or gray literature. Research publication bias is inevitable because poor studies with inadequate designs or unrepresentative samples may be included. The literature search found no systematic reviews or meta-analyses. These studies could have greatly improved the review, but they were not found in the searched databases, which may indicate the novelty of our work in this important field of research.

Further studies that validate test blueprints across all recognized domains of validity, not just content and construct, should be considered for the prospective field of test blueprinting research. Such studies should demonstrate the significance of test blueprints and employ a variety of validation methods. By determining the degree of congruence between test items and the predetermined test blueprint, quality assurance should be incorporated into the test blueprinting process. Test blueprint applications should be examined comprehensively in examinations that utilize the IRT framework. Finally, additional research is required regarding programmatic assessment strategies and computerized adaptive testing that employ test blueprints.

Conclusions

In this scoping review, we analyzed the test blueprinting process and discussed its significance in enhancing test measures. Pertinent studies were selected for comprehensive analysis, and their significant findings were carefully summarized. Along with discussing the primary validity threats in test design and the role of blueprinting in overcoming them, we additionally presented the studies' main findings regarding the recognition of how a test blueprint is constructed and applied in various learning curricula and what are the primary attributes of a successful test.

In this context, we raised a number of recommendations, like providing a systematic approach to blueprint design, ensuring its quality, and emphasizing the need to disseminate the blueprint among participants. In addition, we found that blueprinting for computerized adaptive testing, programmatic assessment, electronic, and a fit-to-purpose blueprint was identified to be needed.

Finally, it is worth noting that the test blueprint is used in conjunction with other context-specific measures to enhance test quality. Of these, post-examination reports, psychometric criteria, assessment utility measures, and expert feedback are important. To address the importance of this domain in assessment and curriculum development, further studies employing variable methodologies and thorough data-collection and analysis tools are recommended.

Orcid

Hussein Abdellatif: https://orcid.org/my-orcid?orcid=0000-0001-5590-5112

Authors' contributions

Conceptualization, writing, review, and editing: HA; Review of the selected literature and data charting (HA, AE, TK, and MRB); Revision and format of the final review (HA and MRB).

Ethics approval

According to the standards of the Institutional Research Ethics Committee at Sultan Qaboos University (SQU), College of Medicine and Health Sciences, review articles do not require ethical approval.

Funding resource

None.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the production of this work, the author (PI) utilized [Wordtune] to a very limited extent to enhance the writing process. After utilizing this tool, the author(s) reviewed and edited the content as necessary and took complete responsibility for the publication's content.

Declaration of competing interest

The authors declare that there are no conflicts of interest associated with this study.

Acknowledgments

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.edumed.2024.100906.

References

1. Meiklejohn S, Anderson A, Brock T, Kumar A, Maddock B, Wright C, Walker L, Kent F. The utility of an interprofessional education framework and its impacts upon perceived readiness of graduates for collaborative practice? A multimethod evaluation using the context, input, process, product (CIPP) model. Nurse Educ Today. 2023 Jan 5:105707.
2. Ahmad RG, Hamed OA. Impact of adopting a newly developed blueprinting method and relating it to item analysis on students' performance. Med Teach. 2014;1;36(sup1):S55–61.
3. Reeves TD, Marbach-Ad G. Contemporary test validity in theory and practice: a primer for discipline-based education researchers. CBE Life Sci Educ. 2016 Spring;15(1):rm1. https://doi.org/10.1187/cbe.15-08-0183. PMID: 26903498; PMCID: PMC4803101.
4. Crocker L, Algina J. Introduction to classical and modern test theory. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887. 1986;32.
5. Burns GN, Morris MB, Periard DA, LaHuis D, Flannery NM, Carretta TR, Roebke M. Criterion-related validity of a Big Five general factor of personality from the TIPI to the IPIP. Int J Sel Assess. 2017;25(3):213–22. https://doi.org/10.1111/ijsa.12174.
6. Denga DI. Educational Measurement, Continuous Assessment and Psychological Testing.Calabar Rapid Educational Publishers Ltd. 1987.
7. Allen MJ, Yen W. Introduction to Measurement Theory. Monterey, CA: Brooks/Cole Publishing Company M; 1979;37.
8. Mohajan H. Two criteria for good measurements in research: validity and reliability. ASHUES. 2017;29(4). https://ssrn.com/abstract=3152355.
9. Michael G, John B, Megan L, Eric F, Mary S, Wagner J. Educator's blueprint: A how-to guide for developing high-quality multiple-choice questions. AEM Educ Train. 2023. https://doi.org/10.1002/aet2.10836.
10. Cesare A. The future of standardised assessment: validity and trust in algorithms for assessment and scoring. Eur J Educ. 2023. https://doi.org/10.1111/ejed.12542.
11. Valentine Joseph O, Delight O, Idika B, Asuquo B. Exploring the potential of artificial intelligence tools in educational measurement and assessment. Eurasia J Math Sci Technol Educ. 2023;19(8):em2307. https://doi.org/10.29333/ejmste/13428.
12. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Social Res Methodol Theory Pract. 2005;8(1):19–32. https://doi.org/10.1080/1364557032000119616.
13. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960 Apr;20(1):37–46.
14. Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qual Res Sport Exerc Health. 2019 Aug 8;11(4):589–97.
15. Zorowitz S, Niv Y. Improving the reliability of cognitive task measures: A narrative review. Biol Psychiatry Cogn Neurosci Neuroimaging. 2023;8:789–97.
16. Kibble JD. Best practices in summative assessment. Adv Physiol Educ. 2017 Mar 1;41(1):110–9.
17. Van Der Vleuten CP. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ Theory Pract. 1996;1:41–67. https://doi.org/10.1007/BF00596229.
18. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, Roberts T. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33:206–14. https://doi.org/10.3109/0142159X.2011.551559.
19. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ. 2015;49(6):560–75.
20. Banerjee Y, Tambi R, Gholami M, Alsheikh-Ali A, Bayoumi R, Lansberg P. Augmenting flexnerism via twitterism: need for integrating social media application in blueprinting pedagogical strategies for undergraduate medical education. JMIR Med Educ. 2019 Mar 25;5(1):e12403.
21. Raymond MR, Grande JP. A practical guide to test blueprinting. Med Teach. 2019 Aug 3;41(8):854–61.
22. Bridge PD, Musial J, Frank R, Roe T, Sawilowsky S. Measurement practices: methods for developing content-valid student examinations. Med Teach. 2003 Jan 1;25(4):414–21.
23. Bloom BS, Krathwohl DR. Taxonomy of Educational Objectives. Longmans. Green & Co. 1956.
24. Ward Educational Consulting. Handbook for Test Development. Florida: Ward Educational Consulting Inc; 1983.
25. Abdellatif H, Al-Shahrani AM. Effect of blueprinting methods on test difficulty, discrimination, and reliability indices: cross-sectional study in an integrated learning program. Adv Med Educ Pract. 2019 Jan 22:23–30.
26. Hamdy H. Blueprinting for the assessment of health care professionals. Clin Teach. 2006 Sep;3(3):175–9.
27. Eweda G, Bukhary ZA, Hamed O. Quality assurance of test blueprinting. J Prof Nurs. 2020 May 1;36(3):166–70.
28. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003;37:830–7. https://doi.org/10.1046/j.1365-2923.2003.01594.x.
29. Malau-Aduli BS, Zimitat C. Peer review improves the quality of MCQ examinations. Assess Eval High Educ. 2011;37:919–31. https://doi.org/10.1080/02602938.2011.586991.
30. Naeem N, van der Vleuten C, Alfaris EA. Faculty development on item writing substantially improves item quality. Adv Health Sci Educ Theory Pract. 2012;17:369–76. https://doi.org/10.1007/s10459-011-9315-2.
31. Wallach PM, Crespo LM, Holtzman KZ, Galbraith RM, Swanson DB. Use of a committee review process to improve the quality of course examinations. Adv Health Sci Educ Theory Pract. 2006;11:61–8. https://doi.org/10.1007/s10459-004-7515-8.
32. Cantrell P. Using test blueprints to measure student learning in middle school science classrooms. Researcher. 2012;24(1):55–71.
33. Abdellatif H. Test results with and without blueprinting: psychometric analysis using the Rasch model. Educ Méd. 2023 May 1;24(3):100802.
34. Camilli G. IRT scoring and test blueprint fidelity. Appl Psychol Measur. 2018 Jul;42(5):393–400.
35. Morera OF, Stokes SM. Coefficient α as a measure of test score reliability: review of 3 popular misconceptions. Am J Public Health. 2016 Mar;106(3):458–61. https://doi.org/10.2105/AJPH.2015.302993 PMID: 26885962; PMCID: PMC4816140.
36. Shieh G. Choosing the best index for the average score intraclass correlation coefficient. Behav Res Methods. 2016 Sep;48(3):

994–1003. https://doi.org/10.3758/s13428-015-0623-y. PMID: 26182855.
37. Van der Linden WJ, Glas CA, editors. Elements of Adaptive Testing. New York: Springer; 2010 Mar 10.
38. Ryan J, Brockmann F. A Practitioner's Introduction to Equating with Primers on Classical Test Theory and Item Response Theory.Council of Chief State School Officers; 2009 Jun.
39. Van Der Vleuten CP, Schuwirth LW, Driessen EW, Govaerts MJ, Heeneman S. Twelve tips for programmatic assessment. Med Teach. 2015 Jul 3;37(7):641–52.