Evaluating user experience with immersive technology in simulation-based education: a modified Delphi study with qualitative analysis

Chris Jacobs¹
Georgia Foote¹
Michael Williams²

¹ University of Bath, Claverton Down, Bath, UK
² Queen’s University of Belfast, University Road, Belfast, UK

Chris Jacobs corresponding author
cj511@bath.ac.uk
Abstract

Background

Immersive technology is becoming more widespread in simulation-based medical education with applications that both supplement and replace traditional teaching methods. There is a lack of validated measures that capture user experience to inform of the technology utility. We aimed to establish a consensus of items and domains that different simulation experts would include in a measure for immersive technology use.

Methods

A 3-stage modified Delphi using online software was conducted to support the conceptual framework for the proposed measure. The first round was informed by prior work on immersive technology in simulation. In the first round, participants were asked to describe what we could measure in simulation-based education and technology. Thematic analysis generated key themes that were presented to the participants in the second round. Ranking of importance in round 2 was determined by mean rank scores. The final round was an online meeting for final consensus discussion and most important domains by experts were considered.

Results

A total of 16 simulation experts participated in the study. A consensus was reached on the ideal measure in immersive technology simulation that would be a user questionnaire and domains of interest would be: what was learnt, the degree of immersion experienced,
fidelity provided, debrief, psychological safety and patient safety. No consensus was reached with the barriers that this technology introduces in education.

Conclusions

There is varied opinion on what we should prioritise in measuring the experience in simulation practice. Importantly, this study identified key areas that aids our understanding on how we can measure new technology in educational settings. Synthesising these results in to a multidomain instrument requires a systematic approach to testing in future research.
Background

Teaching through simulation is well established in medical disciplines. The pedagogy underlying simulation is based on deliberate practice through recreation of clinical scenarios, irrespective of the enabling technology. An educational experience is constructed with elements that resemble relevant environments, with a greater or lesser degree of fidelity, and match the functional task with the learner’s engagement. For example, we might describe a high-fidelity manikin and clinically accurate context, for training for students in basic life support.

In recent decades there are examples of technological innovation enhancing educational effectiveness through increased immersion, associated with improved motivational states. There are many technological modalities of simulation, summarised in table 1.

Table 1. Simulation modalities with mixed reality methods. AR- Augmented reality, VR- Virtual Reality, 360- 360-degree video, CAVE- Cave Automatic Virtual Environments, VP- Virtual patient. Modified from Forrest and McKimm (5)

<table>
<thead>
<tr>
<th>Simulation modality</th>
<th>Mixed reality modalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulated patient</td>
<td>AR, VR, 360, CAVE, VP</td>
</tr>
<tr>
<td>Part task trainer</td>
<td>AR, VR, 360, and CAVE</td>
</tr>
<tr>
<td>Virtual learning environment</td>
<td>VP</td>
</tr>
</tbody>
</table>
The Simulation Process provides an overarching structure to simulation-based education (SBE) (figure 1). Immersive technology integration in SBE requires consideration of all levels of the process, as no technological modality is inherently superior, and each possesses different strengths and weaknesses relevant for different scenarios.

The reality-virtuality continuum describes a range of simulation environments, from real world to virtual world. Within the continuum, the sensory stimuli can be a mix of real and virtual objects. For example, Mixed Reality (MR) is an overarching heading that includes augmentation of real-life (AR), and virtual environments in a head-mounted display (VR).(6) The extent of real-world disconnect and suspension of disbelief is dependent on the technology. Coherence reflects the authenticity or fidelity with which the technology matches to the real-world, and again there is a spectrum of coherence.(7)
There is ever growing evidence that immersive technology benefits healthcare students and qualified health professionals in simulation practice. Studies investigating the effectiveness of differing technologies in varied medical specialties is increasing year on year, with research in both VR and AR doubling in the period of 2020-2021. In a systematic review on immersive technology in healthcare education, over 50 methods were described of measuring healthcare practice, which could be broadly separated into: cognitive objective, procedural skills objective, or subjective behavioural. Capturing this information helps educators judge the performance of participants and the utility of the technology. The quality of the assessment tool used is essential for evaluating how potentially generalisable the results are. However, shortcomings in the validity and evidence supporting the use of many assessment tools in SBE exist. The Delphi technique is widely accepted as an effective group-based exercise, which aims to integrate multiple interpretations and opinions on a given topic. It has 4 defining characteristics: group anonymity, repeating questions (iteration), feedback to participants, and statistical description of responses. The Delphi researchers orchestrate focal issues surrounding the mode of delivery, the threshold for consensus and encouragement of responses. Delphi techniques supplement analyses of published literature on a topic in preparing a comprehensive framework. The overarching aim of this project is to create a de novo brief, valid and reliable multi-dimensional measure that will inform educators and learners on the users’ experiences of technology in simulation practice. The domains that could be considered important in
quantifying a user’s experience are yet not fully understood. The Delphi approach allows for a diverse number of experts in the field of SBE to be brought together to address this.

Materials and Methods

The study was funded by Health Education England (HEE) Southwest Simulation Network as part of a 2-year project collaboration with University of Bath and simulation network. This study will establish a consensus for generating a new measure’s theoretical and conceptual framework, that can be further examined with cognitive interviewing and quantitative exploration in future work.

Modified Delphi method

A steering group consisting of authors (CJ, GF, and MW) guided three-stages of convergent opinion and consensus building(18) between May and August 2022. The conventional Delphi method uses iterative rounds of data collection from groups of experts, with subsequent feedback on responses, with the opportunity for them to rethink their response until consensus is met.(19) Online questionnaire collection facilitates this process for response collection and feedback, in short time frames.(20)

A hybrid of the conventional method, a modified Delphi method was adopted, which collects initial participant opinions online, followed by in-person group meetings for final consensus discussion.(15, 21) An a priori response rate was set at 70% for Delphi round 2,
which was deemed as minimally acceptable to minimise bias. A lower response to the online meeting was deemed acceptable as it required greater commitment, and drop-out linked to additional demands of attending. Figure 2 flow chart illustrates the Delphi process, and a reporting checklist is provided in supplementary material.

Figure 2- Evaluating user experience with immersive technology in simulation-based education Modified Delphi process flow chart

Participants

The target population for this study will be individuals engaged in SBE. Stake holders included doctors, technicians, managers, and administrators involved in simulation for healthcare education (table 2). Purposive sampling from a targeted group allowed creation of a group of experts, collectively offering a range of experience in healthcare simulation. It was the aim to recruit 10 to 20, which has been described as the panel size to allow meaningful statistical analysis. Recruitment was via a Southwest simulation network email list as a single advert.

Table 2. Eligibility criteria- key inclusion and exclusion criteria.
### Key inclusion criteria

<table>
<thead>
<tr>
<th>Experience with simulation in healthcare education setting in the Southwest of England</th>
</tr>
</thead>
<tbody>
<tr>
<td>Able to access the internet (Study sign-up and intervention could only be accessed through the internet).</td>
</tr>
<tr>
<td>Willing to participate in 2 rounds of questionnaires AND/OR willing participate in a virtual meeting.</td>
</tr>
</tbody>
</table>

### Data collection and analysis

**Delphi Round 1**

This was an online questionnaire. Questionnaire design and distribution was done using Jisc Online surveys ([https://www.onlinesurveys.ac.uk](https://www.onlinesurveys.ac.uk)) accessed September 2022, software which meets data protection regulations. Participants’ information and consent were embedded in the survey prior to questions. The Delphi study was approved by Health Research Authority (research ethics committee reference 22/PR/0339 and integrated research application system project ID 312830). Research sponsor was the Great Western Hospital NHS Foundation Trust (UK).

Participants’ age and roles were collected to inform authors of range of their simulation experience. Otherwise, question development was drawn from 2 sources: firstly, a qualitative study exploring themes of participant experience in Virtual Reality within an
education setting, and secondly, a systematic review was undertaken on immersive technology in healthcare education covering literature from 2002 and 2022. In the systematic review, data from 246 papers on learning theory and on measures adopted in simulation practice with immersive technology were analysed. Outcomes from both these sources were discussed between authors CJ and MW to inform questionnaire design and participant information sheets. The questions focused on the modalities of learning in SBE, how this might be measured, and what potential barriers to use of emergent technology.

Round 1 provided qualitative data: panelists responded with free text answers in the form of a non-prioritised list. Thematic analysis was conducted to these open-ended questions with a 6-step process. Inductive reasoning guided the analysis of responses to identify key patterns and relationships. Generation of initial codes was done independently by CJ and GF. Further interpretative analysis was undertaken by CJ and GF, specifically repeated conversations and code sorting to identify overarching themes, that were eventually defined in short phrases or words that unified the narrative of responses. A thematic map was created as a pictorial summary for feedback to the expert panelists.

**Delphi Round 2**

This was an online questionnaire. Those participating in round 1 were invited to respond to round 2, which included the same questions with the newly identified and defined key themes resulting from round 1 as options for ranking items in importance from highest to lowest using a check-list response. Means, median and interquartile range (IQR) were calculated using StatsDirect version 3.3.5 for the ordinal round 2 responses. Statistical significance of data was determined as p <0.05, assessed using non-parametric Friedman’s
test and weighted Kappa to describe strength of agreement between experts ranking.(28) Kappa values were treated as 0.00 to 0.10 conferring poor agreement, 0.11 to 0.20 conferring slight agreement, 0.21 to 0.40 conferring fair agreement, 0.41 to 0.60 conferring moderate agreement, and 0.61 to 0.80 conferring substantial agreement.(29) Mean rankings were complemented with a boxplot visual summary for each question(15), and overarching thematic maps from Round 1 were distributed to participants prior to round 3.

**Delphi Round 3**
The online meeting was held using Zoom. The meeting was chaired by CJ and field notes collated by GF. Revealing of majority opinions prior to the final meeting enabled experts to reflect on group thinking with potential disagreement amongst experts during Round 2 seen, although, consensus was the ultimate aim. In this round sharing of opinions was promoted, and a discussion facilitated clarification of any misunderstandings (30). Chairing of a meeting promotes encouragement of those attending and avoiding dominance of voices.(31) An effort was made to have everyone contribute to the conversations. Consensus criteria, set before round 3, for identification of the top 4 themes from responses to questions 1 and 2 was set at 75%, and for question 4, the top single method of measuring was set at 75% of those attending the meeting(32) with opinions collected using the Zoom polling tool.

**Results**
There were 16 participants recruited from an email sent to 350 on a mailing list. True response rate in relation to inclusion and exclusion criteria could not be calculated. Table 3 shows the demography of the group in this study.

Table 3. Age and roles of participants. SBE- Simulation based education.

<table>
<thead>
<tr>
<th>SBE role</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outreach simulation co-ordinator</td>
<td>33</td>
</tr>
<tr>
<td>Clinical Teaching Fellow</td>
<td>26</td>
</tr>
<tr>
<td>Simulation Medical Lead</td>
<td>37</td>
</tr>
<tr>
<td>Programme Lead for Simulation &amp; TEL for HEE southeast</td>
<td>48</td>
</tr>
<tr>
<td>Paediatric Simulation fellow at University Hospitals Bristol and Weston NHS Foundation Trust</td>
<td>33</td>
</tr>
<tr>
<td>Nurse Education (university)</td>
<td>46</td>
</tr>
<tr>
<td>Simulation Facilitator</td>
<td>33</td>
</tr>
<tr>
<td>Primary Care Simulation Fellow</td>
<td>38</td>
</tr>
<tr>
<td>Clinical Skills Tutor to final Year Medical Students</td>
<td>53</td>
</tr>
<tr>
<td>Resuscitation &amp; Simulation Officer</td>
<td>28</td>
</tr>
<tr>
<td>GP tutor</td>
<td>33</td>
</tr>
<tr>
<td>Previous clinical teaching fellow</td>
<td>32</td>
</tr>
</tbody>
</table>
Clinical teaching fellow | 30
---|---
Senior Technical Instructor | 44
Simulation and Clinical Skills Technician | 46

**Qualitative Round 1**

Questions asked and key themes emerging from round 1 are seen presented in Figure 3 below.

Figure 3. Thematic map with the 4 questions abbreviated. Arrows indicate a 2-way link of response between questions. Dotted line indicates association of responses. Red line represents barriers.

**Quantitative Round 2**

Thirteen panelists from initial 16 recruited (81%) responded to round 2.

The 4 questions were individually analysed. Figure 4 panel A-D are boxplots with medians and IQR for each question.
Question 1- List up to 5 factors that you think are important in learning in simulation?

Table 4. Descriptive results for question 1 with mean ranking.

<table>
<thead>
<tr>
<th>Themes</th>
<th>Engagement</th>
<th>Psychological safety</th>
<th>Debrief</th>
<th>Realism/Fidelity</th>
<th>Teamwork</th>
<th>Repetition</th>
<th>Learning objectives defined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid data</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Missing data</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mean</td>
<td>2.76</td>
<td>3.23</td>
<td>2.85</td>
<td>3.39</td>
<td>5.62</td>
<td>5.00</td>
<td>4.15</td>
</tr>
<tr>
<td>Upper quartile</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Median</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Lower quartile</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>IQR</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>
Top four ranked responses in order: Engagement/immersion (mean rank 2.76), Debrief (mean rank 2.85), psychological safety (mean rank 3.23), and realism/fidelity (mean rank 3.39). Table 4 summarises the descriptive data for question 1.

Friedman's test indicated rank patterns ($p < 0.001$) with a degree of difference in ranking.

Weighted Kappa 0.21 ($p < 0.001$) indicating fair agreement in rankings between participants.

**Question 2- List what you think should be measured**

Table 5. Descriptive results for question 2 with mean ranking

<table>
<thead>
<tr>
<th>Themes</th>
<th>What was learnt</th>
<th>User confidence</th>
<th>User Immersion</th>
<th>Cost effectiveness</th>
<th>Demonstrate clinical skills</th>
<th>Benefits of technology</th>
<th>Technology ease of use</th>
<th>User stress levels</th>
<th>Patient Safety</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid data</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Missing data</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mean</td>
<td>2.54</td>
<td>3.15</td>
<td>3.69</td>
<td>5.54</td>
<td>4.38</td>
<td>4.92</td>
<td>5.38</td>
<td>5.38</td>
<td>3.92</td>
</tr>
<tr>
<td>Upper quartile</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Median</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>
Top four ranked responses in order: What was learnt (mean rank 2.54), user confidence (mean rank 3.15), user immersion (mean rank 3.69), and patient safety (mean rank 3.92).

Table 5 summarises the descriptive data for question 2.

Friedman’s test indicated rank patterns ($p < 0.001$) with a degree of difference in ranking.

Weighted Kappa 0.22 ($p=0.0003$) indicating a fair agreement in rankings between participants.

**Question 3- What measurement methods could you use?**

Table 6. Descriptive results for question 3 with mean ranking

<table>
<thead>
<tr>
<th>Themes</th>
<th>Performance metrics (test of skills)</th>
<th>Participant questionnaire of user experience</th>
<th>Test of Knowledge</th>
<th>Quality improvement outcomes</th>
<th>Adverse event reports</th>
<th>Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>----------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Valid data</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Missing data</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mean</td>
<td>2.92</td>
<td>1.77</td>
<td>3.31</td>
<td>2.69</td>
<td>4.69</td>
<td>5.62</td>
</tr>
<tr>
<td>Upper quartile</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Median</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Lower quartile</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Interquartile range</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Top four ranked responses in order: Participant questionnaire of user experience (mean rank 1.77), quality improvement outcomes (mean rank 2.69), performance metrics (mean rank 2.92), and test of knowledge (mean rank 3.31). Table 6 summarises the descriptive data for question 3.

Friedman’s test indicated rank patterns ($p < 0.001$) with a degree of difference in ranking. Weighted Kappa 0.22 ($p < 0.001$) indicating a fair agreement in rankings between participants.
Question 4- List the barriers you consider in "measuring immersive technology" in simulation and medical education

Table 7. Descriptive results for question 4 with mean ranking

<table>
<thead>
<tr>
<th>Themes</th>
<th>Real world improvements</th>
<th>Validity of measures</th>
<th>Fidelity</th>
<th>Accessibility</th>
<th>Cost</th>
<th>Technical difficulties</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid data</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>Missing data</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mean</td>
<td>3.58</td>
<td>3.67</td>
<td>4.25</td>
<td>3.75</td>
<td>4.25</td>
<td>3.83</td>
<td>3.67</td>
</tr>
<tr>
<td>Upper quartile</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Median</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Lower quartile</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Interquartile range</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>
Top four ranked responses in order: Real world improvements not measurable (mean rank 3.58), validity of measures (mean rank 3.67), cost (mean rank 3.67), and technical difficulties (mean rank 3.83). Table 7 summarises the descriptive data for question 4.

Friedman’s test indicated no significant rank patterns (p=0.97).

Figure 4. Boxplot of theme rankings. A- Question 1, B-Question 2, C-Question 3, and D-Question 4. Key, LQ-lower quartile, UQ- Upper quartile, cross marks mean, rhombus marks median.

Delphi Round 3

All participants of both rounds 1 and 2 were invited to join a virtual meeting hosted on Zoom. Five participants (36%) attended: 2 doctors, 2 nurses, and 1 simulation technician. There was a short presentation with a brief recap on literature surrounding immersive technology in medical education and each question was discussed in turn with results from round 2 available to prompt discussion. Table 8 shows the consensus reached for each of the 4 questions. The lack of spread to data in question 4 meant it was not appropriate for a consensus as each held equal weight.
<table>
<thead>
<tr>
<th>Question</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>List up to 5 factors that you think are important in learning in simulation?</td>
<td>100% agreement with: Engagement/Immersion, Psychological safety, Debrief, Realism/Fidelity</td>
</tr>
<tr>
<td>List what you think should be measured</td>
<td>60% agreement with: What was learnt, User confidence, User Immersion, Patient Safety</td>
</tr>
<tr>
<td></td>
<td>100% agreement if Cost effectiveness is added to above.</td>
</tr>
<tr>
<td>What measurement methods could you use?</td>
<td>100% agreement with: Participant questionnaire of user experience</td>
</tr>
<tr>
<td>List the barriers you consider in “measuring immersive technology” in simulation and medical education</td>
<td>No consensus reached</td>
</tr>
</tbody>
</table>
Field notes taken during the final meeting created discussion points with each question presented. Although, no further thematic analysis was undertaken, selected quotations highlighted the expert opinion on the way we might learn from immersive technology enhanced learning (TEL) and how we might measure it. Complete field notes are available in supplementary material.

Participant J (doctor) in response to question 1.

“I think, fidelity is high up – with enjoyment and immersion, then debrief”

Following this statement, participant DM (simulation technician) echoed J thoughts.

“engagement is instantaneous feedback about simulation...then can consolidate learning with debrief”

Participant SK (doctor) reflecting on question 2.

“...simulation as a concept can be a challenge...cost effectiveness is hard to measure but is very important to collect data on”

Following this participant J responded to SK in question 2.

“This tech is novel and is useful compared to standard methods – so immersion and testing how useful the technology is important”
Expert consensus was quickly reached with question 3 as acceptability on measuring user experience was the most important, although it was acknowledged that this might not capture all relevant factors.

“yes and no – good as it is easy to ask what they think – but whether the tech has made a difference after the session could be measured... difficult to measure maybe best measure is quality improvement and adverse events”

Discussion

Key aspects of relevant domains on measuring user experience of immersive technology in simulation were identified using the modified Delphi method in this study, with a consistency of responses evident across a heterogeneous groups of experts. It was agreed that a measure ideally would collect information on: what was learnt, the degree of immersion experienced, fidelity provided, debrief, psychological safety and patient safety. Additionally, a technological assessment should include a cost-effectiveness evaluation.

There was 100% consensus in round 3 that a psychometric measure of user experience using a participant questionnaire would be the most suitable format to explore use of immersive technology in SBE.

There are numerous measures, subjective and objective, in existence for assessing a particular immersive technology in a certain situation.(33) These are often created for the purpose of the individual study in which they are used, without any argument for their validity.(13) Furthermore, there is a risk that instruments used to collect data become obsolete as emergent technology surpasses existent training methods. An international
Delphi study was conducted to ensure researchers shared definitions of and terminology about instrument properties. (34) This study clarified what was meant by key measurement properties: reliability, validity, responsiveness and interpretability. These properties need to be considered when any outcome measurement is tested. The findings of a robust literature review supported by this Delphi study have led to a conceptual model of how the utility of immersive technology in healthcare education can be measured, indeed what to measure. This will inform development of a new evidence-based measure, which can then undergo field testing with various technologies and settings, that will enable the properties to be tested and the measure refined. The process of validity is not a single task, but testing hypotheses in a continuous process to see whether scores are consistent with the model intended. (35)

Although, various methods of Delphi are possible, a qualitative first round is useful as it allows experts to contribute ideas beyond current established or published knowledge. Reliability and validity of the Delphi study may be improved if the initial experts create the items for consensus reaching. (36) In this study, a relevant systematic review assisted in guiding development of questions to ask the experts.

Nasa et al. (37) described significant variation in how Delphi studies are conducted and reported, which can lead to uncertainty to the conclusions. Using a 9-point qualitative evaluation for a Delphi quality assessment, this study would score 8 out of 9, missing a full score only as group stability could not be evidenced. Transparency of reporting in this study is consistent with all areas in the Conducting and Reporting Delphi Studies (CREDES) checklist (39) and seen in supplementary material.
Sixteen participated in round 1 of a heterogeneous panel and although, no minimum number for Delphi method exists, 15 has been suggested as minimum group size.\(^{(37)}\) The threshold for consensus was achieved for all measure related questions, which was set \textit{a priori}. The number of rounds of Delphi is dependent on the method and the intended outcomes. In another modified-Delphi process, exploring learning and assessment in healthcare education, 2 rounds were used to create a consensus.\(^{(38)}\) Delphi response rate ideally should not fall below 70\% on subsequent rounds\(^{(22)}\): this standard was achieved in this study. Quantitative analysis was not extensive in this study due to methodology to summarise items with thematic analysis. This can be considered as a limitation to the study as Kappa values indicated fair agreement only and subsequent rounds were not conducted to seek convergence. Similarly statistical stability could not be tested for the strength of the item responses.\(^{(39)}\) Final consensus criteria were set by a third round as an in-person virtual meeting, whereby descriptive statistics of mean ranks complemented by visual box plots to indicate spread, and a thematic map, were available to support the discussion. Thematic analysis may introduce bias, potentially inherent in individual’s interpretations.\(^{(40)}\) The primary investigator had a medical educational background, and his and the co-author’s reflexive practice, and being conscious of this in the analysis, aimed to minimise impact of bias of personal experiences. Any measure developed from this process should be tested in different settings, which will reassure with regard to its validity.

The Delphi is not a psychometric instrument in itself\(^{(15)}\) but a practical method for gauging a group-based judgment, more useful for complex concepts and decisions. It can be argued
that there is not a preferred Delphi method(41), however, the process of a groupthink interaction generates ideas and aids the conceptual framework that this study addressed.

Conclusion

This study explained the process of sourcing opinion from a varied cohort of experts to determine how the experience of a learner in healthcare practice using immersive technology might be measured. Synthesising these results into a multidomain instrument requires a systematic approach to testing in future research. There is variation between educationalists in what is regarded as important in learning from this technology, however, it is important to establish relevant experts’ opinions before designing a user experience outcome measure, and this study highlighted a consensus on the topic.

REFERENCES


11. Jacobs C. Immersive technology in healthcare education: a scoping review


39. Nasa P, Jain R, Juneja D. Delphi methodology in healthcare research: How to decide its appropriateness. 2021(2222-0682 (Print)).


Needs Analysis
What are the specific learning outcomes for the intended learners?

Scenario Development
Detailed plan with outline of scenario, logistics, and feasibility testing

Curriculum integration
How does the simulation fit and integrate with existing and future curriculum design

Simulation Practice
Simulation centre and technology. How the simulation facilitates technical and non-technical skills acquisition

Evaluation
Debrief for participants. Principles of assessment and how to measure performance and experience.
Modified Delphi Process of rounds flow chart

0. Existing literature - systematic review
   - Pilot studies on immersive technology
   - Experts recruited
   - Online survey creation and distributed

1. Items defined by group
   - Thematic analysis of responses

2. Panel rank most important from items generated from round 1.
   - Analysis descriptive statistics

3. Consensus in person meeting.
Figure