Computers & Education: X Reality 2 (2023) 100025

Contents lists available at ScienceDirect

Computers & Education: X Reality

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

Development and testing of a scale for examining factors affecting the learning experience in the Metaverse

Emmanuel Fokides

University of the Aegean, Department of Primary Education, 1 Dimokratias str., 85132, Rhodes, Greece

ARTICLE INFO

Keywords:
Immersive virtual environment
Learning experience
Metaverse
Scale

ABSTRACT

The Metaverse, as it gradually becomes a reality, offers great possibilities for education, providing a new dimension of engagement, interaction, and experiences for students and educators alike. Yet, researchers lack tools that would allow them to simultaneously examine a satisfactory number of factors that shape one's experience when engaged in educational applications in the Metaverse. The study at hand is an attempt to fill this gap, as it reports the steps followed for the development and testing of the Metaverse Learning Experience Scale. The statistical analyses of data coming from 462 university students who participated in its testing, established the scale's validity and reliability. Its final version can capture the views of users using forty-three items, examining ten factors (namely, perceived quality of the virtual environment's graphics, perceived cognitive load, perceived ease of use/control of the virtual environment, immersion/presence, perceived feedback and content quality, perceived degree of interaction, motivation to learn and use the virtual environment, perceived usefulness/knowledge gains, simulator sickness, and positive feelings. The scale's factorial structure together with the implications it has for research and practice are also discussed.

1. Introduction-the Metaverse

A catchword circling the Internet in recent years is "Metaverse." The term was initially associated with science fiction; yet, advances in technology, allowed its materialization, although not fully so far. Generally speaking, it describes networks of virtual worlds where users interact with each other and digital objects in shared environments. Ball (2021) defined it as: " … a massively scaled and interoperable network of real-time rendered 3D virtual worlds and environments which can be experienced synchronously and persistently by an effectively unlimited number of users with an individual sense of presence … " Kye, Han, Kim, Park, and Jo (2021) identified four types of Metaverse technologies, namely Augmented reality (AR), Virtual Reality (VR), Lifelogging, and Mirror Worlds, depending on two axes (simulation vs augmentation and intimate vs external). There is also mixed reality (MR) which falls somewhere between AR and VR, combining virtual and real worlds, and extended reality (xR), an umbrella term for present and future VR, AR, and MR technologies. This probably led Lee et al. (2021) to define the Metaverse as a virtual environment that blends the physical and digital, merging Web technologies and xR.

Evidently, there is no consensus on the Metaverse's definition. Then again, because of the demand for highly immersive experiences and high levels of interaction and engagement, people mostly associate it with 3D virtual environments and the use of AR and VR headsets (Doug, 2020). Moreover, it is questionable whether computers, smartphones, and tablets can offer high levels of such experiences. In light of the above, in this study, the Metaverse is defined as the sum of 3D virtual spaces in which users can experience high levels of immersion, engagement, and interactivity through the use of xR headsets/devices.

Nevertheless, setting aside the problems in defining it, experts believe that the Metaverse can connect individuals from all over the world; thus, it can function as a central hub for various activities, including but not limited to gaming, socializing, education, and commerce (Ball, 2021). As a result, it can change how people access information and communicate. Though society has transitioned with the advent of remote workplaces, distance learning, and telemedicine, the Metaverse will expand the boundaries further, by bringing the workplace to the employee, the classroom to students, healthcare to patients, and shopping to consumers via immersive virtual worlds.

As far as education is concerned, the Metaverse offers a new dimension of engagement, interaction, and experiences for both students and teachers. It encompasses a wide range of tools and platforms, including virtual classrooms, labs, and field trips, which allow students to attend lectures, conduct experiments, and explore historical sites, scientific labs,

E-mail address: fokides@aegean.gr.

https://doi.org/10.1016/j.cexr.2023.100025
Received 12 March 2023; Received in revised form 23 April 2023; Accepted 3 May 2023

and natural habitats. Additionally, it functions as a platform for virtual language learning, educational games, virtual reality simulations, peer-to-peer/collaborative learning, distance education, connecting students with mentors and experts, virtual career exploration, and virtual internships, providing a flexible and accessible way for students to gain knowledge and experience from anywhere and at any time (Atsikpasi & Fokides, 2022).

As expected, researchers and practitioners have explored various ways in which the Metaverse can be used to enhance education and learning. Some examples of studies in which devices and applications belonging to the realm of Metaverse were used (specifically, VR headsets offering high levels of immersion) include ones to teach astronomy (Rupp et al., 2019), history (Fabola & Miller, 2016), physics (Pirker, Lesjak, Parger, & Gütl, 2018), and engineering (Shi, Du, & Worthy, 2020). Besides learning, studies have shown that the Metaverse can benefit students in other key areas, such as the acquisition of skills (Queiroz, Nascimento, Tori, & da Silva Leme, 2018), engagement (Bertrand, Bhargava, Madathil, Gramopadhye, & Babu, 2017), and motivation (Rupp et al., 2019). The studies also suggested that the Metaverse has the potential to provide new opportunities for students with different learning styles and to support students with special educational needs (Newbutt, Bradley, & Conley, 2019).

However, the literature also points out that the Metaverse is still *terra incognita*, due to the constant technological developments. More research is needed to understand how learning occurs in virtual spaces, identify the factors that come into play in shaping the learning experience, and identify best practices for its use in education (Atsikpasi & Fokides, 2022). To achieve the above, besides good research (in terms of sample sizes, materials, and settings), instruments are needed able to accurately measure attitudes, beliefs, behaviors, or other characteristics of the individuals or groups who use Metaverse's applications for learning. To this end, as will be elaborated in a coming section, previous studies used a multitude of instruments. There were cases in which validated, well-accepted, and widely used scales/questionnaires examining a specific (but limited) number of factors were used (i.e., the Simulator Sickness Questionnaire, Kennedy et al., 1993; the Mec Spatial Presence Questionnaire, Vorderer et al., 2004; and the Presence Questionnaire, Witmer & Singer, 1998). Several other researchers, for examining VR or AR educational applications, adapted validated scales/questionnaires examining multiple factors (i.e., the Computer Attitude Scale, Selwyn, 1997; the EGameFlow Scale, Fu et al., 2009). Then again, the original scales/questionnaires were meant to be used in different contexts and their adapted versions were not thoroughly validated for the new ones. Other assembled questionnaires using factors/items from different validated scales/questionnaires (mostly used in different contexts). Again, these questionnaires lacked in-depth validation. Finally, there were cases in which the researchers used questionnaires developed *ad hoc*, either by including items from other scales/questionnaires or items developed by the researchers themselves. From the above, it can be concluded that there is a need to develop a more comprehensive tool and validate it. This was the study's main objective, as it focused on the development and testing of the Metaverse Learning Experiences Scale (MLES), which tries to capture as many factors as possible that play a role in learners' experiences. The procedures that were followed for developing it and the results of its testing are presented in the coming sections.

2. Scale development considerations

The development of a rigorous scale is a multi-stage process. Boateng, Neilands, Frongillo, Melgar-Quiñonez, and Young (2018) suggested three main stages with several steps in each:

- Item development, the purpose of which is to come up with an initial pool of items to include. It involves (i) the domain identification and generation of items and (ii) content validity considerations.
- Scale development, in which the individual items are turned into constructs. It includes (i) the pre-testing of items, (ii) sampling and administration of the scale, (iii) item reduction, and (iv) the extraction of factors.
- Scale evaluation. It includes tests of (i) dimensionality, (ii) reliability, and (iii) validity.

Similarly, Hair, LDS Gabriel, Silva, and Braga (2019) suggested a five-stage process: (i) definition of the domain (i.e., what the scale is supposed to measure), (ii) literature review for identifying items or, alternatively, interviews with experts, (iii) face validation, (iv) semantic validation/validation with possible responders, and (v) statistical validation.

The procedure followed for the development of MLES was a combination/adaptation of the above two suggestions. Given that the domain definition/identification, is already elaborated in the previous section (the factors that come into play in one's learning experience in the Metaverse) the remaining steps were:

- Scale development, having three stages:
  - o Item development. The objective was to create an initial pool of items.
    - oFace validation, by asking experts in the field to evaluate the items and whether they are related to their respective constructs.
    - oPre-testing of the scale, by administering it to a number of possible responders (not included in the next step), followed by interviews of these responders.
- Scale evaluation, having three stages:
  - oMethod, including sampling/selection of participants, selection of the learning material to be used, testing procedures, and data screening.
  - oExploratory factor analysis (EFA), for item reduction and factor extraction.
  - oConfirmatory factor analysis (CFA), for accessing the MLES's dimensionality, reliability, as well as its convergent and discriminant validity.

3. Scale development procedures

When developing a questionnaire or scale examining factors affecting or shaping a variable of interest, researches usually have in mind which ones to include, depending on their interests, literature, context, and field of study. Yet, this can be viewed as a form of predetermination. For example, why a researcher decided to include factor A and not factor B or both; what makes a given factor more important than another; what if researcher A defines/views a factor differently than researcher B, resulting in different items examining the same concept; what to do when the phenomenon is too complex and many factors have to be considered? These are valid questions raising concerns not for the validity of the instruments (as this can be established through robust procedures) but for their comprehensibility. Thus, in this study, a different, more holistic approach was followed.

As mentioned in the introduction, in this study, the Metaverse is viewed as the sum of highly immersive and interactive 3D spaces, accessed using xR headsets/devices. On the other hand, according to a recent systematic literature review on the educational uses of the Metaverse, in which studies published from 2011 to 2022 were included, the authors found that, among the Metaverse's technologies, VR is, by far, the most researched one (Alfaisal, Hashim, & Azizan, 2022). There was a similar finding in the review of Tlili et al. (2022). As a result, it was decided to review the instruments used in past research in which educational applications were used together with VR headsets, and the impact on learning (or on factors related to learning) was examined. Besides selecting items to include in the scale under development, the objective was to have a better understanding of which factors other researchers considered important in shaping one's learning or learning

experience in the Metaverse. A scoping review that analyzed eighty-seven studies provided the basis for this endeavor (Atsikpasi & Fokides, 2022). That is because its focus was on the educational uses of immersive VR and the use of VR headsets; thus, it was closely related to how the Metaverse is viewed in this study. The papers included in the review of Alfaisal et al. (2022) were also considered, as a precautionary measure, given that it examined technologies other than VR.

3.1. Item development

The papers included in the above-mentioned reviews were examined in terms of the instruments they used. When a study adapted another study's instrument the original was included, together with any other factors/items the researchers added. Sixty-two questionnaires/scales were located (as some studies used the same instruments), having a total of 1634 items, examining 172 factors and sub-factors.

All the items were inserted in an Excel spreadsheet, together with the factors they belonged. If the authors of a paper did not mention such factors, the items were placed under the umbrella of a single factor which was the questionnaire's general theme/purpose.

Next, a multi-stage elimination process was initiated. The exact duplicates were deleted (293 items). Moreover, items belonging to constructs that were examined just once were also removed. Items that differed in terms of verb tense or word order (e.g., "I knew clearly what I wanted to do" and "I clearly know what I want to do") were excluded as well. It goes without saying that one item from each such case was retained. At this stage, 314 items were removed.

The remaining items were examined one by one to identify the conceptually similar. For example, items such as: "How involved were you in the virtual environment experience?" "Do you easily become deeply involved in computer games or video games?" "How involved were you in the game experience?" and "I can become involved in the game" fell into this category. This is because all examined the degree of the users' involvement in the subject under consideration. As in the previous phase, one item from each such case was retained. Overall, in this phase, 389 items were excluded.

A fairly significant number of conceptually identical or similar items were present in more than one factor. For example, the item "I enjoyed playing the game" according to Calvillo-Gámez et al. (2015) examined the factor of enjoyment, while according to Jennett et al. (2008), examined immersion. A more intricate example, which highlights the complexity of the problem, were the items (i) "Do you ever become so involved in doing something that you lose all track of time?" which according to Chen et al. (2011) examined immersive tendencies, (ii) "When I play this game, I tend to lose track of time", which according to Fang et al. (2013) examined the sub-factor transformation of time which belonged to the factor flow experience, (iii) "While reading the story, I lost track of time", which according to Hartung et al. (2016) examined the sub-factor attention which belonged to the factor immersion, (iv) "Were you involved in the experimental task to the extent that you lost track of time?" which according to Witmer and Singer (1998) examined the sub-factor involvement/control which belonged to the control factors, that, eventually, measured presence, and (v) "I lose track of time", which according to Brockmyer et al. (2009) examined temporal dissociation/absorption a sub-factor of the factor engagement.

The above examples are indications that certain concepts are so closely related that researchers either confused them or while trying to separate them, eventually examined them using the same or similar items. Such are the concepts of immersion and presence, as it is not uncommon for researchers to use these terms interchangeably (Wilkinson, Brantley, & Feng, 2021). In an effort to solve this problem, it was decided to create four larger constructs by moving into them items belonging to (i) dissociation, flow, immersion, and presence, (ii) usability, ease of use, and control (iii) aesthetics, audio aesthetics, visual aesthetics, and audiovisual aesthetics, and (iv) content, guidance, and feedback. After that, the items now belonging to the four consolidated constructs were re-checked and the conceptually similar ones were removed, regardless of the factor to which they initially belonged (374 items). It should be noted that the purpose was to transfer to another stage the control of whether or not an item examines one of the above difficult-to-define factors and whether or not it should be included, instead of deciding this at the present -early- stage, in which the personal opinion of the researcher might influence these decisions. It was also decided to create the following constructs: (i) learning enjoyment and enjoyment while using the virtual environment and (ii) motivation to learn and motivation to use the virtual environment. That is because in some cases the researchers were interested in motivation and enjoyment in relation to learning and in other in relation to the devices or applications that were used.

Finally, questions containing everyday expressions of the English language not having similar ones in Greek were rephrased or excluded (seventeen items). At the end of this phase, the remaining items were translated into Greek. Moreover, they were either reworded regarding the person and/or tense of the verbs (for example, the item "I lose my normal awareness of time" became "I lost my normal awareness of time") or reworded to fit the scope of the scale (for example, the item "Playing the online game was interesting in itself" became "The use of the virtual environment was interesting in itself."

It has to be noted that items examining users' collaboration or social experiences were considered. Then again, given that the application developed for the purposes of the study was not a multi-user one (see section "4.1.2 Materials and apparatus"), it was decided to remove them (eighteen items), but consider them for future inclusion, when testing multi-user applications.

The procedure resulted in MLES's initial draft, consisting of 208 items, coming from thirty-eight questionnaires/scales. Readers can find the complete list of the contributing questionnaires and scales in Appendix I. The items were theorized to belong to fifteen constructs, namely (i) immersion and presence, (ii) motivation to learn, (iii) motivation to use the virtual environment, (iv) learning enjoyment, (v) enjoyment while using the virtual environment, (vi) perceived ease of use/control of the virtual environment, (vii) positive feelings, (viii) negative feelings, (ix) simulator sickness, (x) perceived degree of interaction, (xi) perceived usefulness/knowledge gains, (xii) relevance to personal interests, (xiii) perceived quality of the virtual environment's graphics, (xiv) perceived cognitive load, and (xv) perceived feedback and content quality.

3.2. Face validation

Following a procedure similar to the one the Decision Delphi method suggested (Rauch, 1979), access was granted to the pool of items formed during the previous stage (both to the initial wording in English and the translated version) to five experts in the field. They were asked to thoroughly review the items and provide feedback, including arguments supporting their views, whether:

- The wording was appropriate for the given target audience and suggest alternative phrasing.
- Items should be added or removed.
- Technical terms or other concepts needed clarification or re-wording.
- Each group of items actually belonged to the construct that was theorized to belong.
- They were also instructed to take care so as constructs not to have too many or too few items.

Their comments/suggestions were written on a shared document so that the other experts to view and comment. Following several rounds of reviews-comments-suggestions, a consensus was reached and the panel of experts made its final proposal. Several items were eliminated or reworded. Also, the construct labeled "relevance to personal interests" was removed (together with its items), as the experts suggested that it (i) can be considered as a motivating factor (already included) and (ii) in

was unclear whether it added something important to the learning experience per se. Fourteen factors and sixty-four items remained for the next stage.

3.3. Pre-testing

The purpose of this stage was not to pilot the scale, but rather the examination of the difficulties the responders might face. Thirty university students were recruited. It has to be noted that the participants in this stage were not included in the next one. During a meeting, the scale's application was explained and they were asked to assist in the process of refining it by commenting on how easy is to understand its items.

Since few had a -rather vague-understanding of what the Metaverse is and since none had previously experienced an immersive VR application, they were given devices such as the ones used in the next stage (see section "4.1.2 Materials and apparatus"), were instructed on how to use them, and were allowed to interact with a variety of applications. They were also free to take the devices home and use them for as long as they liked during a three-day period. Following that, the items were presented to them (in print) and they were asked to highlight words or phrases they did not understand or were unclear and write down comments. Given that, at this stage, the scale was not administered anonymously, the participants were then interviewed (individually), so as to further explain the notes and remarks they made.

Reflection on the annotated responses and notes taken during the interviews followed, resulting in the re-wording of some items, while none was removed. As a precautionary measure, the revised items were sent to the panel of experts for their final thoughts.

No further changes were made and the scale's draft was finalized consisting of sixty-four items, belonging to fourteen theoretical constructs. The items were presented to the individuals who participated in the MLES's testing (see next sections), using a five-point Likert-type scale, ranging from 1-strongly disagree to 5-strongly agree. As the items theorized to belong to the constructs simulator sickness, positive, and negative feelings were presented as questions (e.g., "To what degree you felt dizziness?"), 1 represented none/not at all and 5 represented a lot/very much.

4. Scale testing

4.1. Method

4.1.1. Participants

There are no exact demographics regarding who uses the Metaverse (given the multiple definitions of the term), but data suggested that there are around 400 million active monthly users, more than half of them being registered users of a social/gaming platform (230 million). In fact, most users are gamers, out of which 51% are under the age of thirteen (Nikolovska, 2022). Then again, the study focused on the educational uses of the Metaverse. In an educational context, again no data are available, but an educated guess is that systematic users are more likely to be educators and individuals aged sixteen to late twenties (Tlili et al., 2022). Moreover, most studies in this field focused on higher education (Atsikpasi & Fokides, 2022; Tlili et al., 2022).

Another issue that had to be addressed was the sample size. However, this is not an easy-to-solve puzzle, as the sample size depends on a number of factors such as the availability of resources, the study's design, data scaling, indicator reliability, estimation method, and the model's complexity (Brown, 2015).

Reflecting on the above, it was decided the target group to be students studying at a Department of Primary Education in Greece, given that on the one hand they are university students and, on the other hand, they are future educators. A call for participation was issued, informing them about the study and its procedures. There were no prerequisites (e.g., prior experience in using VR headsets) and course credit was granted for participation. 473 students were enrolled, who gave informed consent to the study's procedures (in addition to the permission that was already granted by the Department's Ethical Committee). This number of participants satisfied Kline's (2016) suggestion for at least 200 cases, was within Bentler and Chou's (1987) suggestion of five to ten observations per estimated parameter, and above Hair's, Black's et al.'s (2019) suggestion of five responders per item.

4.1.2. Materials and apparatus

Though there are several educational applications available, just a handful are translated into Greek. A number of them were reviewed and it was concluded that they did not suit the study's needs. For example, some presented too much or complex or specialized information, while others needed too much time to complete. This necessitated the use of a tailor-made application, which was developed using Unity. A virtual guided tour of the city of Rhodes was selected as the application's theme, given that the Department of Primary Education resides in this city. Five virtual spaces were created, all of which simulated an outdoor environment with a gazebo placed in the middle of it. Users could freely walk within the gazebos' boundaries (physically, not virtually, so as to increase the realism of the experience). The first space served as the landing/welcome area, providing users with information (audio and written) on how to interact with the application and the 3D objects. It also had "teleport" buttons for visiting the other spaces. Each of these spaces presented one of the city's landmarks or places of historical/archaeological interest. Freely available 3D models of these landmarks were placed in the gazebos, with which users could interact. Info buttons were included, which provided written information for the place presented. In addition, there was a "video button," for users to view a short 3D-360° video (less than 2 min) of the given place. Audio information about the place was also included. The videos were captured using a Vuze 3D-360° camera.

It has to be noted that for interacting with the application and its objects, controllers were not necessary. Using the hand tracking technology implemented in the devices that were used (see below) and a series of Unity addons, the participants could interact with just their bare hands (e.g., they could pick up objects, rotate or throw them away and they could activate the info and teleport buttons by touching them). This was done for offering a more immersive (and realistic) user experience.

Since in this study, the Metaverse is viewed as the sum of highly immersive 3D virtual spaces that require the use of VR headsets so as individuals to experience them in full, the Meta Quest 2 VR headsets were used for viewing and interacting with the application. This device has several advantages besides hand-tracking technology. It is untethered, meaning that it does not have to be connected to a computer, has good technical specifications, and has an affordable price, rendering it a pretty attractive choice for the average user who wants to experience highly immersive VR. Not only that, but it allows for six degrees of freedom (forward/backward, up/down, left/right) so that users can freely move in space, in contrast to other devices that offer just three degrees of freedom (rotational motion, pitch, yaw, and roll).

For conducting the experiment, a spacious office was available (around 36 m^2), with more than enough area for participants to move around without bumping into the walls or furniture.

4.1.3. Procedure

As none of the participants had prior experience in using VR headsets, after welcoming each, they were given oral instructions on how to move in space (so as not to go beyond the boundaries of the "play" area) and how to use their hands. Next, they wore the headsets and adjusted the interpupillary distance (to achieve good image quality) and the straps. After booting the devices, they were given some time (around ten to 15 min) to familiarize themselves with the virtual environment. Additional instructions were given to participants who faced interaction and navigation problems.

Following that, the participants used the application for twenty-five to 30 min, which was considered enough to visit all five spaces. The

next step was to fill out the MLES online (using a Google form). The whole procedure lasted around 75 min. In case there was an incident of severe simulator sickness, users were instructed to remove the VR headset and rest, but were asked to fill out the MLES, given that simulator sickness is an important side effect of the use of VR headsets. In case of mild simulator sickness or discomfort, the decision to continue or stop the experiment was left to their discretion.

4.1.4. Data screening and preparation

For all the analyses presented in the coming sections, the SPSS 28 statistics software was used. The negatively worded items were reverse-coded for obtaining interpretable factor loadings. There were no missing data, but eleven participants were excluded because of their unengaged responses (i.e., the standard deviation in their responses was 0.00). As a result, the final sample size was 462 individuals. As expected (because the participants were students studying at a department of primary education), the majority belonged to the 20–24 years old group (84.40%), while most were females ($N = 340$).

Next, it was examined whether the data met the assumptions for conducting EFA and CFA. The visual inspection of the full correlation matrix revealed a substantial number of meaningful relationships among the items (Pearson's correlation coefficient >0.30). This was also confirmed by the level of significance of Bartlett's test of sphericity (approx. Chi-square = 13046.893, $p < .001$) and the high value of Kaiser-Meyer-Olkin measure of sampling adequacy (KMO = 0.917) (Tabachnick & Fidell, 2007).

The assumption of multivariate normality, although not required for the EFA (when using principal axis factor analysis, PAF), is required for the CFA, in which maximum likelihood extraction methods are used, although moderate deviations are acceptable (Matsunaga, 2010). The multivariate normality was checked by examining the items' skewness and kurtosis. In several items, skewness and kurtosis were well above the recommended values (for skewness < |2| and for kurtosis <7, Finney & DiStefano, 2013). As all these items were theorized to examine two constructs (simulator sickness and negative feelings), it was decided to transform their data using log(10) and re-examine them. Unfortunately, as these indices remained very high, all five items belonging to negative feelings were dropped, together with two items belonging to simulator sickness. Although the values of skewness and kurtosis in some other items were slightly above the recommended values, it was decided not to transform their data because: (i) avoiding transformation allows for a better interpretation of the results, especially in highly exploratory studies (as in this case), (ii) there is literature suggesting that the factorial analysis can be performed in skewed and kurtotic data (Wang, Fan, & Willson, 1996), and (iii) data transformation affects reliability (Cronbach's α) and, thus, may not always be appropriate (Norris & Aroian, 2004).

The final step at this stage was to randomly split the data into two halves; one half ($n = 231$) was used for conducting the EFA and the other half for the CFA.

4.2. Exploratory factor analysis and polytomous Rasch model analysis

Since the scale was based on selected items from multiple sources that were adapted and translated into Greek, it was essential to conduct an EFA, for uncovering its underlying structure and for establishing the relationship between the variables and the latent constructs. PAF was selected as the extraction method, as the covariation between variables is taken into account (Kline, 2015). Direct Oblimin was selected as the rotation method, given that oblique (nonorthogonal) rotations produce more accurate results when the research involves human behaviors (Costello & Osborne, 2005) and offers a more realistic representation and easier interpretation of the data (Fabrigar & Wegener, 2012).

Several items were removed on the basis of the following criteria: (i) low communality coefficients (<.40), (ii) factor loadings below |0.50|, (iii) cross-loadings on two or more factors greater than |0.30|, (iv) low conceptual relevance to a factor, and (v) conceptual inconsistency with the items in the same factor (Costello & Osborne, 2005; Hair, Black, Babin, & Anderson, 2019; Tabachnick & Fidell, 2007). Every time an item was removed, the analysis was re-run, to ensure that there were no major effects on the scale's structure. This process resulted in the elimination of twelve items, while forty-five items were retained belonging to ten factors.

At this point, a polytomous Rasch model analysis (Andrich, 1978) was run. In general, the Rasch model provides a framework that allows researchers to assess whether an instrument is capable to quantify unobservable human conditions/behaviors and to emulate the properties of fundamental measurement (invariance and unidimensionality). For that matter, the software package Jamovi 2.3.21, together with the eRm R package was used, while the results were estimated by Marginal Maximum Likelihood Estimation. Two fit indices, namely the Infit (Information-weighted mean square statistic) and Outfit (Outlier-sensitive means square statistic) were used to investigate whether the items contributed adequately to their domain. The analysis revealed that two items, theoretically belonging to the factor labeled "simulator sickness," had to be removed because their Infit and Outfit values exceeded the recommended maximum value of 1.5 (Linacre, 2002). The Wright maps indicated that all the remaining items were clustered together in their respective factors, meaning that the items in a factor had -more or less-the same level of difficulty (or easiness).

The EFA was run for a final time using the remaining forty-three items. Kaiser's (1960) criterion (eigenvalue >1), suggested a ten-factor solution. As proposed by Velicer, Eaton, and Fava (2000), a parallel analysis was run (using PAF as the extraction method), following O'Connor's method (2000). The analysis confirmed the existence of ten factors (Table 1).

Moreover, the ten-factor solution appeared to be the most parsimonious and conceptually sound one. The factors were named as follows: motivation to learn and use the virtual environment (Mot), simulator sickness (SimSick), perceived cognitive load (CogLoad), perceived usefulness/knowledge gains (Gains), positive feelings (PosFeel), perceived ease of use/control of the virtual environment (Contr), perceived degree of interaction (Inter), perceived quality of the virtual environment's graphics (Graphics), immersion/presence (Imm/Pre), and perceived feedback and content quality (Feed/Cont).

The items' communalities were more than acceptable (the lowest value observed was 0.516). All the items loaded high in their respective factors (>0.60), while each factor averaged around the 0.70 recommended level (Hair, Black, et al., 2019) (Table 2). There were at least three items per factor, as suggested by Raubenheimer (2004). No issues regarding item cross-loadings were noted. The highest factor correlation that was observed was 0.56, well below the value of 0.80 which implies discriminant validity issues (Brown, 2015) (Table 3). The total variance

Table 1
Parallel analysis.

Number of factors	Random data eigenvalues	Actual data eigenvalues
1	1.275	16.053
2	1.109	4.409
3	1.015	3.017
4	0.940	2.058
5	0.878	1.710
6	0.804	1.514
7	0.758	1.326
8	0.706	1.197
9	0.654	0.997
10	0.607	0.887
11	0.553	0.471
12	0.525	0.374
13	0.477	0.331

Note. The 10th factor was the last one in which the eigenvalue of the actual data exceeded the eigenvalue of the random ones; thus, the analysis suggested a ten-factor solution.

Table 2
The results of the final EFA.

Item	Factor									
	Mot	SimSick	Imm	CogLoad	Gains	PosFeel	Contr	Inter	Graphics	Feed/Cont
MotApp1	.804									
MotLearn3	.788								.108	
MotApp2	.763									
MotApp3	.752									
MotLearn2	.627		−.122	.114					.137	
MotLearn1	.626						.129			−.201
SimSick10		.835								
SimSick3		.832							−.103	
SimSick5		.830								
SimSick4		.825								
SimSick2		.751			−.117					
SimSick1		.652	−.110	−.127						
Imm/Pre4			−.932							
Imm/Pre3			−.871							
Imm/Pre6			−.844							
Imm/Pre5			−.798					−.115	.114	
Imm/Pre1			−.771						.147	
Imm/Pre9	.150		−.714							
CogLoad2				.865						
CogLoad3				.849						
CogLoad1				.790					.118	−.118
Gains3					.906					
Gains3					.902					
Gains1					.672					
PosFeel3						.834				
PosFeel1						.826				
PosFeel2						.807			.111	
PosFeel5	.109					.740			.120	
Contr5							.903			
Contr6							.788	−.125		
Contr4			−.112				.701			
Inter2								−.789		
Inter1								−.743		
Inter3								−.730		
Graphics3									.800	
Enj1									.748	
Graphics4									.738	
Graphics2									.692	
Graphics1						.109			.669	
Enj2							.118		.668	−.106
Feed/Cont1										−.768
Feed/Cont3									.136	−.729
Feed/Cont2	.122				.125					−.678
Initial Eigenvalues	16.165	4.561	3.124	2.192	1.826	1.615	1.433	1.335	1.107	1.006
% variance explained Total = 79.92	37.59	10.61	7.27	5.10	4.25	3.76	3.33	3.11	2.57	2.34
Cronbach's α Total = .953	.932	.904	.945	.884	.913	.949	.894	.870	.934	.888

Notes. Extraction method: PAF; rotation method: Oblimin with Kaiser Normalization; the rotation converged in 11 iterations; coefficients $< |0.10|$ were suppressed for clearance of presentation.

Table 3
Factor correlation matrix.

Factor	Mot	SimSick	Imm	CogLoad	Gains	PosFeel	Contr	Inter	Graphics	Feed/Cont
Mot	1.000	−.174	−.297	.186	.428	.487	.489	−.406	.561	−.428
SimSick	−.174	1.000	−.014	−.270	−.265	−.183	−.197	.051	−.191	.089
Imm/Pre	−.297	−.014	1.000	−.108	−.366	−.264	−.271	.396	−.332	.421
CogLoad	.186	−.270	−.108	1.000	.361	.181	.164	−.135	.217	−.268
Gains	.428	−.265	−.366	.361	1.000	.378	.331	−.345	.288	−.378
PosFeel	.487	−.183	−.264	.181	.378	1.000	.467	−.395	.530	−.429
Contr	.489	−.197	−.271	.164	.331	.467	1.000	−.409	.428	−.275
Inter	−.406	.051	.396	−.135	−.345	−.395	−.409	1.000	−.410	.406
Graphics	.561	−.191	−.332	.217	.288	.530	.428	−.410	1.000	−.500
Feed/Cont	−.428	.089	.421	−.268	−.378	−.429	−.275	.406	−.500	1.000

explained by the ten components was 79.92%. As far as the internal consistency was concerned, as assessed using Cronbach's alpha, it was very good, ranging between 0.870 and 0.949 for the constructs, while the overall score was 0.953 (DeVellis, 2016, Table 2, last row).

4.3. Confirmatory factor analysis

For conducting the CFA, AMOS 28 was employed using the remaining half of the data. Prior to conducting the analysis, the internal consistency was re-checked. No substantial variations from the initial assessment were noted. Maximum Likelihood was the estimation method of choice,

because of its robustness to moderate violations of normality (Matsunaga, 2010), and because the quality of the parameter estimates is not affected (Brown, 2015). As is evident in Table 4, the standardized estimates were very good as they ranged from 0.70 to 0.94 (Hair, Black, et al., 2019). The same applied to the R^2 values, given that all but one were above the 0.50 threshold, meaning that they explained more than half of the variance of the latent factor they belonged. The only exception was SimSick1. Then again, as its value was very close to the above threshold ($R^2 = 0.48$), it was not considered an issue.

The scale's reliability and convergent validity were checked by measuring the Composite Reliability (CR) and the Average Variance Extracted (AVE), which is a more conservative measure of convergent validity (Malhotra & Dash, 2011, Table 5). No issues were found as the CR in all cases was above the 0.70 threshold and the AVE was above the 0.50 level (Hair, Black, et al., 2019). For assessing the scale's discriminant validity, the heterotrait-monotrait ratio of correlations (HTMT) technique was used, as the sensitivity of shared variance in capturing discriminant validity issues between constructs has been questioned (Henseler, Ringle, & Sarstedt, 2015). The discriminant validity was established as all the HTMT values were below the strict threshold of 0.85 (Table 6).

For model fit assessment, the following were estimated: (i) the comparative fit index (CFI), values exceeding 0.95 indicate excellent model fit, (ii) the root mean square error of approximation (RMSEA), values less than 0.06 indicate a very good model fit, (iii) the standardized root mean square residual (SRMR), values less than 0.08 are used as a cut-off point, indicating excellent model fit, (iv) PClose, values above 0.05 indicate good fit, and (v) the minimum discrepancy divided by its degrees of freedom (χ^2/df), in which the acceptable values are between 1 and 3 (Hu & Bentler, 1999). Evidently, all the indices indicated an excellent model fit (Table 7, second column).

To further solidify the results, the EFA was re-run, forcing nine and eight-factor solutions. The fit indices of the three models were then compared. A one-factor model was also used as a baseline. The results, as presented in Table 7, confirmed that the ten-factor solution had the best overall model fit.

Summarizing the results, it can be concluded that both the EFA and CFA established that the scale's validity and reliability were satisfactory. Its final version with the forty-three retained items can be found in Appendix II.

Table 5
Reliability and convergent validity.

Factor	CR	AVE	Factor	CR	AVE
Mot	0.93	0.70	PosFeel	0.94	0.81
SimSick	0.91	0.64	Contr	0.90	0.74
Imm/Pre	0.95	0.74	Inter	0.87	0.69
CogLoad	0.89	0.72	Gains	0.92	0.79
Graphics	0.93	0.70	Feed/Cont	0.89	0.73

5. Discussion

Past research demonstrated that the Metaverse's applications can be used in diverse learning domains (e.g., Fabola & Miller, 2016; Pirker et al., 2018; Rupp et al., 2019) and benefit students in areas that go beyond learning (e.g., Bertrand et al., 2017; Queiroz et al., 2018; Rupp et al., 2019). However, there is no consensus on how to measure users' views, attitudes, and feelings related to the experiences they had while learning in the Metaverse. Furthermore, during the process of developing the MLES, it became evident that researchers attempted to do that, but they used unvalidated tools, examined a limited number of factors, ill-defined others, and used the same items to examine different factors. Contrary to that, the MLES is the result of a rigorous multistage process that followed the steps suggested by the literature related to scale development and testing. Indeed, in the pursuit of its development, an extensive number of resources (i.e., scales and questionnaires used in past research related to Metaverse applications), were screened, resulting in a rather large initial pool of items. A series of iterative phases followed, in which experts suggested modifications and item removal. Individuals belonging to the audience the scale is addressed also participated and provided feedback. The refined scale was administered to 462 university students. The subsequent statistical analyses resulted in a multi-dimensional, yet parsimonious (in terms of the total number of items it has) scale.

Regarding MLES's factorial structure, the following can be noted. Ten factors emerged from the analyses although it was theorized that the items examined fourteen. That is because some factors were eliminated or merged with others. All the items theoretically examining the factor labeled “negative feelings,” were dropped at the early stages of the statistical analysis. The reason was that their data were not suited for the subsequent statistical procedures (extreme data skewness). A quick visual inspection revealed a problem similar to the one found in a number of

Table 4
Results of the CFA.

Item	Est.	SE	t	p	R^2	Item	Est.	SE	t	p	R^2
MotApp1	.85	–	–	–	.67	Imm/Pre9	.82	.07	15.82	<.001	.67
MotLearn3	.91	.06	18.26	<.001	.57	Imm/Pre5	.86	–	–	–	.74
MotApp2	.84	.06	16.29	<.001	.71	Imm/Pre6	.88	.07	18.15	<.001	.78
MotApp3	.83	.05	18.36	<.001	.68	Imm/Pre3	.89	.07	18.39	<.001	.79
MotLearn2	.76	.07	13.52	<.001	.83	Imm/Pre4	.89	.07	18.28	<.001	.78
MotLearn1	.82	.06	14.92	<.001	.71	PosFeel5	.93	.05	21.95	<.001	.87
SimSick10	.83	–	–	–	.69	PosFeel2	.94	–	–	–	.78
SimSick5	.82	.07	14.65	<.001	.67	PosFeel1	.88	.04	22.27	<.001	.70
SimSick2	.75	.09	12.89	<.001	.56	PosFeel3	.84	.05	22.35	<.001	.89
SimSick4	.87	.07	15.95	<.001	.75	Enj2	.87	.06	16.87	<.001	.75
SimSick3	.83	.07	15.02	<.001	.69	Graphics1	.79	.06	14.34	<.001	.62
SimSick1	.70	.11	11.67	<.001	.48	Graphics2	.80	.06	14.63	<.001	.64
CogLoad2	.86	–	–	–	.75	Graphics4	.86	.06	16.46	<.001	.73
CogLoad3	.88	.06	15.81	<.001	.78	Enj1	.86	.06	16.68	<.001	.74
CogLoad1	.80	.07	14.18	<.001	.64	Graphics3	.84	–	–	–	.71
Contr5	.89	–	–	–	.78	Inter3	.83	.08	13.89	<.001	.69
Contr6	.85	.06	16.45	<.001	.72	Inter1	.85	.07	14.13	<.001	.72
Contr4	.85	.06	16.32	<.001	.72	Inter2	.82	–	–	–	.68
Gains2	.92	.05	22.90	<.001	.85	Feed/Cont2	.86	.07	15.09	<.001	.74
Gains3	.94	–	–	–	.88	Feed/Cont3	.87	.07	15.38	<.001	.76
Gains1	.79	.05	16.59	<.001	.63	Feed/Cont1	.82	–	–	–	.68
Imm/Pre1	.84	.05	19.80	<.001	.71						

Notes. -: This value was fixed at 1.00 for model identification purposes; Est.: standardized estimate; SE: standard error.

Table 6
HTMT analysis.

Factor	Mot	SimSick	Imm	CogLoad	Graphics	PosFeel	Contr	Inter	Gains	Feed/Cont
Mot										
SimSick	0.24									
Imm/Pre	0.42	0.03								
CogLoad	0.28	0.32	0.17							
Graphics	0.71	0.26	0.45	0.29						
PosFeel	0.64	0.25	0.39	0.26	0.70					
Contr	0.61	0.25	0.39	0.23	0.58	0.59				
Inter	0.56	0.12	0.51	0.21	0.57	0.55	0.55			
Gains	0.55	0.32	0.46	0.44	0.45	0.50	0.43	0.46		
Feed/Cont	0.62	0.16	0.55	0.36	0.67	0.61	0.44	0.58	0.54	

Table 7
Model fit comparisons.

Measure	10-factor model estimates	9-factor model estimates	8-factor model estimates	1-factor model estimates	Thresholds
χ^2	1146.86	1382.05	1956.68	4861.26	–
DF	809.00	818.00	826.00	854.00	–
χ^2/df	1.42	1.69	2.37	5.69	Between 1 and 3
CFI	.96	.93	.87	.53	>.95
SRMR	.05	.06	.09	.13	<.08
RMSEA	.04	.06	.08	.14	<.06
PClose	.99	.06	.00	.00	>.05

items in the factor simulator sickness, but more intense. Specifically, there was just a handful of responses with values ranging from two to five; almost all were ones. While this result was not good for statistical procedures, it can serve as an indicator that participants' negative feelings (e.g., stress, nervousness, and irritation) were almost nonexistent. Nevertheless, the items in this factor are presented in Appendix II, so that future studies might consider them.

It was initially theorized that motivation to learn and motivation to use the virtual environment were two distinct constructs. As is evident in Table 2, they were merged in a single factor. This indicates that users did not make any distinction between how motivated they were to learn and how motivated they were to continue interacting or viewing the virtual environment. On the basis of this outcome, one might argue that motivation is a single construct in which both sub-constructs contribute as they have a two-way relationship; when individuals are motivated to learn they will be also motivated to use the virtual environment and vice versa.

As in motivation, enjoyment was also theorized to be two distinct constructs. Yet, the items examining learning enjoyment were dropped during the first stages of the EFA, while two of the items examining enjoyment while using the virtual environment were merged with the factor labeled "perceived quality of the virtual environment's graphics." Indeed, by re-examining these items, it was concluded that they conceptually fit into this factor. Although past research indicated that enjoyment plays a significant role in one's experience in VR (Atsikpasi & Fokides, 2022), the above resulted in not having items examining it. On the other hand, the factor labeled "positive feelings" can substitute for this absence. In fact, one item in this factor is directly related to enjoyment (To what degree you felt joy?"), while the remaining three can be considered indirectly related (see Appendix II).

A factor that could have been labeled "presence," did not emerge during the EFA. As mentioned in section "3.2 Item development," immersion and presence cause some confusion among researchers (Wilkinson et al., 2021), and this is reflected in how they are examined (e.g., by using the same items to examine both). Therefore, during the early stages of MLES's development, items belonging to both were merged and the decision to split them was left to be taken during the statistical analysis of the data. Then again, items that could -theoretically- examine presence either remained under the umbrella of the unified factor (immersion/presence) or were dropped as they loaded very low in other factors. On the other hand, it is suggested that, although not identical, the two concepts are very closely related; immersion reflects the technical qualities of the system that aid or lead to the feeling/experience of presence (Witmer, Jerome, & Singer, 2005). In this respect, it can be argued that having a single factor in MLES to examine both concepts, reflects this close relationship. Moreover, because of this close relationship, it can be argued that the same factors influence both, while both influence the same factors.

Coming to the number of items that were dropped, it is more possible that the main reason was the procedures that were utilized, which followed rather strict rules for factor or item retention. In general, the recommendations for satisfactory communality coefficients (>.40), high item loadings (>|0.50|), low cross-loadings (<|0.30|), low factor intercorrelations (<0.80), and satisfactory internal consistency (>0.70) (Brown, 2015; Costello & Osborne, 2005; DeVellis, 2016; Hair, Black, et al., 2019; Tabachnick & Fidell, 2007) were applied.

It is true that more items could have been retained if more relaxed rules were applied. It is also true that more constructs could have been added (together with items examining them), given that the scale's objective was to examine one's learning experience in the Metaverse which is a multidimensional concept. On the one hand, lengthy scales/questionnaires with many items per factor, tend to have greater α values; thus, they are considered more reliable (DeVellis, 2016). On the other hand, they are not that attractive to responders, driving researchers towards developing shorter ones (Olaru, Witthöft, & Wilhelm, 2015). Nevertheless, the MLES has at least three items per factor, which, by some, is considered the minimum required (Raubenheimer, 2004). Consequently, the MLES, having forty-three items examining ten factors is neither too short nor too long.

As for the number of factors, it is an issue of balance between complexity and stability. It is true that more factors could have been included. Collaboration (in the case of multi-user applications) and self-efficacy (commonly included in other questionnaires/scales) are good examples. Readers might come up with even better examples. Then again, during the EFA, factors were merged, items theorized to belong in a factor ended up in different ones, and several items were removed. Thus, it is probable that, in more complex factorial structures, such issues will become even more pronounced, resulting in conceptually obscure ones. As this was not the case in MLES and on the basis of its good attributes, as these emerged during the EFA and the CFA, it can be supported that the scale is quite robust and able to accurately measure a sufficient number of factors that come into play in shaping one's learning experience when using Metaverse applications.

5.1. Implications for research and practice

Experts involved in the development of educational applications, researchers, and educators need a scale psychometrically validated and suitable for a variety of evaluation scenarios in the context of the Metaverse. As discussed in previous sections, the scale's attributes (e.g.,

consistency, validity, and reliability), are indicators of its robustness. Moreover, because the scale is fairly short, individuals can complete it in around ten to 15 min, without being overwhelmed. Thus, it can be argued that the scale's contribution to research is that it provides a condensed tool able to simultaneously measure multiple aspects of one's learning experience. Hopefully, this can lead to the much-desired common methodology for assessing the Metaverse's educational impact.

A logical assumption is that the scale has a modular structure. The strong items' factor loadings and the factors' reasonable cross-correlations imply that factors can be excluded without altering the scale's validity. Indeed, to test this assumption, several EFAs were run, in which up to four randomly selected factors (and their items) were removed (e.g., perceived cognitive load, perceived feedback and content quality perceived degree of interaction, and positive feelings). In all cases, the factorial structure remained unchanged, while no items were removed. In this respect, researchers can use the scale for assessing different affordances, depending on their interests and needs. For example, if the researchers are not interested in interactions, they can remove the corresponding factor.

The scale can be administered to diverse groups of learners, for example, those belonging to different levels of education, having varying degrees of academic performance, and different levels of experiences in the Metaverse. Thus, application developers can either sum the items' scores and have a composite index of their learning experience or focus on certain components. By comparing the results, they can determine what needs to be changed. The same can be done for different versions of the same application. Researchers and educators can also benefit in a similar fashion. By administering different systems, they can evaluate the differences in students' learning experiences.

5.2. Limitations and future studies

The study is subject to limitations, the first being the sample size, given that "the more the merrier." The trustworthiness of the responses is also a concern. While it was theorized that university students are more likely to constitute the target population of Metaverse's educational applications, the study's sample might not be representative. Although the participants interacted with the application for enough time to visit all of its spaces, it might have not been enough to have a comprehensive experience. The application developed for this study falls into just one out of the many categories of educational applications. Thus, it is unknown how participants might react to different types of applications and what impact this might have on the MLES. As discussed in the previous section, some factors were dropped during the various stages of MLES's development. Despite the arguments presented, one might still argue that the exclusion of some factors negatively affected the scale's comprehensiveness. Lastly, the scale was not tested using other xR technologies; therefore, it is possible that other factors, related to a specific technology, were left out of the equation.

The limitations can serve as guidelines for future studies. The scale's reliability and validity have to be tested multiple times before it can be established as a valid tool. Different target populations, a greater variety of applications, and xR technologies will provide evidence of how well the scale functions and what revisions have to be made. The addition of other factors that shape one's learning experience can be considered (e.g., self-efficacy, collaboration, negative feelings, and relevance to personal interests), though one has to be cautious not to disturb the factors' delicate balance. Finally, a research path of great interest is to use the MLES together with tests that measure the learning outcomes produced by a Metaverse educational application. This will offer insights into how the scale's factors interact and what impact they have on learning.

6. Conclusion

The Metaverse is just around the corner and its applications have the potential to become valuable tools in the hands of educators. Yet, researchers and education stakeholders have to first understand this potential. To this end, instruments are needed, able to outline what learners feel and think when using Metaverse applications. This was the study's primary objective. Through the previous sections, the various steps that led to the development and psychometric validation of the MLES were elaborated. In conclusion, it can be argued the study's contribution is the development of a scale that can be used to evaluate a variety of educational applications in the Metaverse context. By being consistent, valid, and reliable, the scale provides a robust tool to measure a substantial number of factors shaping one's learning experience. Furthermore, the scale's brevity makes it accessible to individuals who can complete it in a short amount of time. Given the above, the tool has the potential to provide researchers, educators, and developers with valuable insights into the effectiveness of educational applications in the Metaverse. As such, it is hoped that its use will assist the scientific community in reaching a better understanding of the impact of Metaverse.

Funding

The study received no funding.

Data availability

Data are available on request from the author.

Ethical approval

The study's participants provided their informed consent. The study is part of an ongoing research project for which approval from the Department's Ethical Committee has been granted.

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.

List of acronyms

AR: Augmented Reality
AVE: Average Variance Extracted
CFA: Confirmatory Factor Analysis
CFI: comparative fit index
Contr: perceived ease of use/control of the virtual environment
CogLoad: perceived cognitive load
CR: Composite Reliability
EFA: Exploratory Factor Analysis
Feed/Cont: perceived feedback and content quality
Gains: perceived usefulness/knowledge gains
Graphics: perceived quality of the virtual environment's graphics
HTMT: heterotrait-monotrait ratio of correlations
Imm/Pre: immersion/presence
Inter: perceived degree of interaction
MLES: Metaverse Learning Experiences Scale
Mot: motivation to learn and use the virtual environment
PAF: Principal Axis Factor analysis
PosFeel: positive feelings
RMSEA: root mean square error of approximation
SimSick: simulator sickness
SRMR: standardized root mean square residual
VR: Virtual Reality
Xr: Extended Reality

Appendix I

References to the studies, questionnaires, and scales that contributed to the development of the MLES.

Factor	Reference
Perceived quality of the virtual environment's graphics	IJsselsteijn, W. A., de Kort, Y. A., & Poels, K. (2013). The game experience questionnaire. *Eindhoven: Technische Universiteit Eindhoven*, 3-9 Parnell, M. J. (2009). Playing with scales: Creating a measurement scale to assess the experience of video games [Master's thesis, University College London]. https://uclic.ucl.ac.uk/content/2-study/4-current-taught-course/1-distinction-projects/9-09/2009-parnell.pdf Phan, M. H., Keebler, J. R., & Chaparro, B. S. (2016). The development and validation of the game user experience satisfaction scale (GUESS). *Human Factors, 58*(8), 1217–1247. https://doi.org/10.1177/0018720816669646 Wiebe, E. N., Lamb, A., Hardy, M., & Sharek, D. (2014). Measuring engagement in video game-based environments: Investigation of the User Engagement Scale. *Computers in Human Behavior, 32*, 123–132. https://doi.org/10.1016/j.chb.2013.12.001
Perceived cognitive load	Fu, F. L., Su, R. C., & Yu, S. C. (2009). EGameFlow: A scale to measure learners' enjoyment of e-learning games. *Computers & Education, 52*(1), 101–112. https://doi.org/10.1016/j.compedu.2008.07.004 Ibili, E., & Billinghurst, M. (2019). Assessing the relationship between cognitive load and the usability of a mobile augmented reality tutorial system: A study of gender effects. International Journal of Assessment *Tools in Education, 6*(3), 378–395. https://doi.org/10.21449/ijate.594749 Keller, J. M. (2006). Development of two measures of learner motivation: CIS and IMMS. https://studylib.net/doc/7446614/development-of-two-measures-of-learner-motivation Nijs, L., Coussement, P., Moens, B., Amelinck, D., Lesaffre, M., & Leman, M. (2012). Interacting with the Music Paint Machine: Relating the constructs of flow experience and presence. *Interacting with Computers, 24*(4), 237–250. https://doi.org/10.1016/j.intcom.2012.05.002 Wiebe et al., 2014. See "Perceived quality of the virtual environment's graphics"
Perceived ease of use/control of the virtual environment	Abeele, V. V., Spiel, K., Nacke, L., Johnson, D., & Gerling, K. (2019). Development and validation of the player experience inventory: a scale to measure player experiences at the level of functional and psychosocial consequences. *International Journal of Human-Computer Studies*, *102370*. https://doi.org/10.1016/j.ijhcs.2019.102370 Agarwal, R., & Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS Quarterly, 24*(4), 665–694. https://doi.org/10.2307/3250951 Fang, X., Zhang, J., & Chan, S. S. (2013). Development of an instrument for studying flow in computer game play. *International Journal of Human-Computer Interaction, 29*(7), 456–470. https://doi.org/10.1080/10447318.2012.715991 Fu et al., 2009. See "Perceived cognitive load" Hassenzahl, M., Burmester, M., & Koller, F., 2003. AttrakDiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer Qualität [AttracDiff: A questionnaire to measure perceived hedonic and pragmatic quality]. In J. Ziegler & G. Szwillus (Eds.), *Mensch & computer 2003. Interaktion in bewegung* (pp. 187–196). B. G. Teubner. https://doi.org/10.1007/978-3-322-80058-9_19 Jackson, S. A., & Marsh, H. W. (1996). Development and validation of a scale to measure optimal experience: The Flow State Scale. *Journal of Sport and Exercise Psychology, 18*(1), 17–35. https://doi.org/10.1123/jsep.18.1.17 Nijs et al., 2012. See "Perceived cognitive load" Pavlas, D., Jentsch, F., Salas, E., Fiore, S. M., & Sims, V. (2012). The play experience scale: development and validation of a measure of play. *Human Factors, 54*(2), 214–225. https://doi.org/10.1177/0018720811434513 Phan et al., 2016. See "Perceived qualityof the virtual environment's graphics" Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion, 30*(4), 344–360. https://doi.org/10.1007/s11031-006-9051-8 Selwyn, N. (1997). Students' attitudes toward computers: Validation of a computer attitude scale for 16–19 education. *Computers & Education, 28*(1), 35–41. https://doi.org/10.1016/S0360-1315(96)00035-8 Slater, M. (1999). Measuring presence: A response to the Witmer and Singer presence questionnaire. *Presence, 8*(5), 560–565. https://doi.org/10.1162/105474699566477
Enjoyment while using the virtual environment	Calvillo-Gámez, E. H., Cairns, P., & Cox, A. L. (2015). Assessing the core elements of the gaming experience. In *Game user experience evaluation* (pp. 37–62). Springer. https://doi.org/10.1007/978-3-319-15985-0_3 Fang et al., 2013. See "Perceived ease of use/Control of the virtual environment" IJsselsteijn et al., 2013. See "Perceived quality of the virtual environment's graphics"
Learning enjoyment	Chou, S.-W., & Liu, C.-H. (2005). Learning effectiveness in a Web-based virtual learning environment: a learner control perspective. *Journal of Computer Assisted Learning, 21*(1), 65–76. https://doi.org/10.1111/j.1365-2729.2005.00114.x Keller, 2006. See "Cognitive load" Pavlas et al., 2012. See "Perceived ease of use/Control of the virtual environment"
Immersion/Presence	Abeele et al., 2019. See "Perceived ease of use/Control of the virtual environment" Agarwal & Karahanna, 2000. See "Perceived ease of use/Control of the virtual environment" Baños, R. M., Botella, C., Garcia-Palacios, A., Villa, H., Perpiña, C., & Alcañiz, M. (2000). Presence and reality judgment in virtual environments: A unitary construct? *CyberPsychology & Behavior, 3*(3), 327–335. https://doi.org/10.1089/10949310050078760 Boletsis, C. (2020). A user experience questionnaire for VR locomotion: Formulation and preliminary evaluation. *Proceedings of the International Conference on Augmented Reality, Virtual Reality and Computer Graphics*, 157–167. Springer. https://doi.org/10.1007/978-3-030-58465-8_11 Brockmyer, J. H., Fox, C. M., Curtiss, K. A., McBroom, E., Burkhart, K. M., & Pidruzny, J. N. (2009). The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology, 45*(4), 624–634. https://doi.org/10.1016/j.jesp.2009.02.016 Hartung, F., Burke, M., Hagoort, P., & Willems, R. M. (2016). Taking perspective: Personal pronouns affect experiential aspects of literary reading. *PloS one, 11*(5), e0154732. https://doi.org/10.1371/journal.pone.0154732 IJsselsteijn et al., 2013. See "Perceived quality of the virtual environment's graphics" Makransky, G., Lilleholt, L., & Aaby, A. (2017). Development and validation of the Multimodal Presence Scale for virtual reality environments: A confirmatory factor analysis and item response theory approach. *Computers in Human Behavior, 72*, 276–285. https://doi.org/10.1016/j.chb.2017.02.066 Rheinberg, F., Vollmeyer, R., & Engeser, S. (2003). *Die erfassung des flow-erlebens* [The capture of flow experience]. https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/deliver/index/docId/551/file/Rheinberg_ErfassungFlow_Erleben_mitAnhangFKS.pdf Ryan et al., 2006. See "Perceived ease of use/Control of the virtual environment" Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The experience of presence: Factor analytic insights. *Presence: Teleoperators and Virtual Environments, 10*(3), 266–281. https://doi.org/10.1162/105474601300343603

(*continued on next column*)

(*continued*)

Factor	Reference
	Vorderer, P., Wirth, W., Gouveia, F., Biocca, F., Saari, T., & Jiancke, F. (2004). Mec Spatial Presence Questionnaire: Short documentation and instructions for application. Report to the European Community, Project Presence. https://academic.csuohio.edu/kneuendorf/frames/MECFull.pdf
	Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence, 7*(3), 225–240. https://doi.org/10.1162/105474698565686
Perceived feedback and content quality	Fu et al., 2009. See "Perceived cognitive load"
	Högberg, J., Hamari, J., & Wästlund, E. (2019). Gameful Experience Questionnaire (GAMEFULQUEST): an instrument for measuring the perceived gamefulness of system use. *User Modeling and User-Adapted Interaction, 29*(3), 619–660. https://doi.org/10.1007/s11257-019-09223-w
	Phan et al., 2016. See "Perceived quality of the virtual environment's graphics"
Perceived degree of interaction	Baños et al., 2000. See "Immersion/Presence"
	Ibili & Billinghurst, 2019. See "Perceived cognitive load"
Motivation to learn	Fu et al., 2009. See "Perceived cognitive load"
	Gunawardena, C. N., & Zittle, F. J. (1997). Social presence as a predictor of satisfaction within a computer-mediated conferencing environment. *American Journal of Distance Education, 11*(3), 8–26. https://doi.org/10.1080/08923649709526970
	Högberg et al., 2019. See "Perceived feedback and content quality"
	Keller, 2006. See "Cognitive load"
Motivation to use the virtual environment	Chen, M., Kolko, B. E., Cuddihy, E., & Medina, E. (2011). Modeling but NOT measuring engagement in computer games. *Proceedings of the* 7th *International Conference on Games + Learning + Society Conference*, 55–63. ACM.
	Högberg et al., 2019. See "Perceived feedback and content quality"
	Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., & Walton, A. (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
	Parnell, 2009. See "Perceived quality of the virtual environment's graphics "
Perceived usefulness/knowledge gains	Chou & & Liu, 2005. See "Learning enjoyment"
	Fu et al., 2009. See "Perceived cognitive load"
	Gunawardena & Zittle, 1997. See "Motivation to learn"
	Ibili & Billinghurst, 2019. See "Perceived cognitive load"
	Keller, 2006. See "Cognitive load"
	Selwyn, 1997. See "Perceived ease of use/Control of the virtual environment"
	Vorderer et al., 2004. See "Immersion/Presence"
Simulator sickness	Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G., 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology, 3*(3), 203–220. https://doi.org/10.1207/s15327108ijap0303_3
	Kim, H. K., Park, J., Choi, Y., & Choe, M. (2018). Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment. *Applied Ergonomics, 69*, 66–73. https://doi.org/10.1016/j.apergo.2017.12.016
	Slater, 1999. See "Perceived ease of use/Control of the virtual environment"
Positive/Negative feelings	Bernhaupt, R., & Pirker, M. (2013). Evaluating user experience for interactive television: towards the development of a domain-specific user experience questionnaire. *Proceedings of the IFIP Conference on Human-Computer Interaction*, 642–659). IFIP. https://doi.org/10.1007/978-3-642-40480-1_45
	Boletsis, 2020. See "Immersion/Presence"
	IJsselsteijn et al., (2013. See "Perceived quality of the virtual environment's graphics "
	Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology, 36*(1), 36–48. https://doi.org/10.1016/j.cedpsych.2010.10.002
	Spielberger, C.D., Gorsuch, R.L. & Lushene, R.E. (1970). *Manual for the StateTrait Anxiety Inventory (self evaluation questionnaire)*. Consulting Psychologists Press.
	Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine, 166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092
	Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063. https://doi.org/10.1037/0022-3514.54.6.1063
Relevance to personal interests	Keller, 2006. See "Cognitive load"
	Vorderer et al., 2004. See "Immersion/Presence"

Appendix II

The scale's final version.

Factor	Item
Perceived quality of the virtual environment's graphics	The app was aesthetically pleasing
	I enjoyed the app's graphics
	The app was visually appealing
	I enjoyed the virtual environment
	The graphics of the app were attractive
	I was satisfied with the app's graphics
Perceived cognitive load	The cognitive load of the application was reasonable
	The presentation of too much information prevented the memorization of what was important*
	The effort to study the information that the application presented to me, was mentally tiring*
Perceived ease of use/control of the virtual environment	I used/controlled the app with ease
	I had full control over what I did
	When using the app, I had no problems doing whatever I wanted
Immersion/Presence	I immersed myself in the app
	I forgot/ignored everything around me

(*continued on next column*)

(*continued*)

Factor	Item
	I lost the sense of where I am
	I felt like I was inside the virtual world
	I lost the sense of time
	I felt like I was living in another place and time
Perceived feedback and content quality	Overall, the learning content was well presented
	The app gave me useful feedback regarding what I had to do
	The information provided by the app (e.g., objectives, help messages, images, texts, and audio) was clear and understandable
Perceived degree of interaction	I could interact a lot with the virtual world
	The virtual world responded well to my actions
	The interactions with the virtual objects were similar to the interactions with real objects
Motivation to learn and use the virtual environment	I want to know more about what I saw in the app
	I enjoyed the content so much that I would like to know more about this topic
	The content had things that triggered my curiosity
	I feel motivated to keep using the app
	I was intrigued to see what was in the app
	I wanted to explore the app more
Perceived usefulness/knowledge gains	I understood the basic ideas/issues presented to me within the app
	I learned through the app
	The content increased my knowledge and understanding of the subject presented by the app
Simulator sickness (To what degree you felt …)	Dizziness?*
	Your head being "heavy"?*
	Vertigo?*
	A general discomfort?*
	Nausea?*
	Headache?*
Positive feelings (To what degree you felt …)	Joy?
	Satisfaction?
	Enthusiasm?
	Excitement?
Negative feelings (To what degree you felt …)**	Bored?*
	Disappointment ?*
	Nervous?*
	Stressed?*
	Irritated?*

Notes. * = reverse code this item; ** = this factor was not included in the final version but can be considered for future inclusion; all items are presented using a five-point Likert-type scale.

References

Alfaisal, R., Hashim, H., & Azizan, U. H. (2022). Metaverse system adoption in education: A systematic literature review. *Journal of Computers in Education, 2022*, 1–45. https://doi.org/10.1007/s40692-022-00256-6

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573. https://doi.org/10.1007/BF02293814

Atsikpasi, P., & Fokides, E. (2022). A scoping review of the educational uses of 6DoF HMDs. *Virtual Reality, 26*(1), 205–222. https://doi.org/10.1007/s10055-021-00556-9

Ball, M. (2021). *Framework for the metaverse*. https://www.matthewball.vc/all/forwardtothemetaverseprimer.

Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural modeling. *Sociological Methods & Research, 16*(1), 78–117. https://doi.org/10.1177/0049124187016001004

Bertrand, J., Bhargava, A., Madathil, K. C., Gramopadhye, A., & Babu, S. V. (2017). The effects of presentation method and simulation fidelity on psychomotor education in a bimanual metrology training simulation. In *Proceedings of the 2017 IEEE symposium on 3D user interfaces* (Vol. 2017, pp. 59–68). 3DUI. https://doi.org/10.1109/3DUI.2017.7893318.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6*(149), 1–18. https://doi.org/10.3389/fpubh.2018.00149

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation, 10*(7), 1–9.

DeVellis, R. F. (2016). *Scale development: Theory and applications* (Vol. 26). Sage publications.

Doug, A. (2020). *The technology of the Metaverse, It's not just VR*. https://medium.com/swlh/the-technology-of-the-metaverse-its-not-just-vr-78fb3c603fe9.

Fabola, A., & Miller, A. (2016). Virtual reality for early education: A study. In *Proceedings of the international conference on immersive learning* (Vols. 59–72). Springer. https://doi.org/10.1007/978-3-319-41769-1_5

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780199734177.001.0001

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock, & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 269–314). IAP.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.

Hair, J. F., LDS Gabriel, M., Silva, D. D., & Braga, S. (2019). Development and validation of attitudes measurement scales: Fundamental and practical aspects. *RAUSP Management Journal, 54*, 490–507. https://doi.org/10.1108/RAUSP-05-2019-0098

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science, 43*(1), 115–135. https://doi.org/10.1007/s11747-014-0403-8

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141–151. https://doi.org/10.1177/001316446002000116

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford Press.

Kye, B., Han, N., Kim, E., Park, Y., & Jo, S. (2021). Educational applications of Metaverse: Possibilities and limitations. *Journal of Educational Evaluation for Health Professions, 18*, 1–13. https://doi.org/10.3352/jeehp.2021.18.32

Lee, L. H., Braud, T., Zhou, P., Wang, L., Xu, D., Lin, Z., … Hui, P. (2021). *All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. arXiv preprint arXiv:2110.05352*.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Malhotra, N. K., & Dash, S. (2011). *Marketing research an applied orientation*. Pearson Publishing.

Matsunaga, M. (2010). How to factor-analyze your data right: do's, don'ts, and how-to's. *International Journal of Psychological Research, 3*(1), 97–110. https://doi.org/10.21500/20112084.854

Newbutt, N., Bradley, R., & Conley, I. (2019). Using virtual reality head-mounted displays in schools with autistic children: Views, experiences, and future directions. *Cyberpsychology, Behavior, and Social Networking, 2019*. https://10.1089/cyber.2019.0206.

Nikolovska, H. (2022). *December 13). 31 Metaverse statistics to prepare you for the future*. https://www.banklesstimes.com/metaverse-statistics/.

Norris, A. E., & Aroian, K. J. (2004). To transform or not transform skewed data for psychometric analysis: That is the question. *Nursing Research, 53*(1), 67–71. https://doi.org/10.1097/00006199-200401000-00011

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers, 32*, 396–402. https://doi.org/10.3758/BF03200807

Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale Big-Five assessments. *Journal of Research in Personality, 59*, 56–68. https://doi.org/10.1016/j.jrp.2015.09.001

Pirker, J., Lesjak, I., Parger, M., & Gütl, C. (2018). An educational physics laboratory in mobile versus room scale virtual reality-A comparative study. In *Online engineering & Internet of things* (pp. 1029–1043). Springer. https://doi.org/10.1007/978-3-319-64352-6_95.

Queiroz, A. C. M., Nascimento, A. M., Tori, R., & da Silva Leme, M. I. (2018). Using HMD-based immersive virtual environments in primary/K-12 education. In *Proceedings of the international conference on immersive learning* (pp. 160–173). https://doi.org/10.1007/978-3-319-93596-6_11. Springer.

Raubenheimer, J. (2004). An item selection procedure to maximize scale reliability and validity. *SA Journal of Industrial Psychology, 30*(4), 59–64. https://doi.org/10.4102/sajip.v30i4.168

Rauch, W. (1979). The decision delphi. *Technological Forecasting and Social Change, 15*(3), 159–169. https://doi.org/10.1016/0040-1625(79)90011-8

Rupp, M. A., Odette, K. L., Kozachuk, J., Michaelis, J. R., Smither, J. A., & McConnell, D. S. (2019). Investigating learning outcomes and subjective experiences in 360-degree videos. *Computers & Education, 128*, 256–268. https://doi.org/10.1016/j.compedu.2018.09.015

Shi, Y., Du, J., & Worthy, D. A. (2020). The impact of engineering information formats on learning and execution of construction operations: A virtual reality pipe maintenance experiment. *Automation in Construction, 119*, 1–18. https://doi.org/10.1016/j.autcon.2020.103367

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Pearson Education Inc.

Tlili, A., Huang, R., Shehata, B., Liu, D., Zhao, J., Metwally, A. H. S., et al. (2022). Is Metaverse in education a blessing or a curse: A combined content and bibliometric analysis. *Smart Learning Environments, 9*(1), 1–31. https://doi.org/10.1186/s40561-022-00205-x

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin, & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 41–71). Kluwer Academic. https://doi.org/10.1007/978-1-4615-4397-8_3.

Wang, L., Fan, X., & Willson, V. L. (1996). Effects of nonnormal data on parameter estimates and fit indices for a model with latent and manifest variables: An empirical study. *Structural Equation Modeling: A Multidisciplinary Journal, 3*(3), 228–247. https://doi.org/10.1080/10705519609540042

Wilkinson, M., Brantley, S., & Feng, J. (2021). A mini review of presence and immersion in virtual reality. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 65, pp. 1099–1103). SAGE Publications. https://doi.org/10.1177/1071181321651148. No. 1.

Witmer, B. G., Jerome, C. J., & Singer, M. J. (2005). The factor structure of the presence questionnaire. *Presence: Teleoperators and Virtual Environments, 14*(3), 298–312. https://doi.org/10.1162/105474605323384654

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence, 7*(3), 225–240. https://doi.org/10.1162/105474698565686.