Operationalizing a Weighted Performance Scoring Model for Sustainable e-Learning in Medical Education: Insights from Expert Judgement

Deborah Oluwadele[1,2], Yashik Singh[1] and Timothy Adeliyi[2]
[1]Department of Telemedicine, School of Nursing and Public Health, University of KwaZulu-Natal, Durban, South Africa
[2]Department of Informatics, University of Pretoria, South Africa

deborah.oluwadele@up.ac.za (corresponding author)
singhy@ukzn.ac.za
timothy.adeliyi@up.ac.za

https://doi.org/10.34190/ejel.22.8.3427

Abstract: Validation is needed for any newly developed model or framework because it requires several real-life applications. The investment made into e-learning in medical education is daunting, as is the expectation for a positive return on investment. The medical education domain requires data-wise implementation of e-learning as the debate continues about the fitness of e-learning in medical education. The domain seldom employs frameworks or models to evaluate students' performance in e-learning contexts. However, when utilized, the Kirkpatrick evaluation model is a common choice. This model has faced significant criticism for its failure to incorporate constructs that assess technology and its influence on learning. This paper aims to assess the efficiency of a model developed to determine the effectiveness of e-learning in medical education, specifically targeting student performance. The model was validated through Delphi-based Expert Judgement Techniques (EJT), and Cronbach's alpha was used to determine the reliability of the proposed model. Simple Correspondence Analysis (SCA) was used to measure if stability is reached among experts. Fourteen experts, professors, senior lecturers, and researchers with an average of 12 years of experience in designing and evaluating students' performance in e-learning in medical education participated in the evaluation of the model based on two rounds of questionnaires developed to operationalize the constructs of the model. During the first round, the model had 64 % agreement from all experts; however, 100% agreement was achieved after the second round, with all statements achieving an average of 52% strong agreement and 48% agreement from all 14 experts; the evaluation dimension had the most substantial agreements, next to the design dimension. The results suggest that the model is valid and may be applied as Key Performance Metrics when designing and evaluating e-learning courses in medical education.

Keywords: E-learning evaluation model, Medical education, Content validation, Performance optimization, Expert judgment technique

1. Introduction

E-learning is a comprehensive concept encompassing the asynchronous or synchronous dissemination of knowledge to learners via electronic systems (Güllü, Kara and Akgün, 2024). Recently, e-learning has garnered significant recognition as a mainstream approach in health sciences education (HSE), encompassing medical, dental, public health, nursing, and other allied healthcare disciplines. However, there remains considerable debate surrounding the role of e-learning and its impact on learners' performance and learning enhancement (Regmi and Jones, 2020). In medical education, e-learning is a strategic tool for human resource development in health to combat the burden of diseases and ensure the achievement of the United Nations' Sustainable Development Goals (Oluwadele, Singh and Adeliyi, 2023a). Healthcare professionals play a crucial role in offering essential and dependable assistance to individuals with HIV/AIDS and tuberculosis, and e-learning is appraised as a potent strategy to address the challenges of HIV/AIDS and tuberculosis in sub-Saharan African nations (Ajenifuja and Adeliyi, 2022). Despite these potentials, the debate about the suitability of e-learning in medical education is still heated. Scholars argue that although e-learning is being accepted and used in medical education, it might have been used because of its popularity and novelty rather than for pedagogical evidence, and it is difficult to assess its success (Cook, 2007). Khasawneh et al. (2016) assert that e-learning is not a universally effective teaching tool because it requires careful evaluation when incorporated into established curricula and the dire need to assess the impact of e-learning on a case-by-case basis. The recurring, context-specific global difficulties in medical education and the health profession underscore the lack of agreement on the suitability of e-learning in medical education. These difficulties have been defined as wicked problems that

Cite this article: Oluwadele, D., Singh, Y. and Adeliyi, T. 2024. "Operationalizing a Weighted Performance Scoring Model for Sustainable e-Learning in Medical Education: Insights from Expert Judgement", *Electronic Journal of e-Learning*, 22(8), pp 24-40, https://doi.org/10.34190/ejel.22.8.3427

defy recognized solutions and are perceived differently by different people (Mennin, 2021). Mennin articulated the wicked problems as the effect of the pandemic on the quality of e-learning, the persistent quality concerns in clinical teaching, the enduring resistance to change among faculty members, the intricacies of collaboration, persistent constraints related to time and resources, biases across various dimensions, the intricate balancing act of fulfilling multiple professional and personal roles, the imperative for curriculum integration, pervasive conflicts within health professions institutions, and the perpetual challenge of faculty development. Solving these problems requires pursuing inquiry, pattern recognition, and adaptive action.

Although the COVID-19 pandemic led to the wide adoption of e-learning in medical education, it also created more problems. Despite the ideal approach of developing a well-devised plan, starting with a needs assessment to determine system requirements and usage (Khasawneh et al., 2016), many universities have hastily implemented "emergency e-learning" protocols in response to the pandemic, transitioning from traditional face-to-face learning to e-learning (Murphy, 2020). This unprepared shift has compelled students accustomed to conventional learning into e-learning, a situation referred to as an "imperfect yet quick solution to the crises" with potential repercussions on student performance (Nagar, 2020). The repercussion of this is that although publications on the implementation of e-learning in medical education have sporadically increased since the onset of the pandemic (Oluwadele, Singh and Adeliyi, 2023b), there is a lack of standardization in the evaluation of e-learning in medical education. Performance evaluation is conducted haphazardly, and authors do not focus on metrics that evaluate the technology components of e-learning – a phenomenon termed "evaluating e-learning minus the e" (Oluwadele, Singh and Adeliyi, 2023b), meaning that the learning component of e-learning is being evaluated with little emphasis on the electronic component. Researchers rarely used a performance evaluation framework to guide their evaluation process, and a few used the Kirkpatrick evaluation model, a widely criticized model for its inflexibility in evaluating technology-enabled learning. E-learning evaluation frameworks provide researchers with specific concepts to measure and structures to assess how well e-learning has been implemented. They assist in directing the implementation process and aid in the identification of potential facilitators and inhibitors that need to be addressed.

E-learning implementations are labor- and resource-intensive (Farhan, Talib and Mohammed, 2019); hence, there is a need to rigorously evaluate e-learning implementation to establish its importance, quality, and acceptance, as this gives investors confidence for continued investment and expansion of e-learning implementation (Raspopovic et al., 2014). One of the challenges of e-learning evaluation is the choice of performance criteria to evaluate and ascertain its impact. A solution to this challenge is to develop a systematic e-learning assessment model from the study of existing e-learning models and systems to improve the performance and use of e-learning techniques (Farhan, Talib and Mohammed, 2019, Raspopovic et al., 2014). Various dimensions of e-learning require evaluation; these include educational, personal, institutional, cultural, and technical dimensions, to mention a few. Although all these dimensions are crucial for effective e-learning implementation, this study focuses on the evaluation of e-learning in medical education from a student performance perspective.

This study aims to assess the efficiency of a model developed to determine the effectiveness of e-learning in medical education, specifically targeting student performance using expert judgment techniques. The first section of this study introduces the study by problematizing it based on extant literature. This is followed by Section 2, which delineates the materials and methodology, detailing the appropriateness of the Delphi technique design of our expert judgment method. Subsequently, section 3 encapsulates the outcomes of the conducted process, and ultimately, Section 4 formulates the primary conclusions of the paper, along with recommendations for future research directions.

2. Materials and Methods

This section presents the research questions that informed the research approach for this study and why and how the Delphi method was used to achieve consensus. This study seeks to answer the main research question: "How do experts perceive the efficacy of the weighted performance scoring model for designing and evaluating students' performance in e-learning contexts in medical education? The research questions proposed to drive the study are shown in Table 1:

Table 1: The research questions that drive the study

	Research Question	Purpose
RQ1	To what extent does the model guarantee the design of e-learning courses for optimal student performance in medical education?	This research question explores the model's reliability for designing e-learning courses that students find usable, useful, fit for context, learnable, and valuable.
RQ2	To what extent does the model ensure the evaluation of e-learning courses specifically focusing on students' performance in e-learning in medical education?	This research question explores the model's reliability for evaluating e-learning courses with a specific focus on how satisfied students felt with the course if they acquired knowledge and could apply it to solve problems in real-life scenarios.
RQ3	Does the model provide a coherent approach to designing e-learning courses?	This question presents the gap identified by experts that need to be explored further in research.

2.1 Overview of the Delphi Technique

Expert judgment techniques involve enabling a group of experts within a specific knowledge domain to provide their opinions on a particular subject collectively (Jagatheesan, 2022). Utilizing expert judgment as an assessment approach presents numerous benefits, including the superior quality of judgments provided by experts and the potential to acquire comprehensive information on the subject matter (Almenara and Cejudo, 2013). This process is crucial, as its accurate execution is sometimes the sole indicator of the content validity of a research instrument or consensus (Escobar-Pérez and Cuervo-Martínez, 2008). Various expert judgment techniques are applicable for forecasting, evaluation, or policy design. Some widely recognized techniques include brainstorming, the Nominal Group Technique (NGT), the Delphi method, and didactic interaction.

When brainstorming, the group of experts uses the brainstorming technique to share concepts related to a problem, aiming to diverge from the confines of formal problem-solving sessions to generate many ideas, including groundbreaking ones. Additionally, there is an emphasis on fostering the combination or merging of the suggested ideas (Alonso, 2015). Nominal Group Technique (NGT) involves assembling an expert group to identify problem components, propose potential solutions, and establish priorities. In the NGT approach, experts convene in person and, under the guidance of a facilitator and through a structured process, engage in discussions, voting, and ranking of the elements related to the analyzed problem (Carney, McIntosh and Worth, 1996). The Delphi method involves a panel of experts discussing a chosen topic to achieve consensus. However, in this application, consensus is sought through questionnaires and successive rounds, deliberately preventing experts from mutual identification to mitigate potential group pressures that may arise from direct discussions (Von Der Gracht, 2012). Didactic interaction is typically applied in situations requiring "yes/no" decisions. It involves in-person meetings where the group is divided into two subgroups, each expressing divergent views and engaging in discussions. Subsequently, they switch roles, each group advocating for the opposing position. This exchange facilitates an understanding of each other's perspectives and aids the group in reaching a consensus (Jagatheesan, 2022).

This study aims to conduct content validation of a model that evaluates the performance of e-learning in medical education through expert judgment. The brainstorming technique was unsuitable for the study given that the aim was to seek concession on the quality of the model developed to evaluate performance in e-learning in medical education and not to generate new ideas for a problem. The NTG technique requires experts to convey in person, which was not feasible given the geographic distribution of the experts. The didactic technique was also unsuitable as a yes, or no answer was not required to validate the model; instead, in-depth discussion and triangulation of perspective are necessary to uncover grey areas and propose areas for further improvement. Hence, the Delphi technique is the most suitable for achieving the study's objectives.

The Delphi technique is excellent for eliciting and combining expert judgment. When employing the Delphi technique, the facilitator manages the flow of information among unidentified panelists through multiple iterations, culminating in the average of estimates from the final round serving as the group's collective judgment. This approach is suitable in situations where the application of statistical methods is not fitting, a substantial number of experts are accessible, and the alternatives involve either averaging the forecasts of multiple individuals or employing a conventional group method (Rowe and Wright, 2001). All these situations were related to the context of this study, justifying the choice of expert judgment as the technique of choice.

Expert judgment is characterized by anonymity, iteration, controlled feedback, and statistical group response (Dalkey, Brown and Cochran, 1969, Jagatheesan, 2022). Maintaining anonymity necessitates that panel

members remain unfamiliar with each other to mitigate group pressure. The iteration process mandates multiple rounds until information reaches a point of saturation. Controlled feedback involves structuring the feedback-gathering process under the facilitator's guidance. Statistical group response requires the facilitator to furnish experts with a report incorporating statistical processing of panelists' opinions, comments, mean, median, and standard deviation. This feedback serves as a feedforward for the subsequent iteration.

Hsu and Sandford (2007) Suggest distinct stages (Figure 1) and drawbacks associated with the Delphi method, encompassing the risk of reduced response rates, extensive time commitments, and the susceptibility to influencing opinions.

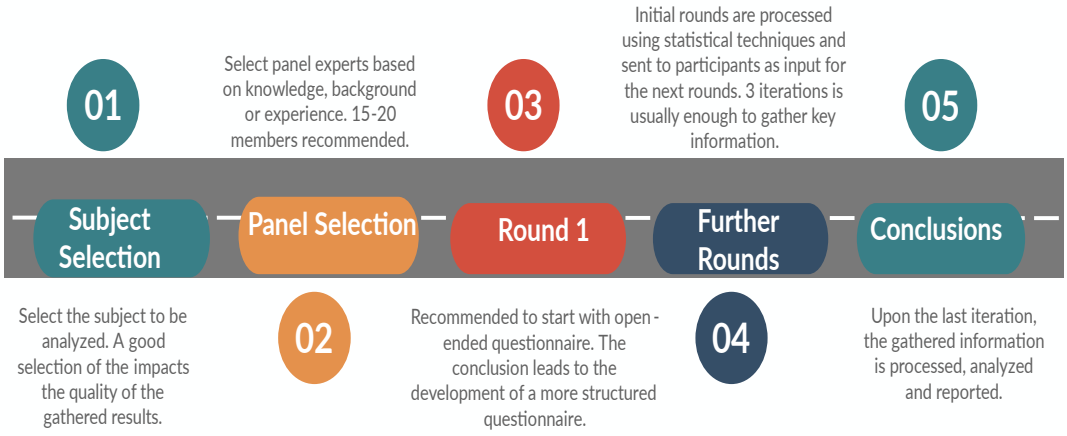

Figure 1: Delphi method overview

Consensus and stability are other vital concepts researchers must consider when using the Delphi method to aggregate expert opinions on future developments and incidents (Von Der Gracht, 2012). Consensus denotes the alignment of opinions toward a specific value, while stability signifies the consistency of values across various iterations. Dajani, Sincoff and Talley (1979) Argued that consensus without stability is meaningless and proposed a decision tree (Figure 2) to guide the achievement of consensus and stability. Stability must be assessed after each round, after which consensus can be examined. The two main approaches to examining consensus include qualitative analysis with descriptive statistics and inferential statistics (Gliem and Gliem, 2003). Qualitative analysis with descriptive statistics involves conducting a subjective analysis to identify a particular level of consensus and incorporating measures such as mean, median, and standard deviation to evaluate and attain consensus among panelists. Inferential statistics, on the other hand, involves diverse statistical metrics to explore various relationships among variables. Metrics such as Chi-square, Cohen's kappa, Fleiss' kappa, or Kendall's W can gauge the level of consensus, contingent on the specified scale.

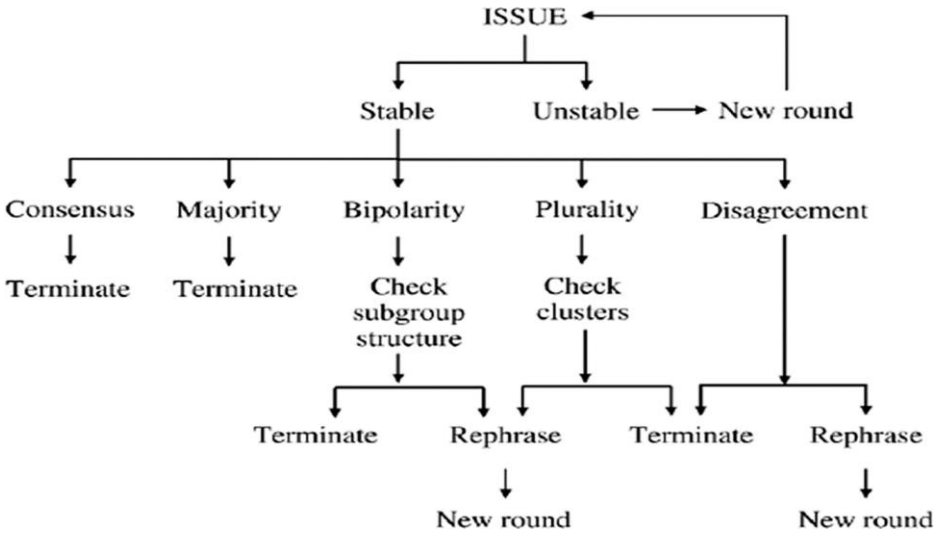

Figure 2: Hierarchical stopping criteria for Delphi methods (DajaniSincoff and Talley, 1979)

2.2 Subject Selection

The first step was to define the subject and design the initial questionnaire. The questionnaire design for the Delphi method involves starting with a well-defined version, clarifying elements in subsequent rounds, and using statistical techniques for ongoing reliability assurance. When formulating a questionnaire for the Delphi method, two acceptable approaches include either utilizing the first round to refine the questionnaire or commencing the first round with an already well-defined version (Hsu and Sandford, 2007). In this study, the latter approach was adopted, given that the elements to be validated were initially clear, though the questionnaire itself may undergo minor modifications based on received feedback. Statistical techniques will be applied throughout the process to maintain reliability.

This study aims to evaluate the Weighted performance Scoring model as a coherent framework that could guarantee the design and evaluation of e-learning courses optimized for student performance in medical education. Therefore, the panel of experts will evaluate three dimensions – first is the extent to which the model guarantees the design of e-learning courses for optimal student performance, and the second is the extent to which the model ensures the evaluation of e-learning courses with a specific focus on students performance in e-learning in medical education. Lastly, the internal coherence of the model is assessed to establish its completeness, coherence, and the presence of conflicting constructs and proposals for further improvement. Hence, the questionnaire is divided into three sections to ensure that each dimension is independent – meaning that the model can help evaluate but not design e-learning courses in medical education or the other way around. Table 2 presents the statements to assess each of the proposed dimensions of the weighted performance scoring model. The design dimension used the user interface and user experience evaluation framework for STEM education (Alomari et al., 2020) to inform the course of inquiry, while the Kirkpatrick evaluation model (Kirkpatrick, 1959) was used to structure questions relative to the evaluation dimension.

Table 2: Proposed statements to assess design, evaluation, and fitness dimensions of the weighted scoring model

ID	Dimension	Statement	Rationale
Q1	Design (UI/UX)	The proposed model helps design an e-learning course that is easy for students to use.	This statement assesses the usability of the online learning environment.
Q2		The proposed model helps to design an e-learning course that satisfies the students' learning needs.	This statement assesses the usefulness of the online learning environment.
Q3		The proposed model helps design an e-learning course that fits the context.	This statement assesses the context of the online learning environment.
Q4		The proposed model helps design an e-learning course that is valuable to students.	This statement assesses the value proposition of the e-learning course to the users.
Q5		The proposed model helps design an e-learning course that uses technology to enhance learning.	This statement assesses how users learn with technology and how technology can facilitate learning.
Q6	Evaluation	The proposed model helps evaluate students' reactions to the e-learning course.	This statement assesses how users felt about the e-learning course.
Q7		The proposed model helps to evaluate how students learned from the e-learning course.	This statement assesses how users acquire knowledge from the e-learning course.
Q8		The proposed model helps to evaluate how students applied what was learned from the e-learning course.	This statement assesses how users applied the acquired knowledge from the e-learning course in a clinical setting.
Q9		The proposed model helps evaluate the benefits of applied knowledge acquired from the e-learning course.	This statement assesses how users applied the acquired knowledge from the e-learning course in a clinical setting.
Q10	Fitness	The model provides a coherent approach to designing e-learning courses in medical education.	This statement measures the effectiveness of the proposed model for e-learning course design in medical education.
Q11		The model provides a coherent approach to evaluating e-learning courses in medical education.	This statement measures the effectiveness of the proposed model for e-learning course evaluation in medical education.

A Likert survey was given to the panel to rate the statements. The scale categorizes responses from "Strongly disagree" (1) to "Strongly agree" (5), allowing individuals to express their agreement or disagreement with a given statement on a continuum. The study employed a unified questionnaire distributed through Google Forms to gather opinions and demographic information from panel members while ensuring their anonymity. The communication and reporting processes maintained strict confidentiality by keeping the panel composition undisclosed and anonymizing results to protect participants' identities.

2.3 Statistical Processing

The analysis employs descriptive statistics, including mean, median, standard deviation, and percentage of agreement and disagreement, to assess experts' consensus on proposed statements, offering an overview of the panel's overall opinion on the framework. Cronbach's alpha is utilized to gauge the reliability of the questionnaire (Cronbach, 1951), examining how the measurement instrument adapts to evaluated magnitudes and how amendments to the questionnaire can impact its reliability across different process rounds. The formula for calculating Cronbach's alpha, denoted as α, is as follows:

$$\alpha = k \times \bar{r} \,/\, 1+(k-1)\times \bar{r} \qquad (1)$$

Where k is the number of items in the test or scale, and $\bar{r}$ is the average of all possible split-half coefficients. The Cronbach alpha reliability interpretation scale categorizes reliability levels as follows: an alpha greater than 0.9 is considered excellent, between 0.9 and 0.8 is deemed good, 0.8 to 0.7 is acceptable, 0.7 to 0.6 is questionable, 0.6 to 0.5 is poor, and an alpha below 0.5 is deemed unacceptable (George, 2011). Simple Correspondence Analysis (SCA) reduces the dimensionality of a matrix while representing the matrix with a two or three-dimensional space for visualization purposes (Nenadic and Greenacre, 2007). SCA calculates the homogeneity of experts' ratings and questions, rates, and tracks how this homogeneity evolves through process rounds, determining whether consensus is achieved among experts. A short distance between ratings or statements implies consensus and homogeneous expert behavior.

2.4 Panel Selection

The literature does not provide standardized recommendations regarding the ideal number and members' profiles in a Delphi method expert panel, as these variables are context-dependent. In our study, experts were selected based on specific criteria to ensure their expertise and relevance to the topic of e-learning in medical education. The criteria included having at least ten years of experience designing and evaluating e-learning courses and programs in medical education. This ensured that the selected experts deeply understood the subject matter and could provide valuable insights during the Delphi study. A total of twenty experts were initially invited to participate in the survey on the 7th of August, 2023. These experts were identified through academic literature, professional associations, and recommendations from other experts in the field. However, only 14 experts responded positively and actively participated in all iterations of the Delphi study. The expert panel consisted of individuals with varying years of experience in designing and evaluating e-learning in medical education. This diversity in experience ensured that the panel represented a wide range of perspectives and insights into the topic.

The Delphi study was conducted in two rounds, each spaced eight weeks apart. In each round, experts were asked to review and provide feedback on statements related to e-learning in medical education. They were also asked to rate their level of agreement with each statement on a scale of 1 to 5. The data collection method involved emailing survey links to the experts on Google Forms. The experts were given a specified period to complete the survey and submit their responses. The process of selecting experts and conducting the Delphi study was carefully designed to ensure the validity and reliability of the results. Figure 3 shows the experiences of experts in designing and evaluating e-learning courses in medical education. The panel members have 12.7 years of experience in designing and 12.4 years in evaluating e-learning courses in medical education.

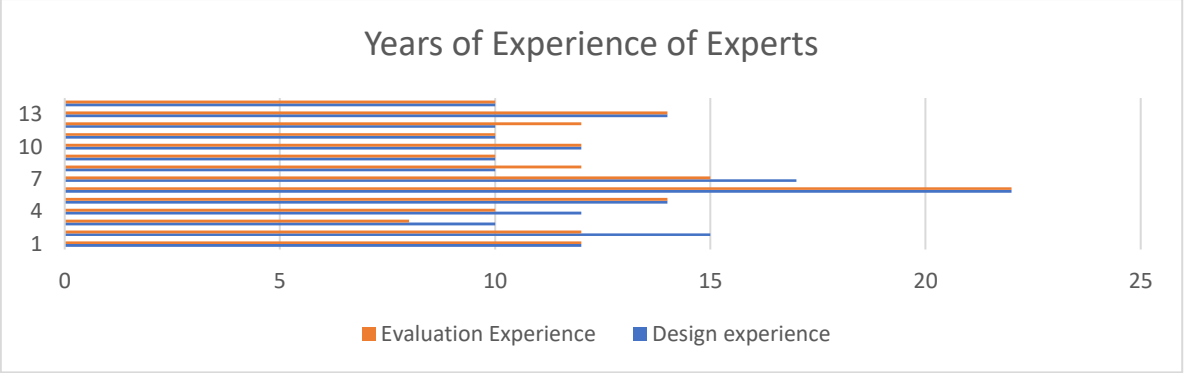

Figure 3: Years of experience of experts

Figure 4 shows the types of organizations of the experts. 86% of the experts work in universities, while 7% work in private organizations and research institutions. Figure 5 shows the work profile of the experts. 50% are professors, 36% are senior lecturers, and 14% are researchers.

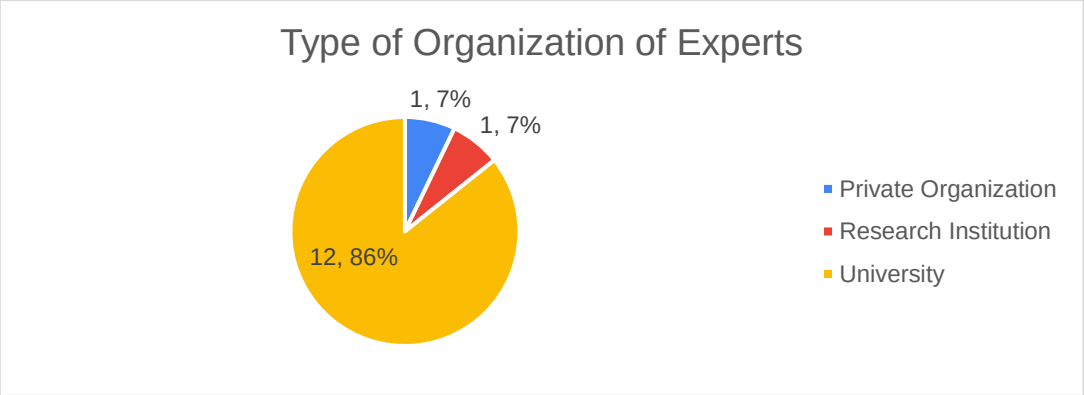

Figure 4: Type of organizations of experts

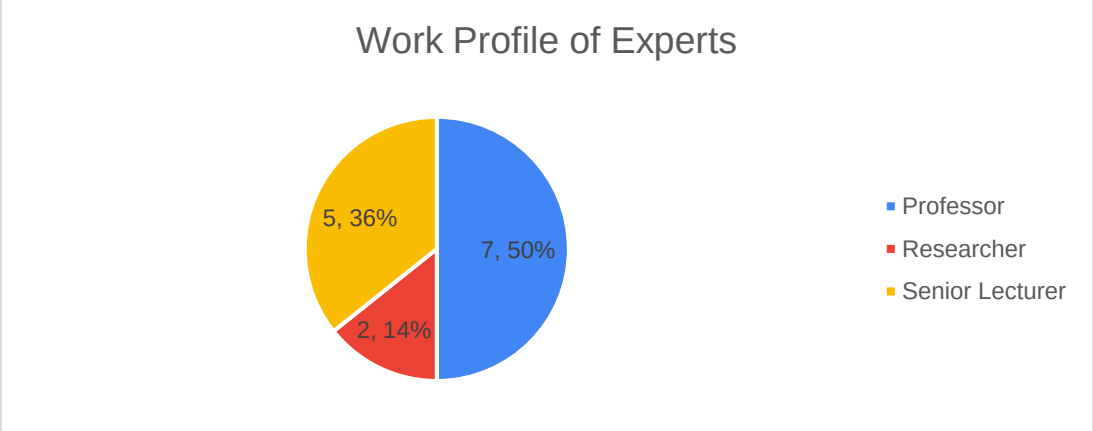

Figure 5: Work profiles of experts

3. Results

3.1 First Round

The first round of the Delphi method was conducted on the 7th of August, 2023, and finished on the 30th of August, 2023. Twenty invitations were sent via email, and only 14 responses were received. Table 3 summarizes the collective results of the phase, including average assessments, medians, standard deviations, and the percentage of agreement and disagreement among the 14 experts for each statement.

Table 3: Collective results for the first round

Dimension	Statement	Average	Median	Standard Deviation	% Agreement	% Disagreement	No opinion
Design	Q1	4.14	4.00	0.52	93	0	7
	Q2	3.57	3.50	0.62	50	0	50
	Q3	3.50	3.00	0.73	36	0	64
	Q4	4.14	4.00	0.52	93	0	7
	Q5	4.21	4.00	0.56	93	0	7
Evaluation	Q6	4.07	4.00	0.46	93	0	7
	Q7	3.86	4.00	0.52	79	0	21
	Q8	3.36	3.00	0.48	36	0	64
	Q9	3.36	3.00	0.48	36	0	64
Fitness	Q10	3.43	3.00	0.49	43	0	57
	Q11	3.57	4.00	0.49	57	0	43

The descriptive analysis revealed that experts do not express any disagreement with the proposed statements as no question received disagreement from all experts. In most statements, the median is 4, indicating that "Agree" is the most frequently chosen value by the raters. However, it's notable that many statements have "Neither agree nor disagree" as the most selected value. The assessment averages show that there is an explicit agreement among experts for seven statements, with values exceeding 3.50. However, four statements exhibit values between 3.50 and 3.36, signaling a lack of consensus. Moreover, the presence of ambiguous answers highlights divergent opinions among experts. After a comprehensive initial descriptive analysis, three primary concerns emerge:

1. Some experts stated that they needed more details concerning the engineering of the model to understand better how the ten constructs were arrived at.
2. Some experts expressed insufficient knowledge to evaluate specific statements, leading them to select Neither agree nor disagree" as a response, signifying a "Don't know/No answer" stance.
3. Other experts highlighted that they could not map each construct of the model to the specific constructs being evaluated in the questionnaire and recommended that a questionnaire be developed to operationalize the measurement of the constructs constituting the model and that the questions on the questionnaire be mapped to the constructs to be evaluated in the next iteration to enhance accuracy in assessment.

Aside from the overall analysis of the questionnaire, the study also performed a detailed descriptive analysis of each dimension. Radar charts were utilized to present the results of each dimension in Figures 6, 7, and 8, while Figure 9 is a visual representation of the experts' agreement for all dimensions using box and whisker graphs.

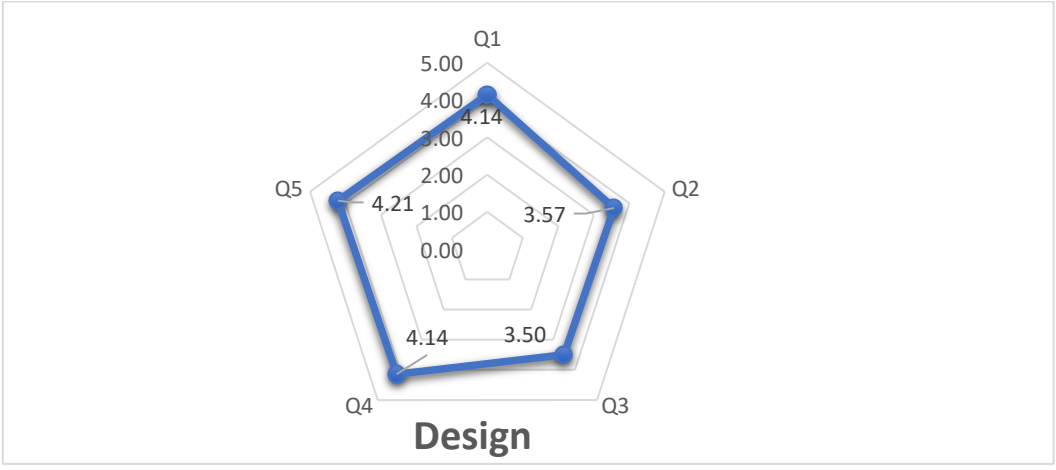

Figure 6: Experts agreement for design dimension

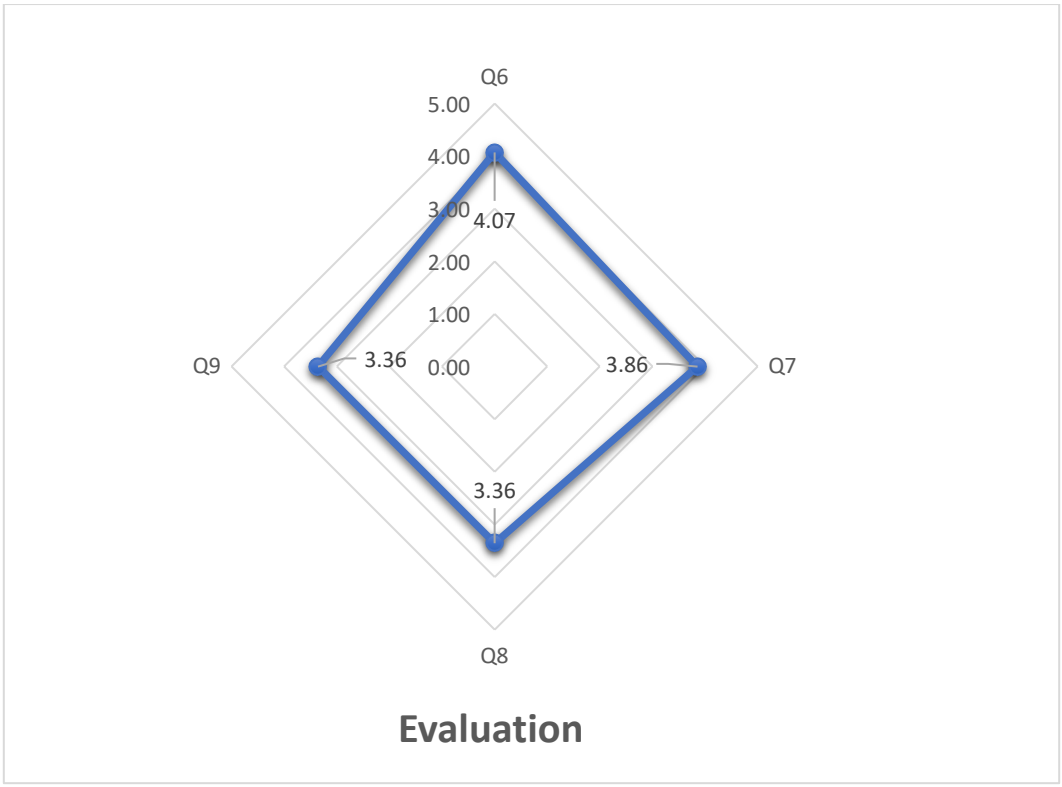

Figure 7: Experts agreement for evaluation dimension

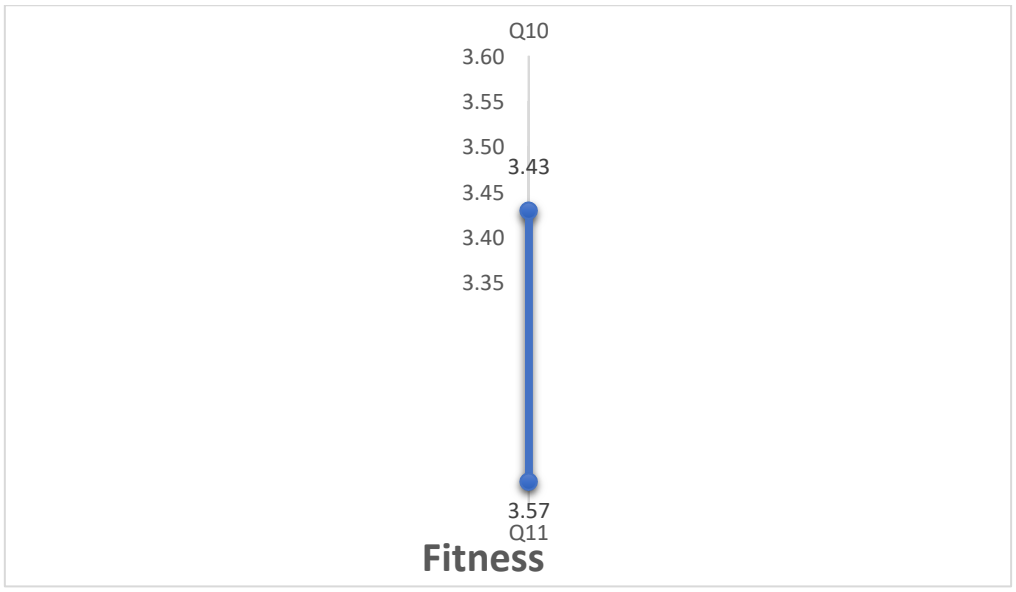

Figure 8: Experts agreement for fitness dimension

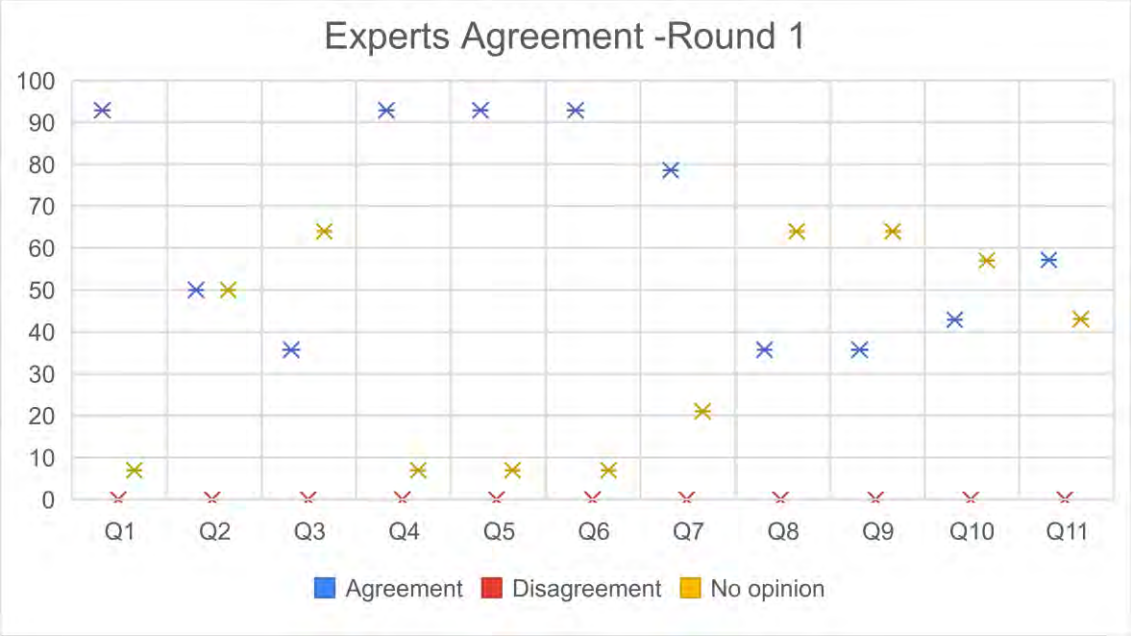

Figure 9: Aggregate of experts' agreement for all statements using box and whiskers graph of the first round

Figures 6 and 9 show that the experts' agreement for the design dimension is high, given that 3 out of 5 statements had agreements of over 90%, the average of all statements in the dimension is 3.91, and the median is 4.14. As reflected in Q2 and Q3, the challenge in this dimension is that experts wanted to understand how learning needs are defined and how the model assists in designing e-learning courses that satisfy students' learning needs. Besides, experts wanted to understand what was meant by context and how the proposed model helps develop an e-learning course that fits the context. Figures 7 and 9 show that experts' agreement for the evaluation dimension was only slightly above average, given that only 25% of the statements had agreements of over 90%, while the average of all statements in the dimension is 3.66, and the median is 3.61. The challenge in this dimension, as reflected in Q8 and Q9, is that experts wanted to understand how the factors in the model can be operationalized to help evaluate how students applied what was learned from the e-learning course and the benefit of the knowledge acquired. Operationalizing the model using questionnaires and mapping factors to the evaluation dimension is crucial for providing clarity in the next round. The fitness dimension was rated comparatively lower than the first two dimensions. Figures 8 and 9 show that experts' agreement for the fitness dimension was average, given that no statements had agreements of over 90% while the average of all statements in the dimension is 3.50 and the median is 3.50. Most experts neither agreed nor disagreed with the statements examining if the model provides a coherent approach to designing and evaluating e-learning courses in medical education. This is not surprising, given the concerns expressed by experts regarding the design and evaluation dimensions.

Homogeneity and concordance analysis

The analysis of homogeneity and concordance followed the descriptive examination to assess reliability. The calculation of Cronbach's alpha was performed on R using two packages, psy and psych, for triangulation and confirmability. This gave a result:

$\alpha = 0.83$ on both packages.

According to George (2011), this value implies good reliability and demonstrates a high internal consistency of the proposed questionnaire. In conclusion of the round, Simple Correspondence Analysis (SCA) was conducted for the entire questionnaire using the R package "ca" and FactoMineR, and the results obtained are depicted in Figure 10. This analysis aimed to visualize the relationships within the comprehensive questionnaire dataset. As illustrated in Figure 11, most experts and statements (except a few experts like experts 2 and 9) are clustered around the central point, signifying a homogeneous perspective among most experts. This outcome is further emphasized when examining the statements that exhibit clustering, with design statements being outliers. In summary, the test reveals a notable level of consensus among experts, indicating a degree of homogeneity in the assessments provided.

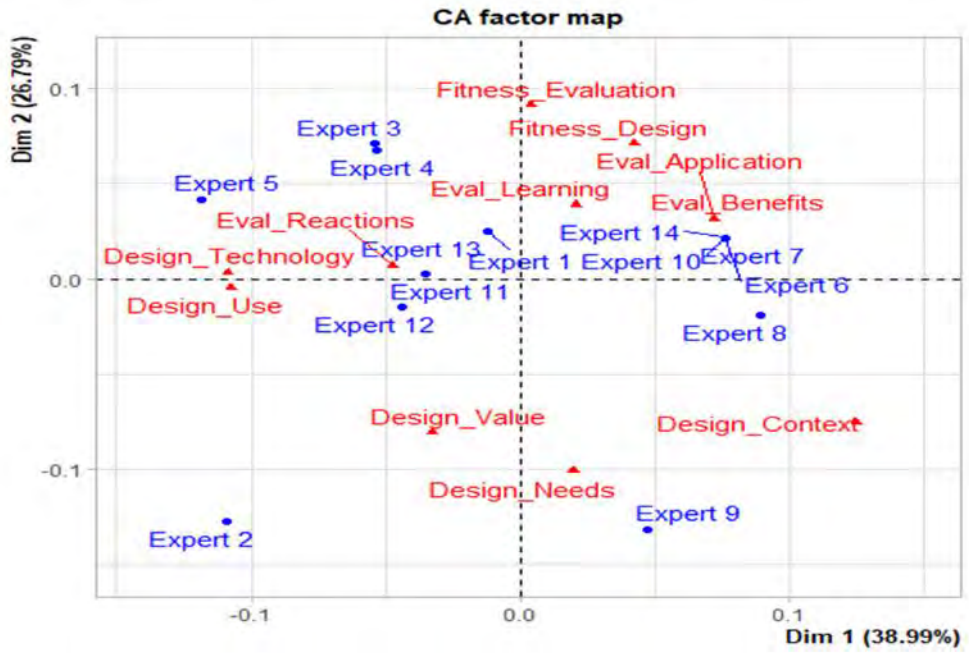

Figure 10: Simple Correspondence analysis of experts' opinions and statements

Round conclusions

The descriptive, homogeneity, and concordance analysis reveal that, while the proposed model demonstrates suitability and a degree of homogeneity in expert responses, the reliability and homogeneity tests affirm that there remain elements requiring improvement and clarification in the experts' opinions. Consequently, a second round of the Delphi method was suggested to the panel. In this round, an anonymized version of the comments and statistical data from Table 3 was provided to the members, the model was operationalized using two types of questionnaires, and modifications were made to the questionnaire based on the feedback received from the experts, as detailed in the next section.

3.2 The Model

The model was developed to standardize e-learning performance evaluation in medical education from the PCA of the factors that constitute the CSF for e-learning in medical education, as shown in Table 4.

Table 4: Correlation and Statistical Significance of Factors used to design the model

Principal Component for Dimension 2 (PC2)	Content Structure	Helpfulness	Appeal	Ease of Navigation	Competence	Interest	Usefulness	Content quality	Suitability	Previous experience
Correlation	8.24E-01	7.05E-01	7.05E-01	6.54E-01	4.54E-01	4.51E-01	4.22E-01	3.43E-01	2.34E-01	2.34E-01
p-value	1.23E-26	9.57E-17	9.57E-17	6.87E-14	1.43E-06	1.78E-06	9.09E-06	3.92E-04	1.72E-02	1.72E-02

Based on these factors' correlation and statistical significance, a system was designed where each dimension contributes to an overall score based on the loading. The performance score for the performance evaluation process or products in e-learning in medical education was calculated using the formula below:

$$P\ Score = \frac{\sum(Correlation\ Weight \times Factor\ Score)}{\sum Correlation\ Weight} \quad (2)$$

Equation 2 is the formula for the weighted performance scoring model. This is further decomposed in Equation 3 by multiplying the correlation weight of each factor by the maximum performance score of each factor.

$$P\ Score = \frac{\begin{array}{c}(0.824 \times \text{Content Structure} + 0.705 \times \text{Helpfulness} + 0.705 \text{ x Appeal} + \\ 0.654 \times \text{Ease of Navigation} + 0.454 \times \text{Competence} + 0.451 \text{ x Interest} + 0.422 \times \text{Usefulness} + \\ 0.343 \times \text{Content Quality} + 0.234 \times \text{Suitability} + 0.234 \times \text{Previous experience})\end{array}}{(0.824+0.705+0.654+0.454+0.451+0.422+0.343+0.234+0.234)} \quad (3)$$

Figure 11 depicts how the weighted scoring model is operationalized to develop various questionnaires to evaluate each factor. The model shows that questions used to assess the factors in the model can be rating scale-type questions. As shown in Table 5, a single question can be used to rate each factor, resulting in a survey containing ten questions. This means that one question might be used to rate each factor. Alternatively, multiple questions can be used to rate each factor, meaning that the researcher might decide to ask as many questions as possible to rate each factor. In this case, the analysis must begin with the calculation of the average score for each of the ten questions used to assess each factor. After this, all factors need to be standardized by calculating the maximum factor score, which is always a maximum of 1. The final step is to calculate the P Score by multiplying the correlation weight of each factor with their factor score, aggregating the result, and dividing the answer by the aggregate of the correlation weight of all ten factors. The maximum possible Performance Score obtained using this formula is 1, while the minimum Score is Zero.

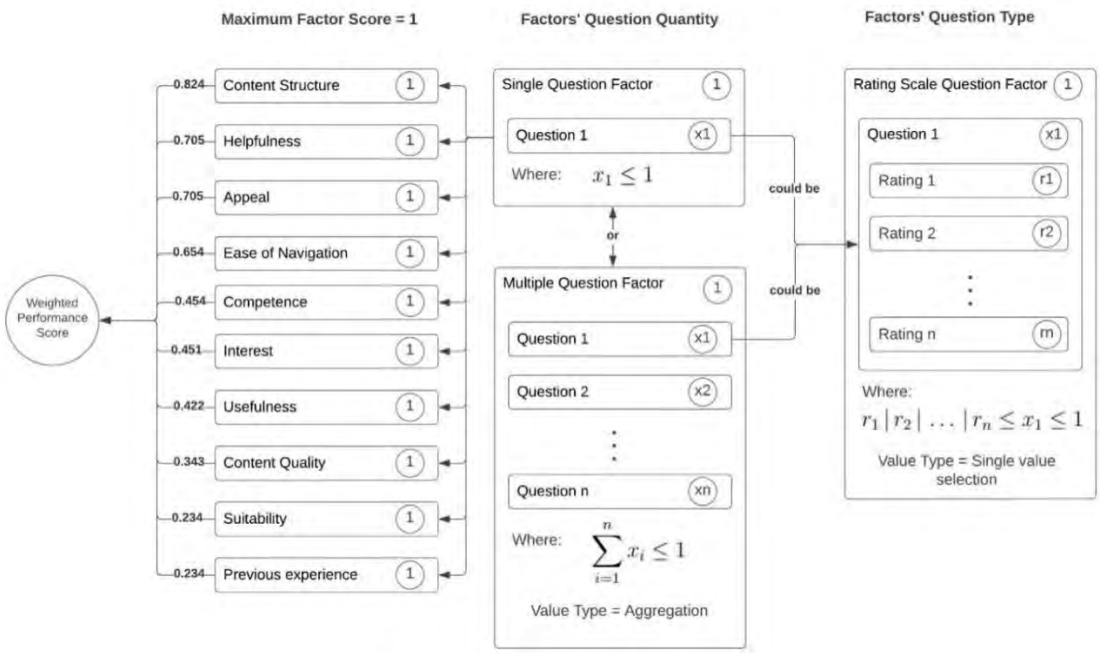

Figure 11: Operationalization of the Weighted Performance Scoring Model

Model operationalization and mapping to assessment statements

The model was operationalized using single rating and multiple rating questions to depict how it could be used for evaluating Key Performance Metrics (KPM) and as CSF for design. For each question, participants choose the rating that best represents their opinions on a scale of 1-5 (1: Strongly Disagree, 2: Disagree, 3: Neutral, 4: Agree, 5: Strongly Agree). The questions were mapped to the proposed statements to assess the design, evaluation, and fitness dimensions of the weighted scoring model as was structured in the questionnaire given to experts. Table 5 presents the operationalization and mapping of the model.

Table 5: Single and multiple rating questions and mapping to proposed statements for expert judgment.

SINGLE QUESTION RATING		
Factors	**Questions**	**Mapping to Statements**
Content Structure	The content of the e-learning course was well-structured.	**Q1, Q6**
Helpfulness	The course content was sufficient to help me prepare for my exam.	**Q2, Q7**

SINGLE QUESTION RATING		
Factors	**Questions**	**Mapping to Statements**
Appeal	The layout of the screen for the online course was appealing.	**Q1**
Ease of Navigation	The platform used for e-learning was easy to navigate.	**Q1**
Competence	I feel confident in my competence in applying the knowledge I acquired from the course.	**Q8**
Interest	My interest in the subject was enhanced significantly after the course.	**Q7, Q4**
Usefulness	I found the content of the online course useful for my future medical work.	**Q9**
Content Quality	The content quality of the online course was high.	**Q6**
Suitability	The course was suitable for online teaching.	**Q3, Q5**
Previous Experience	The online course was not my first online learning experience.	**Q3**
MULTIPLE QUESTION RATING		
Factors	Questions	
Content Structure	The content of the e-learning course was well-structured.	**Q1, Q6**
	The structure of the online course made the course useful for my future medical work.	**Q2, Q8, Q9**
	The structure of the course helped me learn quickly	**Q2, Q7**
Helpfulness	The course content was sufficient to help me prepare for my exam.	**Q2, Q4, Q7**
	The course assessments were sufficient to prepare me for my exam.	**Q2, Q7**
	The course had sufficient information to help me prepare for real-world problems.	**Q9**
	The layout of the screen for the online course was appealing.	**Q1**
	An appealing screen layout inspired me to engage more with the learning resources.	**Q5, Q6**
	An appealing screen layout inspired me to engage more with my peers.	**Q5, Q6**
Ease of Navigation	The platform used for e-learning was easy to navigate.	**Q1**
	I did not encounter any difficulty while moving from one section to another.	**Q1**
	The navigation menus were intuitive.	**Q1**
Competence	I feel confident in my competence in applying the knowledge I acquired from the course.	**Q8**
	I felt competent in my ability to use the e-learning system.	**Q3, Q5**
	I feel more competent taking another e-learning course in the future.	**Q2, Q6**
Interest	I found the online course interesting.	**Q7, Q4**
	My interest in the subject was enhanced significantly after the course.	**Q4**
	I participated in most of the online discussions.	**Q5**
Usefulness	I found the content of the online course useful for my future medical work.	**Q4, Q9**
	The online course was useful in facilitating my learning experience.	**Q2, Q5**
	The online course was useful because it improved my knowledge.	**Q2, Q7**
Content Quality	The content quality of the online course was high.	**Q6**
	The content quality of the online course made my learning experience seamless.	**Q5, Q6**
	The quality of the content made me comfortable with online learning.	**Q1, Q6**

SINGLE QUESTION RATING		
Factors	**Questions**	**Mapping to Statements**
Suitability	The course was suitable for online teaching.	**Q3, Q5**
	The content of the course was easy for me to understand because it was taught online.	**Q3, Q5, Q7**
	The assessments of the course were well-suited for online learning.	**Q3, Q5**
Previous Experience	The online course was not my first online learning experience.	**Q3**
	My previous experience in online learning helped me perform better in the course.	**Q3, Q8**
	I acquired digital skills from the online course that I could use for future courses.	**Q4**

Given the clarity provided on the model to experts, the same questionnaire was used for the second round because there were no complaints concerning the clarity of the questionnaire but rather the clarity of the model.

Descriptive analysis

The second round of the Delphi method was conducted on the 28th of September, 2023, and concluded on the 30th of October, 2023. Fourteen invitations were sent via email, and all 14 responses were received. Table 6 summarizes the collective results of the phase, including average assessments, medians, standard deviations, and the percentage of agreement and disagreement among the 14 experts for each statement.

Table 6: Collective results for the second round

Dimension	Statement	Average	Median	Standard deviation	% Agreement	% Disagreement	No opinion
Design	Q1	4.50	4.50	0.50	100	0.0	0.0
	Q2	4.43	4.00	0.49	100	0.0	0.0
	Q3	4.43	4.00	0.49	100	0.0	0.0
	Q4	4.57	5.00	0.49	100	0.0	0.0
	Q5	4.50	4.50	0.50	100	0.0	0.0
Evaluation	Q6	4.71	5.00	0.45	100	0.0	0.0
	Q7	4.64	5.00	0.48	100	0.0	0.0
	Q8	4.57	5.00	0.49	100	0.0	0.0
	Q9	4.57	5.00	0.49	100	0.0	0.0
Fitness	Q10	4.36	4.00	0.48	100	0.0	0.0
	Q11	4.43	4.00	0.49	100	0.0	0.0

Table 6 shows an increase in mean, median, and standard deviation, implying that experts agree with the proposed statements in all dimensions, with the average of all statements being 4.52 and the median of all statements being 4.50. This means that experts are convinced that the model is fit for efficiently designing and evaluating e-learning courses in medical education. This might be due to the detailed information provided to experts via email concerning the model's need, purpose, journey, and outputs to give more clarity. Explaining and operationalizing the model using two questionnaires and consequent mapping to the statements of the instrument used in round one of the Delphi method proved effective. The box and whisker graph (figure 12) shows the aggregate of experts' agreement for all statements in the second round.

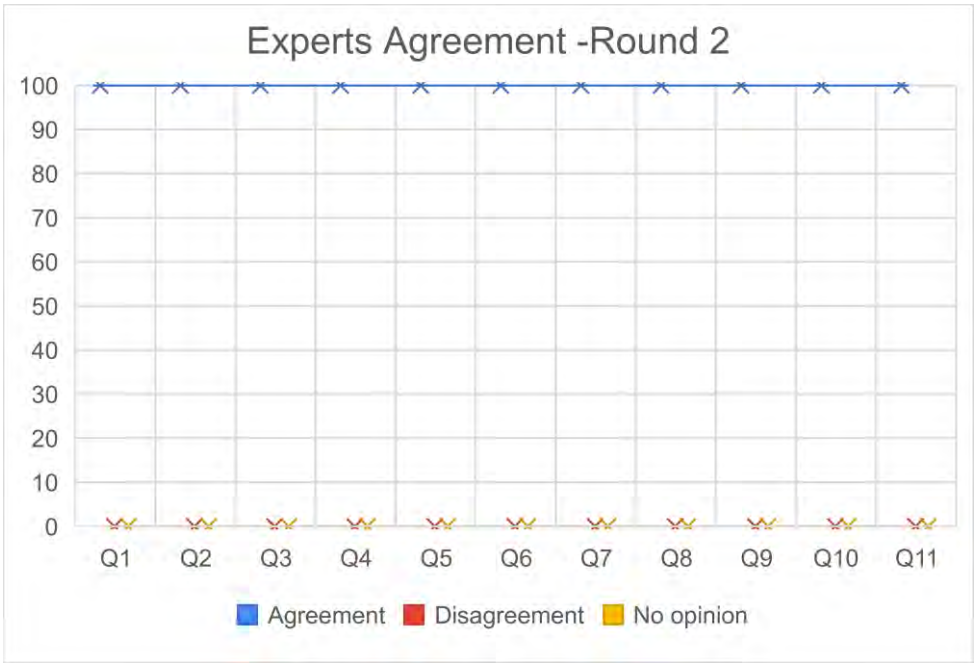

Figure 12: Aggregated experts' agreement for all statements using the box and whiskers graph from the first round

Delphi Conclusion

The panel reached a high level of agreement on all the proposed statements in the second round (figure 13). Strong agreements are confirmed when specific criteria are met, including an average assessment of 3.7 or higher, a median of 4, and at least 60% agreement from a minimum of 12 experts. Slight agreement is confirmed with an average assessment between 3.5 and 3.7, a median equal to or higher than 3.5, and at least 45% agreement from a minimum of 12 experts.

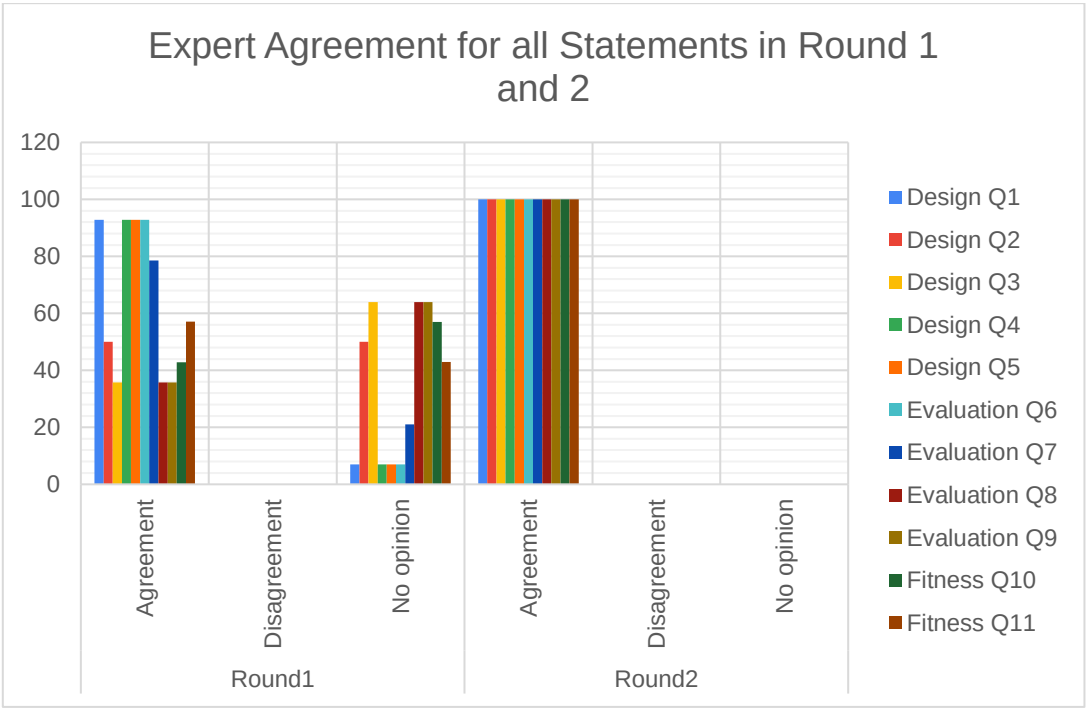

Figure 13: Expert agreement for all statements in rounds 1 and 2.

All 11 statements had an average of 52% strong agreement and 48% agreement from all 14 experts, with the evaluation dimension having the strongest agreement, next to the design dimension. The validity of the fitness

dimensions relies on the results of the first two dimensions. Although all experts agreed that the model presents a coherent and efficient approach to designing and evaluating e-learning courses in medical education, an expert suggested that the results of analysis using the model be correlated to the actual average class performance score in the e-learning course. This is a research area to be considered in the future. Also, another expert proposed that the model be expanded for performance prediction modeling and that the prediction graph be correlated to the performance graph when evaluating e-learning courses in medical education. This is also perceived as a call for future research.

4. Conclusion

E-learning is a comprehensive concept encompassing the asynchronous or synchronous dissemination of knowledge to learners via electronic systems. Recently, e-learning has garnered significant recognition as a mainstream approach in health sciences education (HSE), encompassing medical, dental, public health, nursing, and other allied healthcare disciplines. However, considerable debate remains surrounding the role of e-learning, its impact on learners' performance and learning enhancement. In this study, a comprehensive content validation of a model designed to assess the performance of e-learning in medical education through expert judgment was conducted using the Delphi method. The model underwent rigorous scrutiny from 14 experts. Our analysis reveals that the model demonstrates a high level of agreement among experts, meeting the predefined criteria for both strong and slight agreement. The dimensions related to evaluation and design garnered the strongest agreements, emphasizing the model's coherence and efficacy in designing and evaluating e-learning courses in medical education. The Delphi method proved a suitable and effective expert judgment technique for this study, allowing for anonymity, iteration, controlled feedback, and statistical group response. The analysis incorporated descriptive statistics, Cronbach's alpha reliability prediction, and Simple Correspondence Analysis (SCA) to assess consensus, reliability, homogeneity, and stability among experts. The panel's high level of agreement on the proposed statements after the second round affirms the robustness of the model.

- This study significantly contributes to e-learning in medical education by conducting a content validation of a performance assessment model. It provides a comprehensive overview of expert validation methods, shares practical outcomes, and offers preliminary insights into the model's validity, paving the way for future research directions. The contribution provided by this study includes the following:
- The study provides a thorough overview of the factors essential for crafting an expert validation method tailored explicitly for assessing the performance of a weighted performance scoring model in e-learning.
- It shares practical outcomes of applying the expert judgment method to a weighted performance scoring model in a real-world scenario within the context of medical education. This practical insight adds value by demonstrating the model's applicability and effectiveness in an educational setting.
- It advances our understanding of the validity of the weighted performance scoring model by presenting preliminary findings across its various dimensions. This empirical evidence adds depth to the current knowledge regarding the model's effectiveness and potential areas for refinement.
- It concludes with significant insights from the research findings and proposes future avenues for continued exploration and enhancement of performance evaluation models in e-learning within the medical education domain. This contributes to ongoing discussions and guides future research endeavors in this field.

5. Limitations and Future Work

While our study provides valuable insights into the content validity of the proposed model, certain limitations should be acknowledged. The Delphi method, while effective, relies on expert opinions, and variations in expertise or perspectives may impact the results. Additionally, the study focused on the dimensions of design and evaluation; future research should consider expanding the model to address performance prediction and correlation with actual class performance scores, as suggested by the experts. Future research endeavors should explore the correlation between the results obtained through the model and actual class performance scores in e-learning courses in medical education. Additionally, there is a call for extending the model to encompass performance prediction modeling, with a focus on correlating prediction graphs with performance graphs during the evaluation of e-learning courses. These suggestions pave the way for further refinement and application of the model, contributing to the ongoing discourse on enhancing the effectiveness of e-learning in the medical education domain.

Statement of Open Data, ethics, and conflict of interest

Requests for the data can be addressed to the corresponding author. The approval for conducting this research was received from the University of KwaZulu-Natal, South Africa. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Ajenifuja, D. & Adeliyi, T. 2022. Evaluating the Influence of E-Learning on the Performance of Healthcare Professionals in Providing Support. *International Journal on E-Learning,* 21**,** 201-215.

Almenara, J. C. & Cejudo, M. D. C. L. 2013. La aplicación del juicio de experto como técnica de evaluación de las tecnologías de la información y comunicación (TIC). *Revista Eduweb,* 7**,** 11-22.

Alomari, H. W., Ramasamy, V., Kiper, J. D. & Potvin, G. 2020. A User Interface (UI) and User eXperience (UX) evaluation framework for cyberlearning environments in computer science and software engineering education. *Heliyon,* 6.

Alonso, J. 2015. Cálculo e interpretación del Alfa de Cronbach para el caso de validación de la consistencia interna de un cuestionario, con dos. February.

Carney, O., Mcintosh, J. & Worth, A. 1996. The use of the nominal group technique in research with community nurses. *Journal of advanced nursing,* 23**,** 1024-1029.

Cook, D. A. 2007. Web-based learning: pros, cons and controversies. *Clinical medicine,* 7**,** 37.

Cronbach, L. J. 1951. Coefficient alpha and the internal structure of tests. *psychometrika,* 16**,** 297-334.

Dajani, J. S., Sincoff, M. Z. & Talley, W. K. 1979. Stability and agreement criteria for the termination of Delphi studies. *Technological forecasting and social change,* 13**,** 83-90.

Dalkey, N. C., Brown, B. B. & Cochran, S. 1969. *The Delphi method: An experimental study of group opinion*, Rand Corporation Santa Monica, CA.

Escobar-Pérez, J. & Cuervo-Martínez, Á. 2008. Validez de contenido y juicio de expertos: una aproximación a su utilización. *Avances en medición,* 6**,** 27-36.

Farhan, M. K., Talib, H. A. & Mohammed, M. S. 2019. Key factors for defining the conceptual framework for quality assurance in e-learning. *Journal of Information Technology Management,* 11**,** 16-28.

George, D. 2011. *SPSS for windows step by step: A simple study guide and reference, 17.0 update, 10/e*, Pearson Education India.

Gliem, J. A. & Gliem, R. R. Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. 2003. Midwest Research-to-Practice Conference in Adult, Continuing, and Community ….

Güllü, A., Kara, M. & Akgün, Ş. 2024. Determining attitudes toward e-learning: what are the attitudes of health professional students? *Journal of Public Health,* 32**,** 89-96.

Hsu, C.-C. & Sandford, B. A. 2007. The Delphi technique: making sense of consensus. doi: 10.7275. *PDZ9-TH90*.

Jagatheesan, N. R. D. S. 2022. A Delphi based Expert Judgment Techniques applied with Capability Maturity Model Integration to validate the Agile Scrum based MVP Architecture Framework for Android Mobile Application Development. *NeuroQuantology,* 20**,** 1003.

Khasawneh, R., Simonsen, K., Snowden, J., Higgins, J. & Beck, G. 2016. The effectiveness of e-learning in pediatric medical student education. *Medical education online,* 21**,** 29516.

Kirkpatrick, D. 1959. Four-level training evaluation model. *US Training and Development Journal,* 13**,** 34-47.

Mennin, S. 2021. Ten global challenges in medical education: wicked issues and options for action. *Medical Science Educator,* 31**,** 17-20.

Murphy, M. P. 2020. COVID-19 and emergency eLearning: Consequences of the securitization of higher education for post-pandemic pedagogy. *Contemporary Security Policy,* 41**,** 492-505.

Nagar, S. 2020. Assessing Students' perception toward e-learning and effectiveness of online sessions amid COVID-19 Lockdown Phase in India: An analysis. *UGC Care Journal,* 19**,** 272-291.

Nenadic, O. & Greenacre, M. 2007. Correspondence analysis in R, with two-and three-dimensional graphics: The ca package. *Journal of statistical software,* 20**,** 1-13.

Oluwadele, D., Singh, Y. & Adeliyi, T. 2023a. An Explorative Review of the Constructs, Metrics, Models, and Methods for Evaluating e-Learning Performance in Medical Education. *Electronic Journal of e-Learning,* 21**,** 394-412.

Oluwadele, D., Singh, Y. & Adeliyi, T. T. 2023b. Trends and insights in e-learning in medical education: A bibliometric analysis. *Review of Education,* 11**,** e3431.

Raspopovic, M., Jankulovic, A., Runic, J. & Lucic, V. 2014. Success factors for e-learning in a developing country: A case study of Serbia. *International Review of Research in Open and Distributed Learning,* 15**,** 1-23.

Regmi, K. & Jones, L. 2020. A systematic review of the factors–enablers and barriers–affecting e-learning in health sciences education. *BMC medical education,* 20**,** 1-18.

Rowe, G. & Wright, G. 2001. Expert opinions in forecasting: the role of the Delphi technique. *Principles of forecasting: A handbook for researchers and practitioners***,** 125-144.

Von Der Gracht, H. A. 2012. Consensus measurement in Delphi studies: review and implications for future quality assurance. *Technological forecasting and social change,* 79**,** 1525-1536.

