Computers & Education: X Reality 2 (2023) 100027

Contents lists available at ScienceDirect

Computers & Education: X Reality

journal homepage: www.journals.elsevier.com/computers-and-education-x-reality

Cognitive interviewing as a method to inform questionnaire design and validity - Immersive Technology Evaluation Measure (ITEM) for healthcare education

Chris Jacobs [a,*], Joshua Wheeler [a], Michael Williams [b], Richard Joiner [a]

[a] *University of Bath, UK*
[b] *Queen's University Belfast, UK*

ABSTRACT

Research of immersive technology in education is rapidly expanding with potential to educate future students in healthcare disciplines. Despite increasing literature there is a lack of validated instruments to investigate the effects of these technologies. Cognitive interviewing is a valuable evaluation method to check comprehension of a measure and was applied to a new measure of user experience of immersive technology for healthcare education (ITEM). A 5 domain self-reported measure of: immersion, intrinsic motivation, cognitive load, system usability, and debrief. Prior to the interview 9 participants were allocated to augmented reality and virtual reality educational activities. Verbal probing and think aloud techniques through semi-structured cognitive interviews were conducted. The ITEM was found to have high content validity index scores and relationships between domains were further explored through qualitative analysis. The results indicate high clarity of understanding for those completing the ITEM and supports future research as part of an ongoing validation process.

1. Introduction and measure development

There is a demand to provide learning environments that simulate complex clinical scenes and foster learning of key knowledge and skills related to the health sciences. Traditionally, high fidelity simulation-based education, for example using mannequins, has provided this successfully (Lateef, 2010). However, resourcing (both of personnel and equipment) can be challenging.

One method of creating high-fidelity clinical experiences is using wearable immersive technology. Immersive technology creates multiple facets of a simulated environment as if it is real, which can be achieved by devices which create multimodal sensory stimuli (Forrest & McKimm, 2019). Fig. 1 expands on this definition to include Mixed Reality (MR), Virtual Reality (VR) and Augmented Reality (AR).

Immersive technology in healthcare education is a rapidly growing area of application and research (Jacobs, Foote, Joiner, & Williams, 2022; Tang, Chau, Kwok, Zhu, & Ma, 2022). Health Education England (HEE) outlined the need evidence-based approach when we consider the integration of these technologies (Baxendale, Shinn, Munsch, & Ralph, 2020).

Immersive technology in healthcare education has the advantages of replicating clinical experiences in a virtual or augmented environments (Ryan et al., 2022). The technologies have unique attributes that can improve learning outcomes (Logeswaran, Munsch, Chong, Ralph, & McCrossnan, 2021), whether this is in VR surgical simulation (Andersen et al., 2020), simulating emergency medicine triage in VR (Mills et al., 2020), or anatomical learning in AR (Moro et al., 2021).

Learner-centred pedagogy can observe improvements in procedural skills with deliberate practice to acquire new microskills to a mastery learning approach in VR (Andersen, Konge, Thomasen, & Sorensen, 2016) and AR (Lia et al., 2018, p. 10576). Kolb's experiential learning whereby experience is converted to knowledge and skill has been considered as a theoretical pedagogy for airway management learning in VR (Samosorn et al., 2020). Cognitive frameworks involve an internal mechanism whereby information is re-organised with working and long-term memory (Khalil & Elkhider, 2016). Mayer's cognitive theory considers the sensory processing from dual process of audio and visual to multimedia (Mayer, 2014), which can be interpreted as a system with limited cognitive demand (cognitive load). Immersive media has been described as multi-sensory and the Cognitive Affective Model of Immersive Learning (CAMIL) proposes important factors to learning that included: immersion, fidelity, agency, interest, motivation, and cognitive load (Makransky & Petersen, 2021). Immersive technology affords cognitive, procedural, and affective domains first described in Bloom's taxonomy (Adams, 2015).

A scoping review that summarised the sources of evidence supporting

[*] Corresponding author.
E-mail address: cj511@bath.ac.uk (C. Jacobs).

https://doi.org/10.1016/j.cexr.2023.100027
Received 31 March 2023; Received in revised form 16 May 2023; Accepted 16 May 2023

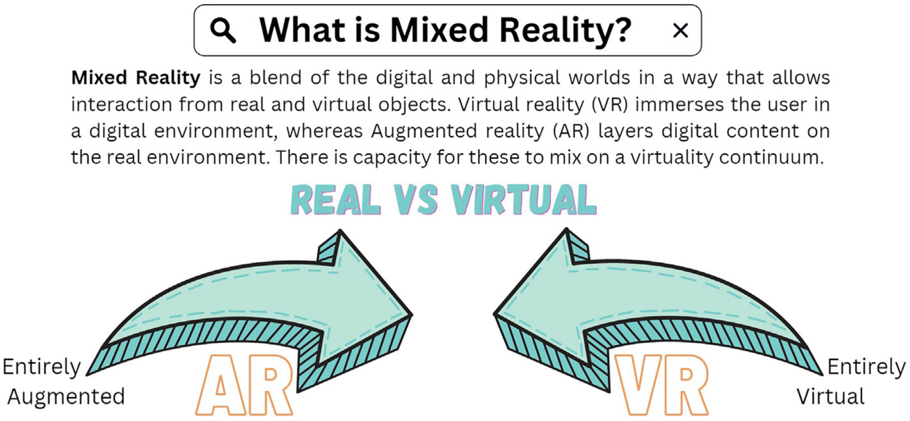
What is Mixed Reality?
Mixed Reality is a blend of the digital and physical worlds in a way that allows interaction from real and virtual objects. Virtual reality (VR) immerses the user in a digital environment, whereas Augmented reality (AR) layers digital content on the real environment. There is capacity for these to mix on a virtuality continuum.
REAL VS VIRTUAL
Entirely Augmented
AR
VR
Entirely Virtual

Fig. 1. Reality-virtuality continuum adapted from Milgram and Kishino's framework.

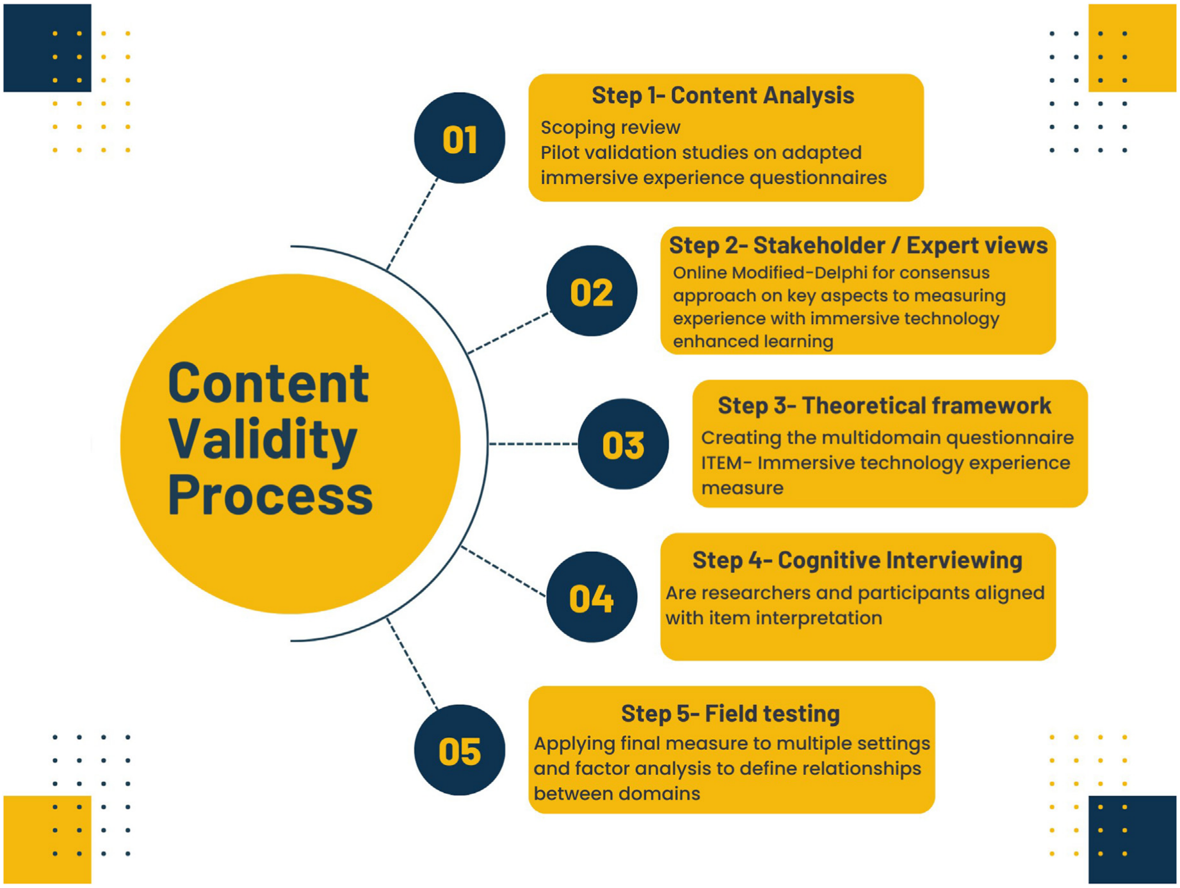
Content Validity Process
01
Step 1- Content Analysis
Scoping review
Pilot validation studies on adapted immersive experience questionnaires
02
Step 2- Stakeholder / Expert views
Online Modified-Delphi for consensus approach on key aspects to measuring experience with immersive technology enhanced learning
03
Step 3- Theoretical framework
Creating the multidomain questionnaire ITEM- Immersive technology experience measure
04
Step 4- Cognitive Interviewing
Are researchers and participants aligned with item interpretation
05
Step 5- Field testing
Applying final measure to multiple settings and factor analysis to define relationships between domains

Fig. 2. Conceptual framework to Immersive Technology Evaluation Measure (ITEM) development with content validity considered at every stage.

the use of immersive technologies in healthcare education was a review of literature informing on this study. Using the Medical Education Research Study Quality Instrument (MERSQI) found a general lack of validated measures, however, researchers that proposed instruments with theoretical background scored highest (Jacobs, Foote, Joiner, & Williams, 2022). Similarly, there is a need for a “validity argument” and to apply a framework to validation of a new instrument(David A. Cook & Hatala, 2016). Hence, validation is a process of collecting evidence to evaluate the relevance of any outcome measures, utimately influencing the confidence that can be placed in decisions made based on those outcomes.

1.1. *Measure development*

The authors applied a conceptual framework to formulate a measure that captures the user experience of immersive technology enhanced learning (TEL). Fig. 2 outlines a content validity approach. The new measure was called the ‘Immersive Technology Evaluation Measure’ (ITEM). This study focused on the cognitive interviewing aspect of the process.

Messick (American Educational Research, 2018; Messick, 1995) proposed a framework that encompasses five potential sources of validity. These sources are compared in Table 1 with the gathering evidence for the validity of the new ITEM measure.

Table 1
Classical and contemporary validity frameworks in relation to Immersive Technology Evaluation Measure (ITEM). AIEQ- Adapted Immersive Experience Questionnaire, AIMI- Abridged Intrinsic Motivation Inventory. MCQ-Multiple choice questions.

Source and type of validity evidence	Definition	Publication findings
Content	Relationship between test items and the construct it is represents	The 'Immersive Experience Questionnaire' (AIEQ) and 'Intrinsic motivation Inventory' (AIMI) were amended for healthcare use. In a mixed method study reviewing VR in healthcare simulation, significantly higher scores in user immersion and intrinsic motivation with immersive technology enhanced learning were found. (Jacobs & Maidwell-Smith, 2022). AR healthcare simulation pilot study reported moderate level of immersion using AIEQ. Additionally, high levels of enjoyment and usefulness reported using AIMI (George, Foster, Xia, & Jacobs, 2023). A scoping review (Jacobs, Foote, Joiner, & Williams, 2022) informed a modified-online Delphi study to create a 5 domain questionnaire. This paper describes the results of verbal probing in subsequent cognitive interviewing, on overall content.
Internal structure and construct	How the items within the questionnaire relate to overall construct	Internal reliability of the AIEQ and AIMI, and criterion validity of immersion questioning were all high (Jacobs & Rigby, 2022). Future work is planned with ITEM exploratory factor analysis.
Relationship with variables	Can test scores be interpreted with underlying proposed construct	Correlation with immersion and motivation scores was high (Jacobs & Rigby, 2022). Qualitative analysis of immersive TEL users was modelled on Kolb's theory of learning. However, no correlation was found with learning scores (as tested by MCQ), nor AIEQ and AIMI scores with learning outcomes. limitations to the study existed with small sample size and potential ceiling effect of MCQ (Jacobs & Maidwell-Smith, 2022).
Response process	The engagement of users of tool and how it fits in detail with performance	Verbal probing method of cognitive interviewing of the new measure, 'ITEM', is described in this study.
Consequences	The impact, whether benefitting or harming of assessment.	Administering or taking test with analysis of the impact of their scores on those being assessed is planned. Feedback was sought on verbal probing experience.

All aspects of validity are important. It is essential for example that the content assessment of items maps to the overall construct that they are intended to measure. The internal structure relates to the individual items, such as, reliability and item difficulty (David A. Cook & Hatala, 2016). Preliminary work on the design of ITEM incorporated the domains of immersion and motivation of participants experiencing immersive TEL: internal reliability and criterion validity evaluated. Furthermore, a positive correlation was found between immersion and motivation of participants that suggests a relationship exists.

The scoping review identified over 100 measures used by researchers in evaluating MR in healthcare education. VR was the most established group of technology in simulation research, with all MR groups evaluated for their validity assessment using MERSQI tool. Measures were appraised and itemised as: objective, subjective, and physiological measurements, or instruments. The most common method of recording outcome was a procedural score generated from a VR system. Secondary measures commonly adopted to assess learning were a multiple-choice questionnaire. Both subgroups scored low on the validity MERSQI assessment. Additionally, these both require input from the immersive media content to construct the test. Numerous subjective measures were identified in the scoping review, which were also presented to the group Delphi.

An online-modified Delphi was undertaken to solicit stakeholder views, to guide and support ITEM domain construct (Jacobs, Foote, & Williams, 2022). Educationalists involved in simulation from a range of backgrounds took part: doctors, technicians, researchers, and administrators, and were invited to consider the way that the experience of those engaged in immersive TEL can be measured. The consensus with 100% agreement were 7 factors should be measured: what was learnt, the degree of immersion experienced, fidelity provided, debrief, psychological safety, patient safety, and cost effectiveness. There was 100% agreement that participant experience should be used as the most efficient method of collecting data. Additionally, there was support for a generalisable measure that can be used with different technologies and settings.

After the Delphi process, a multidomain instrument was created that incorporated the results of prior work (Table 1) and also stakeholder views into a theoretical framework.

Five domains of a 40-question instrument forming the developing measure, ITEM, were:

(1) Adapted immersion experience questionnaire (AIEQ)- contains additional questions relating to fidelity.
(2) Abridged intrinsic motivation inventory (AIMI)- collates thoughts on learning and value of activity for user.
(3) NASA Task Load Index (NASA TLX)- measures the mental and physical strain of the activity that forms the cognitive load (Hart, 2006).
(4) System Usability Scale (SUS) of technology-confidence and accessibility of technology are evaluated (Brooke, 1996).
(5) Prompts for Engaging and Reflective Learning in Simulation (PEARLS)- a debriefing tool that can be used in healthcare simulation to facilitate reflective learning and promote the transfer of learning from the simulation to the clinical setting (Eppich & Cheng, 2015). This domain is optional, as not all immersive TEL involves debrief.

Early work testing and adapting measures of motivation and immersion (AIMI and AIEQ) for use in healthcare education were conducted using VR technology and simulation. These demonstrated high internal validity scores (Jacobs & Rigby, 2022). Alternative questionnaires exist for quantifying these concepts. For example, Core Elements of the Gaming Experience (CEGE), and instructional materials motivation survey (IMMS). CEGE domains of interest are game-centric and feature scales on puppetry and gameplay less suitable for healthcare aims (Calvillo Gamez & Cox, 2010). IMMS based on Keller's motivational theory (Henssen et al., 2020) has similarities to self-determination theory that is the foundation to AIMI, however, IMMS validity has not been evaluated with newer technology and limited use in healthcare education (D. A. Cook, Beckman, Thomas, & Thompson, 2009).

NASA TLX and SUS were both the most commonly used psychometrics to evaluate the cognitive load and technology usability identified in the scoping review, hence, the choice of these in 2 domains.

The validity of a measure is dependent on the user's understanding of its content, and so problems can include the use of complex or inappropriately technical language and incomprehension, altered interpretation of question meaning, and cognitive processing challenges (Oksenberg & Kalton, 1991).

Part of establishing evidence for validity is its pre-test evaluation. The preliminary version of the tool was reviewed in a focus group of medical educators, and questionnaire item revision through direct feedback was undertaken prior to cognitive interviewing (CI). A mixed methods approach was chosen to include a quantitative analysis, using the CVI (Rodrigues, Adachi, Beattie, & MacDermid, 2017), with clarity of assessment of items assessed using qualitative CI (Egger-Rainer, 2019).

1.2. Research aims

This study aims to present a method of testing comprehension of a new self-reported measure evaluating the subjective experience of immersive technology in healthcare education. Furthermore, to propose a model that incorporates the theoretical perspectives outlined in the background with determinants related to immersive technology and learning.

1.3. Judgmental evidence- cognitive interviewing

CI as a method of evaluating sources for potential response error in survey design was developed in the 1980s through interdisciplinary work by methodologists and psychologists (G. Willis, 2005). CI requires a respondent to complete a questionnaire whilst an interview adds verbal information. Tourangeau (Council, 1984) proposed four cognitive operations that can be addressed: comprehension, recall, judgment, and response. CI can be used to inform item wording and provide richer insights than quantitative data alone, as respondent's thought processes are revealed as they respond to items. The multistep process can help identify misalignment of item intent and shed light on relationship of items to overall construct.

Semi-structured interviewing of respondents can be conducted using verbal probe (VP), think-aloud (TA), or combined techniques. VP involves spontaneous or scripted questions most commonly asked concurrently with respondents answering an item (Buers et al., 2014; G. Willis, 2005). Cognitive operations can be mapped on to scripted questions that cover item terminology, item understanding, item response and combining items for desired construct (Peterson, Peterson, & Powell, 2017). Scripting of probes and using a coding system has the advantage of interview consistency between different interviewers. By contrast, TA is with minimal prompting and interviewers listen to respondents articulating their conscious thought stream. Interviewer and respondent training on this method is recommended to avoid hesitation. There is no uniform method of conducting CI and some support a combined VP and TA approach to gain most insights, with TA preferencing to concept understanding and VP to exploring specific items (Priede & Farrall, 2011).

2. Methods

2.1. Design

This was a mixed methods study that employed semi-structured CIs using VP and TA qualitative-descriptive analysis of question meaning, applying Tourangeau's four cognitive operatives as a structure with question-and-answer model (Collins, 2003). Furthermore, CVI explored content validity of the ITEM construct.

2.2. Setting and participants

The study was conducted during November 2022 in a clinical teaching department at a regional hospital in the United Kingdom. Purposive sampling was undertaken to include 1 doctor and 2 medical students in each group that experienced immersive technology, prior to questionnaire responses and CI. Six female and 3 male English speakers took part in the study and the mean age was 24.

The nine participants were recruited via email and no incentives were given. Three immersive experiences were used, that were typical of Mixed Reality (Virtual reality and Augmented reality) and the target media for ITEM use. Allocation to one of three groups was done opportunistically: VR bystander cardiopulmonary resuscitation training, AR HoloLens holographic respiratory simulation, or 360-degree video in head mounted display of paediatric respiratory illness, to ensure sampling of 3 to each group (n = 9). The VR and AR experiences were selected to provide a medical education experience that were different in both technology and content, which ranged in time between 8 and 15 min in length. Five to 15 participants are recommended for this type of method, as sufficient to capture problems with survey design, and increasing sample sizes has diminishing returns on exposing new problems (Peterson et al., 2017; G. Willis, 2005). All interviews lasted up to 1 h in length with additional time spent consuming immersive media prior to interview.

2.3. Data collection

CIs were undertaken by 2 trained authors to reduce confirmation bias (Peterson et al., 2017; Woolley, Bowen, & Bowen, 2006). A Qualtrics survey that included VP of open and closed questions to each of the 40 questions forming the ITEM was used. Interviewers introduced the process to participants whereby verbalising conscious thought was encouraged and responses to VP questions were transcribed for all 40 items. A 4 question VP was designed that included: understanding of the measure item, repeating the question in the participants own words, key words in ITEM question were explored for further comprehension, alternative wording to any parts of question probed, and finally field notes for further qualitative analysis.

2.4. Study ethics

The study was approved at the regional hospital undergraduate medical education committee (reference number CJ-112022). All participants received written information and completed and signed a consent form.

2.5. Data analysis

Analytical technique of transcribing during interview for follow up coding and rapid analysis for confirmation of domain construct (Taylor, Henshall, Kenyon, Litchfield, & Greenfield, 2018). The team undertook a within-case analysis for each respondent's transcript and identified key phrases. Discrepancies of respondent's understanding from the intending meaning was probed with follow up questioning, for item development and rewording. Each item was matched to the pre-existing domain and charted in a matrix that had all respondents' answers. Authors met for an iterative process of coding and evaluation of domains for their degree of agreement with the intended construct. Furthermore, inductive/deductive hybrid thematic analysis was selected to synthesise the mixed methods data for theory generative retroduction (Proudfoot, 2022).

Data collected indicated minor revision to the wording of ITEM measures was needed. Saturation of question interpretation and intent deduction was apparent, so a second round of CI was not required.

There exists numerous methods for quantitative testing of CVI (Rodrigues et al., 2017). Item-CVI is calculated from the proportion of participants giving a high relevance rating. Acceptable CVI values for studies with 9 or more participants are at least 0.78 (Yusoff, 2019).

A trichotomous Likert scale was used, for clarity: 0 = *not understood,* 1 = *maybe understood* 3 = *understood;* and for relevance*:* 0 = survey not relevant to content, 2 = survey maybe relevant to content, 3 = relevant to content.

Table 2
Content validity domain scores for clarity for ITEM.

Domain	Average clarity score	Range
Immersion (9 questions)	0.98	0.0 to 1.0
Motivation (10 questions)	0.98	0.0 to 1.0
Cognitive Load (6 questions)	0.98	0.5 to 1.0
System Usability (10 questions)	0.94	0.0 to 1.0
Debrief (5 questions)	0.92	0.0 to 1.0
ITEM (all 40 questions)	0.96	0.0 to 1.0

3. Results

Following single round of CI the 40 questions forming the multidimensional ITEM were analysed and a number of interpretative and comprehension errors were present. Two questions required editing and 38 questions remained unchanged. Final ITEM is available in supplementary material.

3.1. Content validity results

The relevancy of the construct of questionnaire was calculated as 0.94 for the 9 participants, ranging from 0.5 to 1.00.

Domain related CVI are seen in Table 2, with the overall average CVI clarity score for 40 items being 0.96, ranging from 0.0 to 1.0. Thus, the ITEM was considered as being very clear.

Clarity scores for the three questions that performed lowest in the survey were in three domains. Immersion domain question 7, 'I wanted to learn more about the outcome following the activity', scored 0.78. System usability domain question 8, 'I found the technology cumbersome to use', scored 0.55. Debrief domain question 4, 'Was one or more of your performance domains explored', scored 0.78.

3.2. CI descriptive analysis

A matrix map was created (Fig. 3) with coded response to each item, which related to each domain of the ITEM for the response process.A hybrid deductive and inductive approach was purposively chosen with deductive reasoning adopting key themes that represent the subscale theories, which are represented as the central bubbles in Fig. 3. Coding of transcripts by authors identified numerous subthemes that related to the underlying theory and these are visually seen surrounding the key theme. Importantly subtheme overlap to other domains displayed a degree of relationship to other variables seen with dotted lines crossing domains. Progressing through the extraction of participant perspectives surrounding the underlying key concepts allows analysis of how they might support these concepts and to generate new models via abductive reasoning.

Participant interview comments support the analysis and are recorded as P with allocated anonymous numbering.

Comprehension was demonstrated with VP on understanding of question interpretation. Question 10 explores immersion with responses of "been in the environment simulated and not being disturbed by other external factors" (P1) or "simulation world feels like the only place you are in" (P9). Furthermore, the concept of fidelity or realism was expressed as "mimics as if you were there" (P4) and "what I see and perceive and very interesting and real" (P1). A detachment from external surroundings was described: "like you are focused on the actual scenario, felt what was in the scenario" (P3). These are examples of key factors when considering the degree of immersion experienced by someone.

Question 7 had the lower clarity score, of 0.78, and asked participants to consider the outcome beyond the activity. The word 'outcome' created

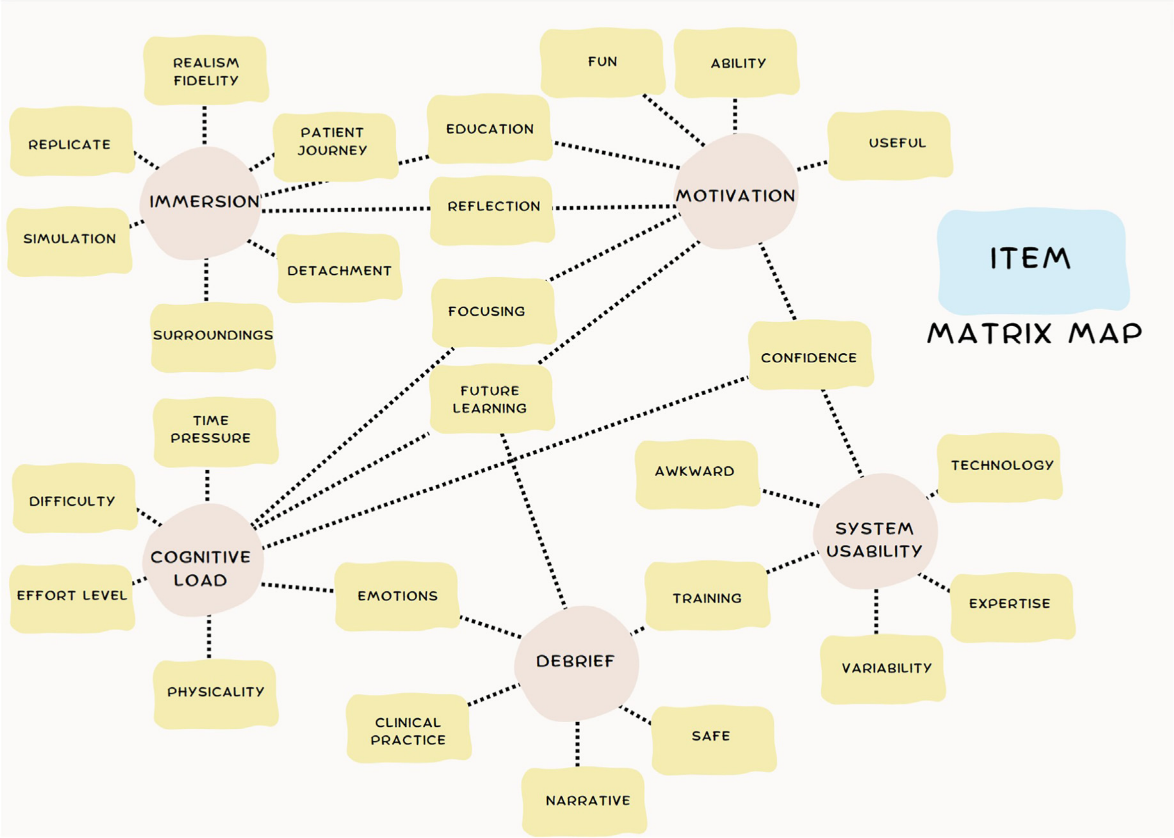

Fig. 3. Five domain matrix map with subcodes to main themes and linking between domains. Pink indicating Immersive Technology Evaluation Measure (ITEM) subscale. Yellow relates to subthemes. Lines indicate linking between themes. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

some interpretative confusion. For example, some interpreted this as the clinical outcome: “what his further treatment and whether he survived … realism” (P2), while others referred to what the user might have learnt: “split between outcome could be education, outcome of fake patient” (P8). These both hold value. The dichotomous interpretation could impact on the apparent self-reported immersion Likert score.

The intrinsic motivational state is explored by the AIMI. This is a tool for measuring an individual's intrinsic motivation to engage in a specific activity or task. It can be used to assess the factors that drive individuals to participate in these activities, such as their interest in the subject matter, their sense of personal accomplishment, and their enjoyment of the learning experience. The processes that promote learning in immersive technology were apparent in CI. Illustrative quotes include: “something captures my attention and to pursue more to find out … fun” (P2), “how this can essentially benefit me in my learning and progression in medicine” (P3), and “Improve to find out my weaknesses and get better” (P7).

Also, motivation is part driven by attention to the subject matter, “How much focus you're giving it compared to other things going on” (P6). Participants judged their understanding of motivation and the related items as a relationship to the activity and a surrogate for competence “sure of my abilities to do the right thing when encountering a similar situation” (P2), “confidence in my management” (P1).

The NASA Task Load Index (TLX) is a tool for measuring the subjective workload of individuals as they perform a task. It assesses six dimensions of workload: mental demand, physical demand, temporal demand, performance, effort, and frustration. The workload index provoked a different set of responses when compared to those from the AIEQ and AIMI. For example, participants commented: “was it taxing or difficult to current level of knowledge” (P4), “The amount of exertion it takes to carry it out” (P5) and “How much effort I need to put in, the amount of understanding I need to know the task” (P7). The activity is time pressured, that could differ from the immersive experience qualitative assessment of time passing, “Very hectic you feel the stress of time, need to complete it as soon as possible or consequences” (P6).

SUS is a tool for measuring and evaluating the usability of a system or technology. This domain explores the complexity and completeness the technology affords to the user. Learning in this setting assumes a degree of technical ability and the participants reflected on this: “Someone who has experience or training with using it” (P5), and “aspects not working together or fitting together in a way you would expect them to be” (P2). This triangulated with NASA TLX and AIMI in perception of confidence, however, in relation to the technology itself, “strong sense of accuracy of what I am doing and justify it” (P1) and “Believing in your own ability to do something” (P6).

SUS is also seen as a functional domain that reviews the technology integration and controls, for example, “Different aspects, how they complement each other, how they form a package, merged into one” (P7). Although SUS is a widely used instrument, question 8 was not fully understood. Four of the nine participants were unsure of the meaning of ‘cumbersome’. The comprehension error prompted several participants to suggest the word “awkward” (P2, P3, and P9) was more appropriate.

The final domain of debrief was evaluated in a 5 item newly created questionnaire based on PEARLS. This included an assessment of psychological safety, which participants justified as “not dangerous to myself and in health and learning environment that there is no judgement” (P3), and allowing for errors “free from judgment and able to make mistakes” (P4). NASA TLX related the activity to a statement of negative emotions, whereas the debrief tool was nonspecific to the negativity or positivity experienced, “emotions, excitement and awe was there” (P1), and “any difficult emotions you might experience with an unwell patient” (P4). In prompting a self-assessment of strategy in clinical practice, question 4, phrase ‘performance domains’ was ambiguous. Some participants were “not sure” (P2) and “did not understand” (P8), whereas others correctly identified the intended meaning: “Different areas that you took part in and how they were ranked compared to each other” (P5), and “different areas and facets of competence” (P9). Participants 5 and 7 suggested greater detail defining the concept, suggesting phrases like “.different areas facets of competence”, “competence in domains needs description.” This question was updated to reflect this.

Future learning by reflection is important to any educational activity and this shared concept is also measured in the AIMI, whilst in debrief there is structure to the lessons to take away, “learning points that are given to go home on and reflect on and focus future learning on” (P1) and “points to reflect on and use to improve future performance” (P4).

4. Discussion and conclusions

A cognitive interview is a method of questionnaire design and content analysis that can be used to understand how individuals process and interpret questions when completing a survey or questionnaire. In this study structured interviews were used as a technique that involves asking participants to think aloud as they completed a questionnaire, allowing researchers to identify any difficulties or ambiguities they encountered while answering the questions. Patterns of divergence of respondents' understanding from intended meanings were seen, which informed question modification (G. B. Willis & Artino, 2013).

The ITEM is the first multidomain instrument that can assess the user's immersion, perceived learning, and usability of immersive technology in the context of healthcare education. This study demonstrated high correlation in the content analysis to the 5 domains of the ITEM. In the qualitative analysis, a process of TA and VP, the participants reviewed different modalities of immersive media in healthcare education, and provided insights into their comprehension of questions as they responded accordingly to their perceived experience of the media (Council, 1984). There were 2 questions that introduced confusion by the language used to describe the phenomena of interest. Understanding of the language used for the target population of a questionnaire is important when establishing the validity (Schildmann et al., 2015). The analysis suggested relationships exist between some variables, such as motivation and confidence, which is not surprising given the subjectivity underlying responses, and less understood process of how we learn and the complexities of linking this to individual and external factors (Franz, Oberst, Peters, Berger, & Behrend, 2022).

The quantitative CVI scores support good comprehensibility and acceptability of the ITEM, but also identified the 2 questions requiring refinement. The high levels of comprehension were seen, with each domain achieving average clarity scores of above 0.92, and 0.96 overall. When asking participants about the relevance to the content of the items to their experience of the immersive technology, and whether they thought the questionnaire captured their experience, the participants found the ITEM highly relevant (0.94). CVI of 0.78 or higher is considered excellent (Rodrigues et al., 2017) and only one question score was below this. This provides evidence for the internal structure and final construct.

An advantages of a self-reporting measure is the relatively short administration time (Jeon et al., 2020), however, given the heterogeneity of immersive technology, the measure requires a degree of generalisability and ability to compare scores between technology modes. The ITEM is supported by prior validation studies on the IEQ (Jacobs, Foote, Joiner, & Williams, 2022; Jacobs & Maidwell-Smith, 2022; Jennett et al., 2008; Rigby, Brumby, & Gould, 2019), IMI (Alverson et al., 2004; Jacobs & Maidwell-Smith, 2022; Jacobs & Rigby, 2022; Sattar et al., 2019), SUS (Anderson et al., 2021; Berg & Steinsbekk, 2021; Issleib, Kromer, Pinnschmidt, Suss-Havemann, & Kubitz, 2021), and NASA TLX (Bric, Connolly, Kastenmeier, Goldblatt, & Gould, 2014; Chan et al., 2020; Gupta, Cecil, & Pirela-Cruz, 2018; Schoeb et al., 2020; Yu et al., 2021) in the field of immersive technology. A recent systematic review evaluating VR in education settings supported both SUS and NASA TLX use as existing operationalisations that have been applied in multiple contexts (Strojny & Dużmańska-Misiarczyk, 2023).

A common form of debriefing after simulation or simulation-based

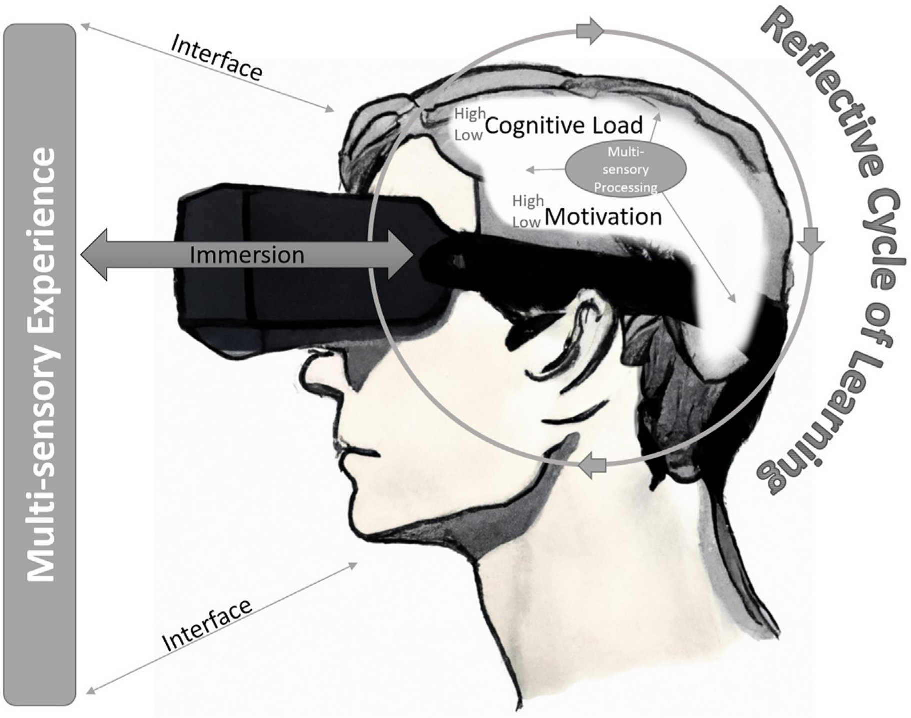

Fig. 4. Model of immersive technology in healthcare education (MITHE).

courses is to allow learners to reflect on the task under simulation and discuss with others. Self-debriefing can occur in immersive simulation design when a facilitator is not present (Cheng et al., 2020). Alternatively, virtual debriefings can occur with physical separation over web-based applications (Verkuyl et al., 2018). Strategies that facilitate learning through debrief are grounded in feedback and encouragement of reflection. The PEARLS debrief tool provided the foundation to the 5-question domain creation. This domain had the lowest score of clarity, and although, all questions were above a CVI of 0.78, rewording to aid understanding was judged to be required in one question. PEARLS is a commonly used debriefing tool (Eppich & Cheng, 2015) and in pilot studies has shown to lessen cognitive load (Meguerdichian et al., 2022), and this strengthens the understanding of the relationship between variables. However, there needs to be further validation studies in both immersive technology use and standard simulation to ensure this domain appropriately captures the important elements of debrief.

A theoretical framework of a model that helps visualise the relationships between concepts has been developed from work on immersive technology and learning by authors. Model of Immersive Technology in Healthcare Education (MITHE) explores from earlier published work with reflection being an act of thought and this construct is not limited to one theory as Fig. 4 highlights. MITHE contains elements of cognitivism, self determinism, and experiential theories.

This theoretical framework is not limited to a dual-process channel for sensory information.

It is a multi-sensory experience, and this will impact on the cognitive load. The arrows represent a two-way exchange, for example, the technology interface is the usability of the technology for an individual or the immersion experienced is related to the content and the individual. The cognitive factors of processing tasks are also depicted as a varying level (between high and low), as does the intrinsic reward mechanisms of motivation. Multi-sensory experience is controlled by the immersion and interface enables multi-sensory processing, which occurs in multiple cerebral areas, however, is limited by cognitive load and our motivational states. Finally, the experience is conceptualised with reference to prior learning and allows for interaction again with the media anchored through immersion and the interface. This promotes a repetition of experiences with adjustments of understanding and potential for mastery through practice.

4.1. Limitations

Bias exists with the selection of measures that forms the complete ITEM. Although, selected through a process of literature review and adapting for healthcare use in pilot studies. The process of CI is to support the content and how we might trust the accuracy of feedback participants will return. Numerous other measures exist that investigate the domains in ITEM and these remain available to study.

Several limitations existed in the methodology applied in CI, which are possible errors in comprehension of interviewer, interviewee, or both (Lenzner, Hadler, & Neuert, 2022). There can be a reactivity bias whereby responses are created just because they have been asked. Furthermore, the cognitive demand of interviews create can influence survey response and understanding (G. B. Willis & Artino, 2013). Thirdly, the CI can fail to detect the problems with the survey, perhaps if there were too few interviews to uncover all problems. Applying a framework and inquiring through a standardised method of design, as in this study, can help minimise these errors (Conrad & Blair, 2004).

This study did not extrapolate the performance scores of participants to infer or imply meaning, or even calculate the participants scores for ITEM (Kane, 2013). For example, it's not established in this study whether a high ITEM score reflects a better experience or technology assessment. A comprehensive approach to validation needs to consider this in different settings and educational outcomes given the simulation-based assessments are surrogates to real-world performance (David A. Cook & Hatala, 2016). Future study of exploratory factor analysis on the ITEM with a larger group of participants in these settings is recommended.

4.2. Conclusion

Through cognitive interviewing the response processes and content are better understood for a new measure that aims to evaluate immersive technology in healthcare education (ITEM). The instrument proved comprehensible and adds to the accumulating evidence to support the interpretation of the intended indirectly measurable concepts. Two questions were amended to improve questionnaire design. The ITEM development explored the evidence that supported a theoretical framework. Findings indicate the need for further work in assessing if the instrument has value for assessing a growing field of technology use in educating current and future healthcare professionals. In particular, more extensive application of the instrument to multiple different MR medium and media to establish the degree to which inferences are made on the operationalisations of the construct and attempt to evidence the correspondence of theory with reality.

Statements on open data and ethics

The participants identity was protected and personal information was not revealed in this study. Participants freely consented and informed of study withdrawal at any point. Written consent was provided for data to be used for publication, however, personal data is not available to be shared publicly. The study was approved by the undergraduate research office at Great Western Hospital, United Kingdom.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cexr.2023.100027.

References

Adams, N. E. (2015). Bloom's taxonomy of cognitive learning objectives. *Journal of the Medical Library Association, 103*(3), 152–153. https://doi.org/10.3163/1536-5050.103.3.010

Alverson, D. C., Saiki, S. M., Jacobs, J., Saland, L., Keep, M. F., Norenberg, J., … Caudell, T. P. (2004). Distributed interactive virtual environments for collaborative experiential learning and training independent of distance over Internet2. Jan 15-17. In *Paper presented at the 12th conference on medicine meets virtual reality* (Newport Beach, CA).

American Educational Research, A. (2018). *Standards for educational and psychological testing*. American Educational Research Association.

Andersen, S. A. W., Konge, L., Thomasen, C., & Sorensen, M. S. (2016). Retention of mastoidectomy skills after virtual reality simulation training. *JAMA Otolaryngology-Head & Neck Surgery, 142*(7), 635–640. https://doi.org/10.1001/jamaoto.2016.0454

Andersen, S. A. W., Mikkelsen, P. T., Sorensen, M. S., Andersen, S. A. W., Mikkelsen, P. T., & Sorensen, M. S. (2020). The effect of simulator-integrated tutoring for guidance in virtual reality simulation training. *Simulation In Healthcare-Journal Of The Society For Simulation In Healthcare, 15*(3), 147–153.

Anderson, M., Guido-Sanz, F., Diaz, D. A., Lok, B., Stuart, J., Akinnola, I., et al. (2021). Augmented reality in nurse practitioner education: Using a triage scenario to pilot technology usability and effectiveness. *Clinical Simulation in Nursing, 54*, 105–112. https://doi.org/10.1016/j.ecns.2021.01.006

Baxendale, B., Shinn, S., Munsch, C., & Ralph, N. (2020). Enhancing education, clinical practice and staff wellbeing. A national vision for the role of simulation and immersive learning technologies in health and care. *Retrieved from Simulation and immersive technologies*.

Berg, H., & Steinsbekk, A. (2021). The effect of self-practicing systematic clinical observations in a multiplayer, immersive, interactive virtual reality application versus physical equipment: A randomized controlled trial. *Advances in Health Sciences Education, 26*(2), 667–682. https://doi.org/10.1007/s10459-020-10019-6

Bric, J., Connolly, M., Kastenmeier, A., Goldblatt, M., & Gould, J. C. (2014). Proficiency training on a virtual reality robotic surgical skills curriculum. *Surgical Endoscopy and Other Interventional Techniques, 28*(12), 3343–3348. https://doi.org/10.1007/s00464-014-3624-5

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry, 189*(194), 4–7.

Buers, C., Triemstra, M., Bloemendal, E., Zwijnenberg, N. C., Hendriks, M., & Delnoij, D. M. J. (2014). The value of cognitive interviewing for optimizing a patient experience survey. *International Journal of Social Research Methodology, 17*(4), 325–340. https://doi.org/10.1080/13645579.2012.750830

Calvillo Gamez, E., & Cox, A. (2010). *Chapter 4 assessing the Core elements of the gaming experience* (pp. 47–71).

Chan, M., Uribe-Quevedo, A., Kapralos, B., Jaimes, N., Jenkin, M., & Kanev, K. (2020). A preliminary usability comparison of augmented and virtual reality user interactions for direct ophthalmoscopy. Aug 12-14. In *Paper presented at the IEEE 8th international conference on serious games and applications for health (SeGAH)* (Vancouver, CANADA).

Cheng, A., Kolbe, M., Grant, V., Eller, S., Hales, R., Symon, B., … Eppich, W. (2020). A practical guide to virtual debriefings: Communities of inquiry perspective. *Advances in Simulation, 5*(1), 18. https://doi.org/10.1186/s41077-020-00141-1

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research, 12*(3), 229–238. https://doi.org/10.1023/A:1023254226592

Conrad, F. G., & Blair, J. (2004). Data quality in cognitive interviews: The case of verbal reports. In *Methods for testing and evaluating survey questionnaires* (pp. 67–87).

Cook, D. A., Beckman, T. J., Thomas, K. G., & Thompson, W. G. (2009). Measuring motivational characteristics of courses: Applying keller's instructional materials motivation survey to a web-based course. *Academic Medicine, 84*(11), 1505–1509. https://doi.org/10.1097/ACM.0b013e3181baf56d

Cook, D. A., & Hatala, R. (2016). Validation of educational assessments: A primer for simulation and beyond. *Advances in Simulation, 1*(1), 31. https://doi.org/10.1186/s41077-016-0033-y

Council, N. R. (1984). *Cognitive aspects of survey methodology: Building a bridge between disciplines*. Washington, DC: The National Academies Press.

Egger-Rainer, A. (2019). Enhancing validity through cognitive interviewing. A methodological example using the Epilepsy Monitoring Unit Comfort Questionnaire. *Journal of Advanced Nursing, 75*(1), 224–233. https://doi.org/10.1111/jan.13867

Eppich, W., & Cheng, A. (2015). Promoting excellence and reflective learning in simulation (PEARLS): Development and rationale for a blended approach to health care simulation debriefing. *Simulation in Healthcare, 10*(2). Retrieved from https://journals.lww.com/simulationinhealthcare/Fulltext/2015/04000/Promoting_Excellence_and_Reflective_Learning_in.7.aspx.

Forrest, K., & McKimm, J. (2019). *Healthcare simulation at a glance*. Hoboken, NJ: John Wiley & Sons.

Franz, A., Oberst, S., Peters, H., Berger, R., & Behrend, R. (2022). How do medical students learn conceptual knowledge? High-, moderate- and low-utility learning techniques and perceived learning difficulties. *BMC Medical Education, 22*(1), 250. https://doi.org/10.1186/s12909-022-03283-0

George, O., Foster, J., Xia, Z., & Jacobs, C. (2023). Augmented reality in medical education: A mixed methods feasibility study. *CUREUS, 15*(3), Article e36927. https://doi.org/10.7759/cureus.36927

Gupta, A., Cecil, J., & Pirela-Cruz, M. (2018). A virtual reality enhanced cyber physical framework to support simulation based training of orthopedic surgical procedures. Aug 20-24. In *Paper presented at the 14th IEEE international conference on automation science and engineering*. Munich, GERMANY: IEEE CASE). Tech Univ Munich.

Hart, S. G. (2006). Nasa-task load index (NASA-TLX); 20 Years later. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting, 50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Health Education England (hee.nhs.UK)..

Henssen, D., van den Heuvel, L., De Jong, G., Vorstenbosch, M., Van Walsum, A. M. V., Van den Hurk, M. M., … Bartels, R. H. M. A. (2020). Neuroanatomy learning: Augmented reality vs. Cross-sections. *Anatomical Sciences Education, 13*(3), 350–362.

Issleib, M., Kromer, A., Pinnschmidt, H. O., Suss-Havemann, C., & Kubitz, J. C. (2021). Virtual reality as a teaching method for resuscitation training in undergraduate first year medical students: A randomized controlled trial. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, 29*(1). https://doi.org/10.1186/s13049-021-00836-y

Jacobs, C., Foote, G., Joiner, R., & Williams, M. (2022). A narrative review of immersive technology enhanced learning in healthcare education. *International Medical Education, 1*(2), 43–72. https://doi.org/10.3390/ime1020008

Jacobs, C., Foote, G., & Williams, M. (2022). Evaluating user experience with immersive technology in simulation-based education: A modified Delphi study with qualitative analysis. In *bioRxiv*.

Jacobs, C., & Maidwell-Smith, A. (2022). Learning from 360-degree film in healthcare simulation: A mixed methods pilot. *Journal of Visual Communication in Medicine*, 1–11. https://doi.org/10.1080/17453054.2022.2097059

Jacobs, C., & Rigby, J. M. (2022). Developing measures of immersion and motivation for learning technologies in healthcare simulation: A pilot study. *J Adv Med Educ Prof, 10*(3), 163–171. https://doi.org/10.30476/jamp.2022.95226.1632

Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., et al. (2008). Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies, 66*(9), 641–661. https://doi.org/10.1016/j.ijhcs.2008.04.004

Jeon, D.-W., Jung, D.-U., Oh, M., Moon, J.-J., Kim, S.-J., & Kim, Y.-S. (2020). Correlation between performance-based and interview-based cognitive measurements in patients with schizophrenia. *Psychiatry Investig, 17*(7), 695–701. https://doi.org/10.30773/pi.2020.0085

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Khalil, M. K., & Elkhider, I. A. (2016). Applying learning theories and instructional design models for effective instruction. *Advances in Physiology Education, 40*(2), 147–156. https://doi.org/10.1152/advan.00138.2015

Lateef, F. (2010). Simulation-based learning: Just like the real thing. *Journal of Emergencies, Trauma, and Shock, 3*(4). Retrieved from https://journals.lww.com/onlinejets/Fulltext/2010/03040/Simulation_based_learning__Just_like_the_real.9.aspx.

Lenzner, T., Hadler, P., & Neuert, C. (2022). An experimental test of the effectiveness of cognitive interviewing in pretesting questionnaires. *Quality and Quantity*. https://doi.org/10.1007/s11135-022-01489-4

Lia, H., Paulin, G., Yeo, C. T., Andrews, J., Yi, N., Haq, H., … Fichtinger, G. (2018). *HoloLens in suturing training. medical imaging 2018: image-guided procedures, robotic interventions* (p. 10576). AND MODELING.

Logeswaran, A., Munsch, C., Chong, Y. J., Ralph, N., & McCrossnan, J. (2021). The role of extended reality technology in healthcare education: Towards a learner-centred approach. *Future Healthc J, 8*(1), e79–e84. https://doi.org/10.7861/fhj.2020-0112

Makransky, G., & Petersen, G. B. (2021). The cognitive affective model of immersive learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educational Psychology Review, 33*(3), 937–958. https://doi.org/10.1007/s10648-020-09586-2

Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The cambridge handbook of multimedia learning* (2 ed., pp. 43–71). Cambridge: Cambridge University Press.

Meguerdichian, M., Bajaj, K., Ivanhoe, R., Lin, Y., Sloma, A., de Roche, A., … Walker, K. (2022). Impact of the PEARLS healthcare debriefing cognitive aid on facilitator cognitive load, workload, and debriefing quality: A pilot study. *Advances in Simulation, 7*(1), 40. https://doi.org/10.1186/s41077-022-00236-x

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Mills, B., Dyks, P., Hansen, S., Miles, A., Rankin, T., Hopper, L., et al. (2020). Virtual reality triage training can provide comparable simulation efficacy for paramedicine students compared to live simulation-based scenarios. *Prehospital Emergency Care, 24*(4), 525–536. Retrieved from https://www.tandfonline.com/doi/full/10.1080/10903127.2019.1676345.

Moro, C., Phelps, C., Redmond, P., Stromberga, Z., Moro, C., Phelps, C., … Stromberga, Z. (2021). HoloLens and mobile augmented reality in medical and health science education: A randomised controlled trial. *British Journal of Educational Technology, 52*(2), 680–694.

Oksenberg, L., & Kalton, G. (1991). New strategies for pretesting survey questions. *Journal of Official Statistics, 7*(3), 349.

Peterson, C. H., Peterson, N. A., & Powell, K. G. (2017). Cognitive interviewing for item development: Validity evidence based on content and response processes. *Measurement and Evaluation in Counseling and Development, 50*, 217–223. https://doi.org/10.1080/07481756.2017.1339564

Priede, C., & Farrall, S. (2011). Comparing results from different styles of cognitive interviewing: 'verbal probing' vs. 'thinking aloud'. *International Journal of Social Research Methodology, 14*(4), 271–287. https://doi.org/10.1080/13645579.2010.523187

Proudfoot, K. (2022). Inductive/deductive hybrid thematic analysis in mixed methods research. *Journal of Mixed Methods Research, 0*(0), Article 15586898221126816. https://doi.org/10.1177/15586898221126816

Rigby, J. M., Brumby, D. P., & Gould, S. J. J. (2019). Development of a questionnaire to measure immersion in video media: The film IEQ. *Association for Computing Machinery*, 35–46. https://doi.org/10.1145/3317697.3323361

Rodrigues, I. B., Adachi, J. D., Beattie, K. A., & MacDermid, J. C. (2017). Development and validation of a new tool to measure the facilitators, barriers and preferences to exercise in people with osteoporosis. *BMC Musculoskeletal Disorders, 18*(1), 540. https://doi.org/10.1186/s12891-017-1914-5

Ryan, G. V., Callaghan, S., Rafferty, A., Higgins, M. F., Mangina, E., & McAuliffe, F. (2022). Learning outcomes of immersive technologies in health care student education: Systematic review of the literature. *Journal of Medical Internet Research, 24*(2), Article e30082. https://doi.org/10.2196/30082

Samosorn, A. B., Gilbert, G. E., Bauman, E. B., Khine, J., McGonigle, D., Samosorn, A. B., … McGonigle, D. (2020). Teaching airway insertion skills to nursing faculty and students using virtual reality: A pilot study. *Clinical Simulation In Nursing, 39*, 18–26.

Sattar, M. U., Palaniappan, S., Lokman, A., Hassan, A., Shah, N., & Riaz, Z. (2019). Effects of Virtual Reality training on medical students' learning motivation and competency. *Pakistan Journal of Medical Sciences, 35*(3), 852–857. https://doi.org/10.12669/pjms.35.3.44

Schildmann, E. K., Groeneveld, E. I., Denzel, J., Brown, A., Bernhardt, F., Bailey, K., … Murtagh, F. E. M. (2015). Discovering the hidden benefits of cognitive interviewing in two languages: The first phase of a validation study of the Integrated Palliative care Outcome Scale. *Palliative Medicine, 30*(6), 599–610. https://doi.org/10.1177/0269216315608348

Schoeb, D. S., Schwarz, J., Hein, S., Schlager, D., Pohlmann, P. F., Frankenschmidt, A., … Miernik, A. (2020). Mixed reality for teaching catheter placement to medical students: A randomized single-blinded, prospective trial. *BMC Medical Education, 20*(1). https://doi.org/10.1186/s12909-020-02450-5

Strojny, P., & Dużmańska-Misiarczyk, N. (2023). Measuring the effectiveness of virtual training: A systematic review. *Computers & Education: X Reality, 2*, Article 100006. https://doi.org/10.1016/j.cexr.2022.100006

Tang, Y. M., Chau, K. Y., Kwok, A. P. K., Zhu, T., & Ma, X. (2022). A systematic review of immersive technology applications for medical practice and education - trends, application areas, recipients, teaching contents, evaluation methods, and performance. *Educational Research Review, 35*, Article 100429. https://doi.org/10.1016/j.edurev.2021.100429

Taylor, B., Henshall, C., Kenyon, S., Litchfield, I., & Greenfield, S. (2018). Can rapid approaches to qualitative analysis deliver timely, valid findings to clinical leaders? A mixed methods study comparing rapid and thematic analysis. *BMJ Open, 8*(10), Article e019993. https://doi.org/10.1136/bmjopen-2017-019993

Verkuyl, M., Lapum, J. L., Hughes, M., McCulloch, T., Liu, L., Mastrilli, P., … Betts, L. (2018). Virtual gaming simulation: Exploring self-debriefing, virtual debriefing, and in-person debriefing. *CLINICAL SIMULATION IN NURSING, 20*, 7–14.

Willis, G. (2005). *Cognitive interviewing*. https://doi.org/10.4135/9781412983655

Willis, G. B., & Artino, A. R., Jr. (2013). What do our respondents think we're asking? Using cognitive interviewing to improve medical education surveys. *Journal of Graduate Medical Education, 5*(3), 353–356. https://doi.org/10.4300/JGME-D-13-00154.1

Woolley, M. E., Bowen, G. L., & Bowen, N. K. (2006). The development and evaluation of procedures to assess child self-report item validity educational and psychological measurement. *Educational and Psychological Measurement, 66*(4), 687–700. https://doi.org/10.1177/0013164405282467

Yu, P., Pan, J. J., Wang, Z. X., Shen, Y., Wang, L. L., Li, J. L., et al. (2021). Cognitive load/flow and performance in virtual reality simulation training of laparoscopic surgery. Mar 27-Apr 03. In *Paper presented at the 28th IEEE conference on virtual reality and 3D user interfaces*. Electr Network: IEEE VR).

Yusoff, M. S. B. (2019). ABC of content validation and content validity index calculation. *Education in Medicine Journal, 11*(2), 49–54. https://doi.org/10.21315/eimj2019.11.2.6