Classifying Presence Scores: Insights and Analysis from Two Decades of the Igroup Presence Questionnaire (IPQ)

TANH QUANG TRAN, TOBIAS LANGLOTZ, and JACOB YOUNG, University of Otago, Dunedin, New Zealand
THOMAS W. SCHUBERT, University of Oslo, Oslo, Norway and Instituto Universitário de Lisboa (ISCTEIUL)/CIS-IUL, Lisbon, Portugal
HOLGER REGENBRECHT, University of Otago, Dunedin, New Zealand

Presence, or the experience of being present in a computer-generated environment, is a defining element of virtual reality. While there are different methodologies to measure presence, questionnaires remain the most popular, particularly the Igroup Presence Questionnaire (IPQ). In this article, we analyse the results of over 20 years of IPQ usage to develop a new comparative means of reporting presence scores and comparing them across existing and future work. We additionally report on correct and problematic usage of the questionnaire and, through this, present guidelines on how to administer the IPQ in future to aid further analysis. Finally, we present a new web-based tool to streamline the analysis and reporting of IPQ results, which we hope will facilitate more standardised usage of the questionnaire in future research.

CCS Concepts: • **General and reference** → **Surveys and overviews**; **Measurement**; • **Human-centered computing** → **User studies**; **Virtual reality**; *Usability testing*; • **Computing methodologies** → **Mixed / augmented reality**; **Perception**; • **Applied computing** → **Computer games**;

Additional Key Words and Phrases: presence, usability, virtual reality, perception, empirical studies, survey, ranking scale, meta analysis

ACM Reference format:
Tanh Quang Tran, Tobias Langlotz, Jacob Young, Thomas W. Schubert, and Holger Regenbrecht. 2024. Classifying Presence Scores: Insights and Analysis from Two Decades of the Igroup Presence Questionnaire (IPQ). *ACM Trans. Comput.-Hum. Interact.* 31, 5, Article 61 (November 2024), 26 pages.
https://doi.org/10.1145/3689046

1 Introduction

At the heart of **Virtual Reality (VR)** is presence: the feeling of "being there" in the virtual environment [35], which is widely assumed to be a prerequisite for successfully using virtual

This work was partially supported by an MBIE Endeavour Research Program Grant (UOOX2308).
Authors' Contact Information: Tanh Quang Tran, University of Otago, Dunedin, New Zealand; e-mail: tqtanh@outlook.com; Tobias Langlotz (corresponding author), University of Otago, Dunedin, New Zealand; e-mail: tobias.langlotz@otago.ac.nz; Jacob Young, University of Otago, Dunedin, New Zealand; e-mail: jacob.young@otago.ac.nz; Thomas W. Schubert, University of Oslo, Oslo, Norway and Instituto Universitário de Lisboa (ISCTE-IUL)/CIS-IUL, Lisbon, Portugal; e-mail: schubert@igroup.org; Holger Regenbrecht, University of Otago, Dunedin, New Zealand; e-mail: holger.regenbrecht@otago.ac.nz.

ACM 1557-7325/2024/11-ART61
https://doi.org/10.1145/3689046

environments. Technologically, VR is constructed using computer-generated, three-dimensional environments that can be explored by users in interactive real-time. However, this sense of presence is what sets a VR experience apart from traditional 3D interfaces. Or, in short: without presence there is no VR.

While different approaches have been developed to measure this sense of presence, the administration of post-experience questionnaires is by far the most popular as they are easy to administer, and their analysis requires standard statistics. However, individual measurements obtained with these questionnaires are difficult to compare across different studies without some kind of objective rating scale, such as used for the **System Usability Scale (SUS)** [4]. This limitation is also true for the **Igroup Presence Questionnaire (IPQ)**, which enjoys increasing popularity within VR [56], Telepresence [42] and Human–Computer interaction research [58] as a way to evaluate immersive experiences [56, 58]. This makes comparing new systems against existing work impractical, especially in industry or field settings where A/B tests aren't always possible.

In this article we analyse 1,771 papers that utilised the IPQ in order to develop an effective way to compare results across studies. We identified 243 studies that reliably reported their results; this is usually given in the form of an aggregated, averaged IPQ presence score, sometimes supported with sub-scale scores or additional statistical information.

One particular issue encountered is the inconsistent quality of reporting IPQ scores, including incomplete datasets and non-standardised questionnaire administration. These inconsistencies prevent application of a traditional meta-analysis; however, it is still useful to place presence scores into context. We have thus compiled the results of these studies into a distribution curve and developed a ranking scale based on percentiles of this distribution.

Using this distribution, we define five classes to cover the range of scores possible; for a convenient and quick contextualisation of a calculated mean presence score, one thus only has to refer to Table 4 and Figure 4 of this article. For the first time, researchers can easily compare presence measures obtained with the IPQ against existing and future work. Note that this can now be applied not only to multi-condition studies in the lab but also to single-factor evaluations such as those often seen in the field.

In addition to providing a distribution for overall presence scores, we provide distributions for the three sub-scales of the IPQ: **spatial presence (SP)**, **involvement (INV)**, and **experienced realism (REAL)**, as well as for the single-item score for **general presence (GP)**. Finally, because the immersive characteristics vary across studies, we also analysed the display modalities of all studies and provided the distribution curves for three commonly used display types: head-mounted displays, monitors, and projection systems. Hence, in the same way as with the overall presence score, authors can compare their measures against previous work.

We have made the findings of this work available online through a new web-based tool, *IPQ Cal*,[1] which provides a detailed overview of existing IPQ results. Researchers can also upload the results of studies using the IPQ and have all relevant analyses performed for them; the distribution curve used for analysis will be updated to reflect the new data, meaning that the classes we have identified will adapt to future experiments. Finally, IPQ Cal will also produce a report researchers can use to disseminate IPQ results. With this tool, we aim to (a) make the analysis and reporting of IPQ data more consistent and convenient, (b) dynamically update the IPQ scoring system, (c) collect and provide raw data on IPQ measures, and finally, (d) conduct a meta-analysis of presence scores in the future for more robust, comparative measures.

[1] https://hci.otago.ac.nz/ipqcal/

In summary, our contributions are:

—An analysis of prior IPQ use in the literature and recommendations for more consistent reporting in the future;
—The development and provision of an IPQ-based presence score with an associated rating scale, including sub-scales for presence and visual display characteristics;
—Guidelines on how to administer and report the IPQ in a standardised manner to facilitate future analysis;
—A web-based tool to calculate, compare, and report on presence measures across studies to encourage simpler and more consistent usage and analysis of the IPQ in future studies.

2 Background

The sense of presence has been widely studied during the past few decades, though what exactly is meant by "presence" is still a point of debate. Within communication studies, presence is commonly defined as the "perceptual illusion of non-mediation" while amongst VR scholars, presence is more commonly denoted as the sense of being in the mediated environment or simply the sense of "being there." With those different definitions and conceptualisations come different ways to measure the sense of presence, including observations, physiological measures, and self-reporting through questionnaires or interviews.

It is important to note the distinction between presence and immersion. There are several opinions on the difference between these two concepts: Witmer and Singer [76] consider them as one and the same, while Slater [61] and others consider them as distinct, with presence as a subjective experience of the user and immersion a technological aspect of the system. We follow Slater et al.'s definition: immersion, as the richness and "surroundedness" of the technical environment, can be objectively described by specifications or technical measures. The sense of presence, as a subjective experience, cannot be so easily quantified.

Despite its subjective nature, presence questionnaires have been a large focus since the early days of VR research, not just for the evaluation of different VR systems and techniques but also to better understand the concept of VR itself. Early instances stemmed from pioneering work by Barfield and Weghorst [5], Sheridan [59], and Steuer [62]. When only considering the number of citations, not analysing the context of their citation, the most popular instruments are the ones developed by Witmer and Singer [75, 76], Usoh et al. [70], Lessiter et al. [33], and Schubert et al. [56]. Schwind et al. [58] produced an excellent tabular overview of presence questionnaires which we present in Table 1 in an amended and updated form.

The Igroup Presence Questionnaire has remained popular since it was first introduced over 20 years ago, with over 2,000 citations as of the time of writing, and has recently been recommended as the presence measure of choice due to its high reliability [58]. It is a compact and efficient instrument and, according to Schwind et al. [58], takes about 2.5 minutes on average to be answered—less than half the time needed for Witmer and Singer's scale [75, 76].

The IPQ is based on the model that two main cognitive processes lead to a sense of "being there": (1) possible (bodily) actions in the environment and (2) the suppression of incompatible sensory input. The possible actions process was mainly based on Glenberg's concept of "meshed sets of patterns of actions" [22] and the suppression process, including the necessary "suspension of disbelief," built on Bystrom et al.'s [8] work on allocation of attention to virtual stimuli.

The questionnaire is designed to evaluate these two factors of presence as a psychological experience and provide a score on a continuous scale for how intensely this sensation was experienced. The questionnaire is composed of 14 items, each presented as a 7-point Likert-type scale to denote

Table 1. Overview and Comparison of 16 Published Presence Questionnaires

Authors	Year	Citations	Items
Banos et al. [3]	1998	319	77
Barfield and Weghorst [5]	1995	378	5+1
Cho et al. [17]	2003	47	4
Dinh et al. [16]	1999	725	13+1
Gerhard et al. [21]	2001	125	19+4
Hartman et al. [24]	2015	228	20
Kim and Biocca [30]	1997	1,106	8
Krauss et al. [31]	2001	21	42
Lombard and Ditton [36]	2000	403	103
Lombard and Weinstein (TPI) [37]	2009	393	4–8
Lessiter et al. (ITC-SOPI) [33]	2001	1,502	44
Nichols et al. [43]	2000	314	9
Nowak and Biocca [44]	2003	1,204	9
Schubert et al. (IPQ) [50, 56, 57]	**2001**	**2,697**	**14**
Usoh/Slater et al. (SUS) [60]	1994	1,692	3/6
Witmer and Singer (WS) [76]	1998	7,473	32

Original papers describing the IPQ are set in bold. Corrected and updated version with data from the end of 2023 (original version by Schwind et al. [58]).

disagreement or agreement with a statement about the participant's experience in the virtual environment. These items fall into three sub-scales: SP, INV, and REAL, plus one GP question. Together, these sub-scales indicate the participant's experience of presence in the virtual environment. A full list of items and their assignment to sub-scales in the IPQ is presented in Table 2.

Currently, researchers can use presence measurements to compare conditions or correlate them with other data. Both applications are relative: one condition relative to another or one individual relative to another. However, we are lacking a solution to assess and observe where the reported values are located in the general distribution of presence scores across the academic literature. If available, we could answer questions about the observed values such as: "Did manipulations in the study lead to values that are at the upper end of the distribution?" or "Did a condition destroy presence in such that it dropped to the very bottom?". In a correlational study, one might ask: "Did the individuals vary across the whole spectrum of the potential distribution?" or "Did they cluster in a small segment?". Vitally, one could finally answer the question: "How well has my system invoked a sense of presence, and how well does this compare to other systems?"

A common approach for this type of comparative rating is to construct a percentile scheme, which is widely used to analyse and classify data in various fields and applications. For example, Falker et al. [19] applied percentiles to categorise different levels blood pressure and determine when it might be considered "high," identifying the 90th and 95th percentiles as thresholds to determine normal blood pressure, prehypertension, and hypertension, respectively. Percentiles can likewise be used to classify body mass index into different categories [73]; for children's weight, the 95th, 85th and 5th percentiles are assigned as cut-off points for weight classes: obese, overweight, normal weight, and underweight [14]. In bibliometric studies, there are four different schemas used to classify bibliometric data [6]: two two-class percentile schemas and two six-class percentile schemas. The first two-class schema uses the 90th percentile as the threshold value, while the

Table 2. The Igroup Presence Questionnaire (IPQ) with Sub-Scales: Spatial Presence (SP), Involvement (INV), Experienced Realism (REAL), and General Presence (GP)

Sub-scale	Question
GP1	In the computer generated world, I had a sense of "being there". `fully disagree [−3, . . . , 3] fully agree`
SP1	Somehow I felt that the virtual world surrounded me. `fully disagree [−3, . . . , 3] fully agree`
SP2	I felt like I was just perceiving pictures. `fully disagree [−3, . . . , 3] fully agree`
SP3	I did *not* feel present in the virtual space. `did not feel [−3, . . . , 3] felt present`
SP4	I had a sense of acting in the virtual space, rather than operating something from outside `fully disagree [−3, . . . , 3] fully agree`
SP5	I felt present in the virtual space `fully disagree [−3, . . . , 3] fully agree`
INV1	How aware were you of the real world surrounding while navigating in the virtual world? (i.e. sounds, room temperature, other people, etc.)? `extremely aware [−3, . . . , moderately aware, . . . , 3] not aware at all`
INV2	I was not aware of my real environment. `fully disagree [−3, . . . , 3] fully agree`
INV3	I still paid attention to the real environment. `fully disagree [−3, . . . , 3] fully agree`
INV4	I was completely captivated by the virtual world. `fully disagree [−3, . . . , 3] fully agree`
REAL1	How real did the virtual world seem to you? `completely real [−3, . . . , 3] not real at all`
REAL2	How much did your experience in the virtual environment seem consistent with your real world experience? `not consistent [−3, . . . , moderately consistent, . . . , 3] very consistent`
REAL3	How real did the virtual world seem to you? `about as real as an imagined world [−3, . . . , 3] indistinguishable from the real world`
REAL4	The virtual world seemed more realistic than the real world. `fully disagree [−3, . . . , 3] fully agree`

Items highlighted gray need reversing (multiplied by −1) before being combined with the other items to calculate the total score for the questionnaire.

second schema uses the 50th percentile. One six-class percentile schema was proposed by Thomson Reuters' Essential Science Indicators. The thresholds for the classes are the 50th, 80th, 90th, 99th, 99.9th and 99.99th percentiles. In contrast, Bornmann and Mutz [7] proposed a six-class percentile schema with thresholds for classes at the 50th, 75th, 90th, 95th and 99th percentiles.

In the following, we present our work analysing usage and reported results of the IPQ to create such a percentile scheme, thereby opening a pathway for providing classifications and baselines when assessing presence using the IPQ.

3 Method

We start our analysis of the IPQ by reviewing past studies found in the academic literature through the PRISMA methodology [45]. This initial step also helps us to get a better understanding of how the IPQ has been used. Refer to Figure 1 for a flowchart depicting our publication aggregation process.

3.1 Identification of Relevant Publications

In the first step, we identified all publications that cite or refer to the IPQ. To achieve this, we used Harzing's Publish or Perish tool,[2] which uses data from sources such as Google Scholar and Microsoft Academic Search, to facilitate complex queries and analysis. After some initial tests, we identified and used the following keywords for our search: "Igroup Presence Questionnaire," "I-Group Presence Questionnaire," "IPQ," "Presence," and "Schubert". We identified 2208 publications using these keywords. Besides publications that referenced the original paper describing the IPQ, we also found papers that cited other research using the IPQ. Similarly, apart from the original paper, several other papers published by the original IPQ authors that described the prior studies and the design of the IPQ [50, 57] were also occasionally mentioned.

After manually reviewing all 2,208 publications, we excluded 222 that appeared in multiple searches. We also excluded 215 papers not written in English as the IPQ is also available in German, French, Dutch, Portuguese, and Japanese. This left us with 1,771 publications for further screening.

3.2 Exclusion Criteria

In our analysis, we are only interested in publications in which it is clear that the IPQ was utilised in a study to measure presence. Subsequently, we excluded:

- —713 papers which only mention the questionnaire without actually using it
- —22 publications which report study results but do not make it clear whether the IPQ was administered
- —One paper which provided a Portuguese translation of the IPQ
- —The original paper from Schubert et al. defining the questionnaire.

One hundred five additional publications were excluded as we could not access the original document. This left 929 publications in which the IPQ had been used to measure the experience of presence. In selecting a subset for our analysis, we took a very conservative approach to ensure that our data would be reliable. We thus excluded:

- —20 studies that were reported in multiple papers, in which case we kept the original paper and removed the duplicates.
- —249 papers that only included some items or sub-scales of the questionnaire instead of using it in its entirety.

[2] https://harzing.com/resources/publish-or-perish

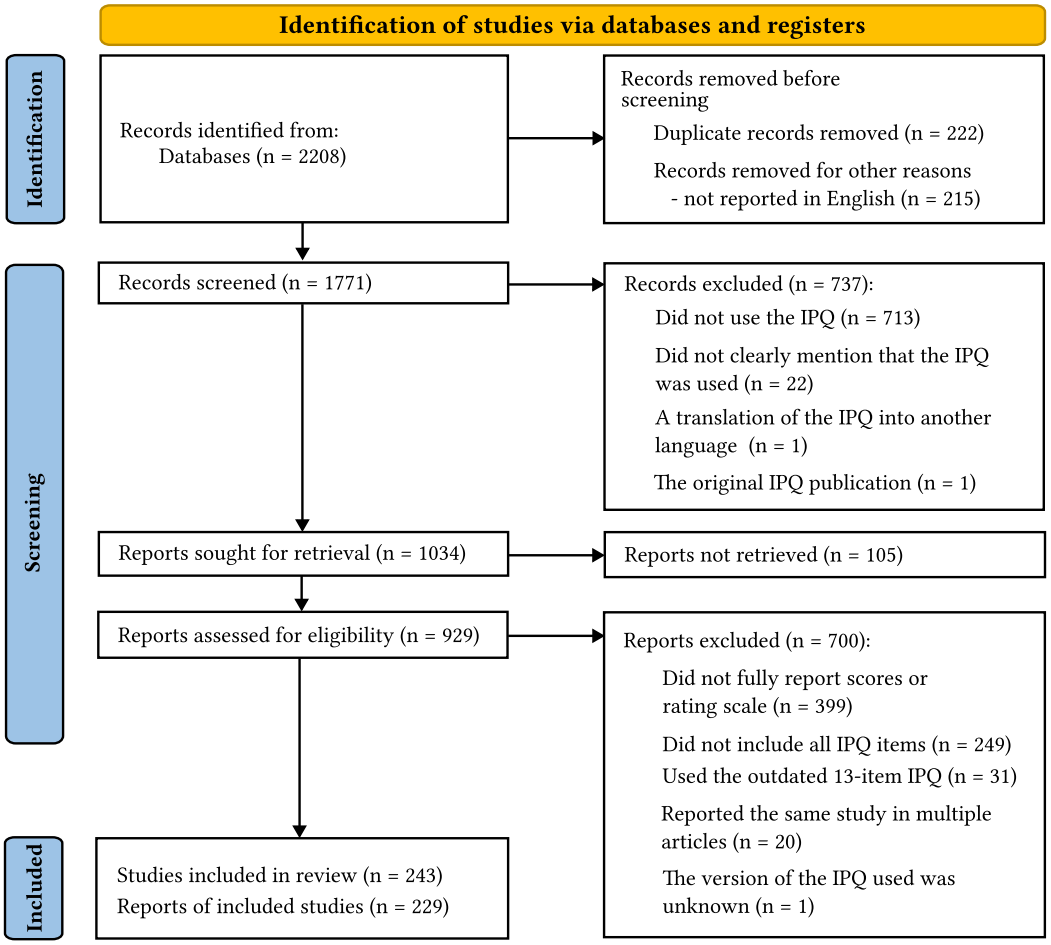

Fig. 1. PRISMA flowchart illustrating our publication aggregation procedure.

—399 papers that used the IPQ but did not fully report their results or did not mention which rating scale was applied for the questionnaire. We focused on collecting mean and standard deviation statistics for presence scores, so papers reporting only median scores were excluded. We also removed papers where the necessary statistical estimates could not be extracted.
—31 papers that used an older 13-item version of the IPQ instead of the official and most popular version with 14 items.
—One paper whose version of the IPQ could not be verified.

3.3 Included Publications

These exclusions left us with 229 suitable papers for our analysis. These reports are on 243 individual studies, with 12 papers containing more than one study. Between them, these studies involved a total of 9,354 participants who completed the IPQ. The selected papers report on studies conducted between the years 2003 and 2023, with an increasing number in more recent years (see Figure 2).

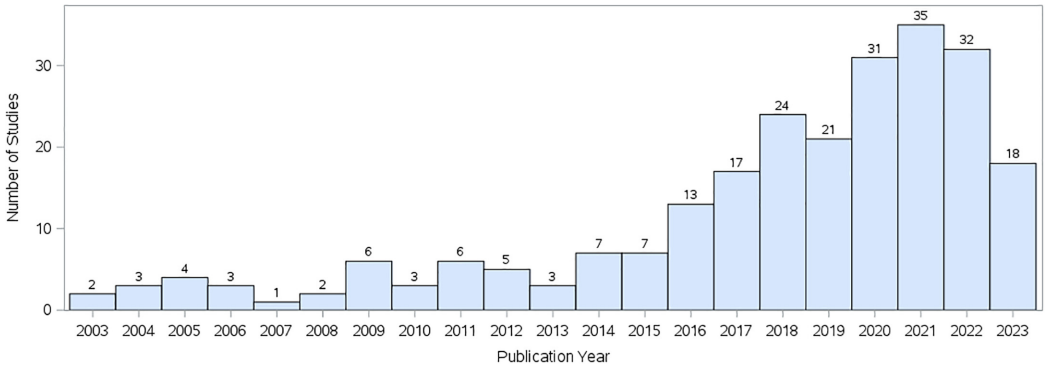

Fig. 2. Number of selected studies published per year from 2002 to Q3 2023.

The papers we identified have several characteristics that are beneficial for our analysis. Each reported on at least one unique study and minimally reported the means of scores for the questionnaire and/or its sub-scales and/or its items. In addition, these papers report on the rating scales used for the IPQ, allowing us to transform all reported scores into a common scale. These papers also provided details of experiments with the number of participants for each conducted user study.

4 Initial Observations

A first analysis of the selected papers already allows for some first observations regarding the use of IPQ but also identified inconsistencies when using the IPQ and reporting on the results.

4.1 General IPQ Usage

When looking at the aggregated publications, we can already see a wide range of publication venues ranging from psychology and behavioural science outlets (e.g., Cyberpsychology, Behavior, and Social Networking; Computer in Human Behaviors; CyberPsychology & Behavior; CyberTherapy and Rehabilitation; and Anxiety Disorders) to classic venues for VR research (e.g., Presence: Teleoperators & Virtual Environments, Computers & Graphics, the IEEE Virtual Reality Conference, and the ACM Symposium on Virtual Reality Software and Technology). The IPQ has, in particular, seen increasing usage from the HCI community (e.g., the ACM CHI Conference on Human Factors in Computing Systems).

This wide usage is even more apparent when looking at the application domains that the IPQ has been applied to. We found that the IPQ was a popular measure for evaluating Virtual Reality Exposure Therapy (VRET) (e.g., [41, 64]); here, the IPQ was used to investigate "... whether the virtual world in this experiment established a reasonable level of presence to evoke the anxiety..." [23, p. 7] or "...different components of presence are associated with the experience of fear and treatment response to VRE[T]..." [47, p. 7]. Furthermore, there are studies investigating the impact of VR technology and how users perceive and perform in education and training applications [12, 77]. The IPQ was also utilised in game and entertainment applications (e.g., [34, 46]), social and psychological studies [20, 40], design and visualisation systems [32, 55], and experiments with different interaction techniques [29, 67]).

As a post-study measurement, the IPQ is commonly administered after participants have completed their tasks in the experimental environment, meaning after they have left the virtual environment. While some studies are unclear about when and how the IPQ was used to assess presence

in their experimental procedures, recent papers have started to explore the application of questionnaires within VR or inside the experimental environment [1, 58]. For the IPQ and presence questionnaires in general, the authors have observed advantages in completion time when the questionnaire is administered within VR. However, the key message seems to be that administering the IPQ within does not change the measured feeling of presence but can reduce error variance [58].

When looking at the number of participants using the IPQ, we observe a significant variation between studies. We found that, on average, 38.5 (SD = 34.8) participants responded to the questionnaire in each study. However, in some cases, fewer than ten participants completed the questionnaire [10, 18], while in others, more than 150 participants completed it [2, 78].

Most of the selected studies used the original IPQ 7-point Likert-type scale, but some studies used different versions of scales, e.g., 5-point or 4-point Likert-type scales. Of the 243 studies included in our analysis, we observed the following Likert-type scales in use:

— $[0, 6]$: 118 studies
— $[1, 7]$: 75 studies
— $[-3, 3]$: 30 studies
— $[1, 5]$: 13 studies
— $[1, 6]$: three studies
— $[0, 3]$, $[0, 5]$, $[0, 10]$, $[1, 10]$: one study each

4.2 Identified Inconsistencies in IPQ Usage

One particular issue in analysing IPQ scores is the inconsistency in the use of Likert-type scales. The IPQ was originally designed with a 7-point Likert-type scale in mind, but various alternative scales have been applied to the IPQ since its introduction. Some studies used a scale from 1 to 7 [2, 48], while others adopted a 5-point Likert-type scale, such as in a translation of the IPQ into Portuguese [71]. There are also studies that did not mention the scale used. These inconsistencies can result in misinterpretation of the level of presence experienced when trying to compare different publications and their reported presence scores.

These inconsistencies are also present in how scores are aggregated. Some authors report means and standard deviations of scores for the whole questionnaire [63, 68], while others report values for each item [25, 53]. Median values for records are also sometimes provided [9].

As with several other presence questionnaires, the IPQ treats the concept of presence as a multifactor phenomenon. Consequently, the IPQ is constructed from four components, namely GP, SP, INV, and Realism. According to the original IPQ proposal, each component plays its role in the overall feeling of presence [56]. However, some papers have extracted one or more components or sub-scales of the IPQ in order to evaluate only the component they are interested in [49, 66]). For example, the INV sub-scale was extracted from the IPQ in order to investigate the level of immersion of participants in a study with a VR simulator by Hock et al. [26]. Reinhard et al. [52] examined the impact of SP on the usability of navigation systems. However, by taking only a subset of the questionnaire, one has to acknowledge that what was measured is not presence but only one part of the overall experience, and thus, no direct conclusion with respect to presence can be drawn.

We furthermore identified research that applied the IPQ outside of VR to measure presence in **Mixed Reality (MR)** or **Augmented Reality (AR)** [11, 38, 65]). Presence is relatively well understood in VR but comparatively underexplored in MR or AR. Only a few studies provided measurements for these environments (e.g., [51]), and thus it is tempting to use the IPQ. However, the IPQ has not been validated for use outside of VR, and so one has to be careful with the results and their interpretation [69].

In conclusion, we identified 243 unique studies that utilised and reported the results from the entire IPQ, resulting in presence measures from 9354 participants across multiple disciplines and applications. However, as expected, this broad utilisation also results in usage not necessarily foreseen when the IPQ was originally proposed. While some implementations of the IPQ may not affect its general validity (e.g., administering the questionnaire virtually) or can make the questionnaire more widely available (e.g., by translating it into different languages), unvalidated use, such as isolating individual sub-scales or applying the questionnaire to MR/AR experiences [69] requires further research and any results obtained should be treated with caution.

5 Developing Baselines for Comparatively Reporting Presence Measurements

After looking into the general usage of IPQ, we aimed to establish general baselines and categories for future studies based on reported IPQ presence scores. To provide indicative baselines or absolute scores, we conducted a two-step approach on the 243 studies analysed: (1) we collected the aggregated mean scores of the IPQ for the entire questionnaire and its sub-scales, and (2) we formed distribution curves from the aggregated data and created ranking classification scales.

Based on the aggregated and cleaned data, the next step toward developing an IPQ presence ranking scale is the analysis of prior measurements. More specifically, this section will address how we transform the available data to create distribution curves that we visualise to give a first insight into past presence measurements achieved with the IPQ. We provide details for the chosen approach, which attempted to account for some of the limitations that arise from incomplete data in prior studies.

5.1 Data Distribution

After conservatively selecting 243 individual studies to examine, we extracted the reported mean IPQ scores from each, transformed these scores to a consistent $[-3, 3]$ scale, and created the new distribution of IPQ rating scales shown in Figure 3. Although there are differences in the number of participants for the aggregated studies, not all reported the distribution of their presence measurements and so we weight all means equally when establishing the distribution. From the distribution and results of several normality tests, it can be observed that the collected mean values are not normally distributed (see Table 3).

The data is approximately symmetric ($skewness = 0.18$, $kurtosis = -0.55$); however, the distribution is platykurtic. In addition, the centre of the data distribution is shifted to the positive anchor of the rating scale. The mean ($M = 0.4$) and its confidence intervals (95% CI [0.28, 0.42]) are all higher than the middle value of the scale at "0." On the 7-point Likert-type scale (from −3 to 3), 50% of the overall average presence scores range from 0.01 to 0.72.

Researchers who utilise the IPQ can now use this distribution to determine how their evaluated experiences compare to existing work. If, for example, an overall average presence score of 1.25 is calculated, it can be located between the scores of the 230th and the 231st study in the distribution, starting from the first study at the left end of the distribution. This implies that the score is higher than 94.2% of the scores reported in previous studies, or in other words, it is in the 94th percentile of the collected data.

5.2 Ranking Classification Scale

The data distribution presented in the previous section provides an initial overview of presence measurements, which can be useful for interpreting IPQ results. In this section, we demonstrate

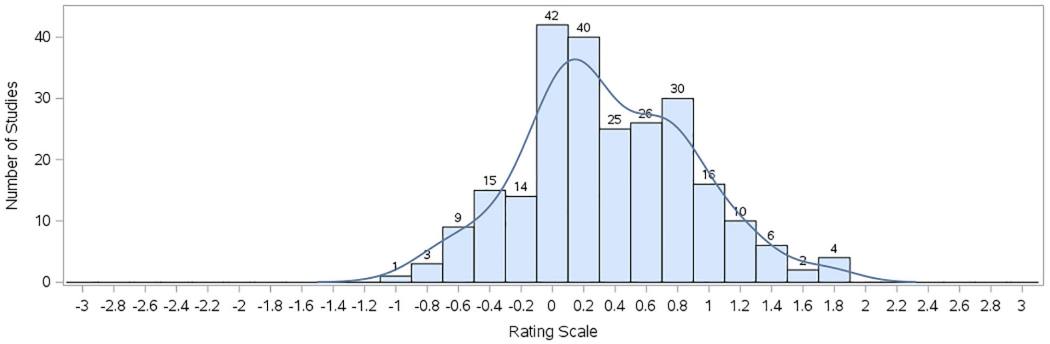

Fig. 3. Distribution with a kernel density estimation curve (bandwidth = 0.19 and asymptotic mean integrated square error (*AMISE*) = 0.01) of aggregated mean values of rating scores for the entire questionnaire from 243 user studies ($M = 0.4$, $SD = 0.5$, $Mdn = 0.3$, $Q_1 = -0.01$, $Q_3 = 0.72$).

Table 3. Results of Normality Tests for the Aggregated Mean Scores from 243 Studies

Test	Statistic	p-value
Shapiro–Wilk	$W = 0.99$	$p = 0.20$
Cramer–von Mises	$W\text{-}Sq = 0.14$	$p = 0.04$
Anderson–Darling	$A\text{-}Sq = 0.68$	$p = 0.08$

how we created a first-ranking scale for IPQ presence measurements using the distribution. Similar to the IPQ itself, which gains some of its popularity due to its ease of use and simplicity, the aim of the ranking scale is to provide an easier and simpler method to interpret and report IPQ presence scores in a constructive way.

When establishing the classification we would ideally like to follow the statistical technique of a meta-analysis. This technique is for integrating or synthesising results from different "independent" and "combinable" studies to quantitative and pooled outcome data [27, 39]. A meta-analysis of the IPQ scores would include results for the questionnaire from studies using this measurement. However, not all of the studies using the IPQ report the scores for each item or for both means and standard deviations for each sub-scale or for the whole questionnaire. For that reason, we applied a specialised approach using only mean values. This parametric method might be biased if distribution assumptions are violated, but it is usually fairly robust. For our analysis, we only considered papers that reported on studies which applied the IPQ in the prescribed way and reported on their results in a way that the results could be replicated. Our specialised approach had to be based on aggregated data of reported IPQ scores instead of actual raw data, which is rare to find in the literature.

To interpret IPQ measures of a given sample in an absolute manner without a specific comparison sample, we suggest comparing them to all previously published IPQ scores, either using the whole questionnaire or specific sub-scales depending on what is being evaluated. Recognising the characteristics of the collected data, we created a classification based on the reported overall average presence scores in each study and organised them into ranked bins (or classes) of percentiles.

Table 4. Overall Ranking Scale for the IPQ Rating Scores with Ranges of Scores and Statistics for Each Ranking Class

Ranking Class	Score Range	Number of Reported Studies	Mean	SD
Exceptional	$[P_{95th}, 3]$ [1.30, 3]	12 (4.9%)	1.53	0.21
Very High	$[P_{90th}, P_{95th})$ [1.07, 1.30)	12 (4.9%)	1.17	0.07
High	$[P_{75th}, P_{90th})$ [0.73, 1.07)	38 (15.6%)	0.86	0.09
Moderate	$[P_{50th}, P_{75th})$ [0.28, 0.73)	60 (24.7%)	0.51	0.14
Low	$[-3, P_{50th})$ [−3, 0.28)	121 (49.8%)	−0.08	0.30

We propose the following class names and percentile thresholds based on the schema by Bornmann and Mutz [7]:

—*Exceptional* : Above the 95th percentile $[P_{95\text{th}}, 3]$
—*Very High* : Above the 90th percentile $[P_{90\text{th}}, P_{95\text{th}})$
—*High* : Above the 75th percentile $[P_{75\text{th}}, P_{90\text{th}})$
—*Moderate* : Above the 50th percentile $[P_{50\text{th}}, P_{75\text{th}})$
—*Low* : Below the 50th percentile $[-3, P_{50\text{th}})$

There are various methods to calculate percentiles or 100-quantiles (or quantiles in general) for a data sample. Hyndman and Fan [28] presented nine different definitions of quantiles with three discontinuous functions and six linear continuous functions. These definitions and functions are widely implemented for calculating percentiles and quantiles in statistical packages. Wicklin [74] compared these definitions for calculating sample quantiles in SAS. Among the definitions, we chose the sixth definition (R-6) of sample quantiles for calculating percentiles for our ranking scale.[3] This definition is based on linear interpolation and was also presented by Weibull [72].

Table 4 presents an overview of reported scores from the aggregated user studies on the IPQ, while Figure 4 illustrates the ranking scale with classified categories. This provides, for the first time, different baselines or absolute measures for the sense of presence that can be used to compare future IPQ presence scores with past studies and their reported presence scores.

5.3 Visual Display Modalities

Researchers and practitioners often want to know the effects of immersion (technological surroundings) factors on presence, and so we were also interested in investigating the impact and contribution of different visual display modalities.

[3]When there are n non-missing values in the order $x_1, x_2, \ldots, x_n$ for an evaluated variable in the sample, where y is the kth percentile and quantile $p = \frac{k}{100}$. As a result, the percentiles at 50th, 75th, 90th, 95th and 99th are 0.5, 0.75, 0.9, 0.95 and 0.99 quantiles, respectively. Set $(n+1)p = j + g$, where $j = \lfloor (n+1)p \rfloor$ and $g = (n+1)p - j$. The percentile y is computed as in the equation below where x_n is applied for x_{n+1} (where x_j is the jth order statistics.): $y = (1-g)x_j + gx_{j+1}$.

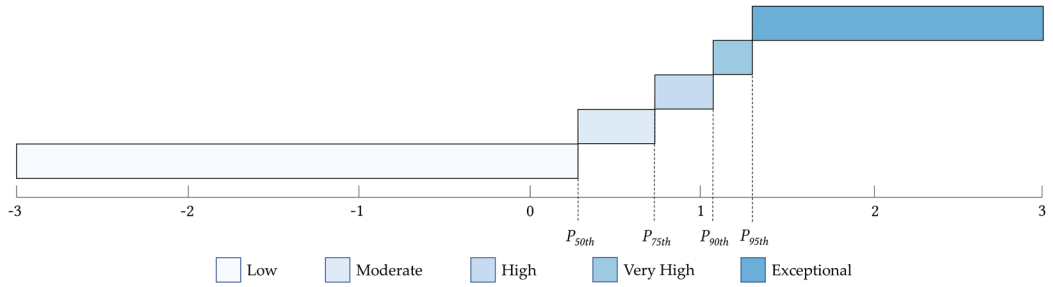

Fig. 4. Illustration of our ranking scale with the representations for each class: Low, Moderate, High, Very High, and Exceptional, and its range on the corresponding 7-point Likert-type scale.

Table 5. Descriptive Statistics of Rating Scores for Each Type of the Visual Displays

Visual Display	Mean	SD	P_{50th}	P_{75th}	P_{90th}	P_{95th}
3D-HMD	0.5	0.5	0.40	0.83	1.17	1.35
3D-Monoscopic	−0.2	0.6	−0.10	0.17	0.47	0.68
Projection Display	0.2	0.5	0.23	0.60	0.73	0.85
Overall Score	0.3	0.6	0.28	0.73	1.05	1.29

Overall Score represents statistics of scores from studies reporting scores for each visual display.

Sixteen of the 237 studies in our sample compared display modalities directly thus, they have IPQ scores for different display types. Consequently, we have 253 IPQ score entries from 223 publications presenting the results of 237 user studies.

Three different display modalities were identified:

—(3D-HMD): **Three-dimensional head-mounted displays (3D-HMD)**, including virtual and MR HMDs such as the HTC Vive and Microsoft Hololens. These were used in 180 user studies.
—(3D-Monoscopic): Monoscopic displays such as desktop monitors or mobile phone screens. These were used in 39 user studies.
—(Projection Display): Displays that are projected onto their surroundings, e.g., CAVEs. These were used in 34 user studies.

Levene's test for homogeneity shows that the variance in presence scores for visual displays met the assumption of homogeneity ($F = 1.49$, $p = 0.23$). We used the obtained Welch's adjusted F-ratio (23.37), which a one-way ANOVA ($\alpha = 0.05$) found to be statistically significant ($p < 0.001$), indicating a significant difference in IPQ rating scores between visual displays. We applied Tukey-Kramer adjustments for multiple comparison tests when there were significant differences between the display characteristics. The test results showed significant differences in IPQ rating scores between 3D-HMD and 3D-Monoscopic ($p < 0.001$), between 3D-HMD and Projection Display ($p = 0.003$), and between 3D-Monoscopic and projection displays ($p = 0.02$). Note that these comparisons were calculated at the sample level.

In general, 3D-HMD had the highest mean presence score ($M = 0.5$, $SD = 0.5$), while 3D-Monoscopic had the lowest average presence score ($M = -0.2$, $SD = 0.6$). Table 5 presents statistics for the visual display modalities, and Figure 5 shows their ranking class ranges.

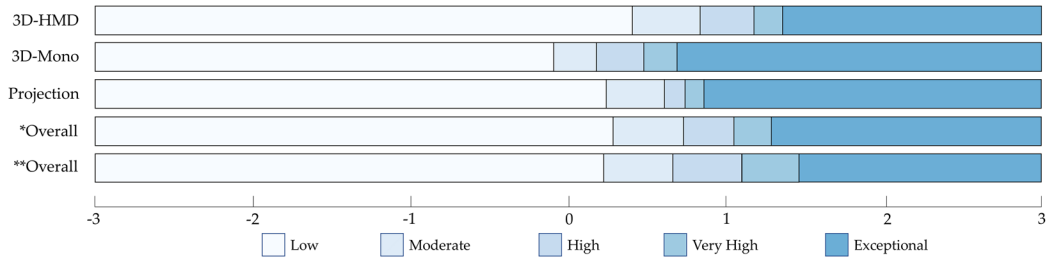

Fig. 5. Distribution of rating scores for each type of visual display. *Overall score from 237 studies reporting scores for each visual display, and **Overall score from all 243 studies. (3DHMD: 3D-HMD, 3DMono: 3D-Monoscopic, and Projection: Projection Display).

5.4 Sub-Scales

We additionally consider different aspects of presence: GP, SP, INV, and REAL. The GP component has only one item which is used as a direct question to assess the extent to which participants feel presence in the experiment; statistical analyses have shown that this sub-scale has strong correlations with other sub-scales in the questionnaire. However, one question is not sufficient to evaluate a multi-dimensional phenomenon such as presence, which is why there are other sub-scales: SP, INV, and REAL. In this section, we present the classification schema for these sub-scales and additional correlations between them, showing a strong connection between the IPQ sub-scales.

After aggregating the data, we handled and analysed the distribution of mean scores for all sub-scales (sub-scales average scores) using SAS software. We found that the distributions of sub-scales mean scores are roughly symmetric ($-1 < skewness < 1$) and (with the exception of REAL) platykurtic with light tails ($kurtosis < 0$). Normality tests using three different methods showed that the mean scores for the REAL sub-scale are not normally distributed, while those for the GP, SP, and INV sub-scales are normally distributed. The skewness value for the SP average scores is closest to 0, while INV has the closest kurtosis value to 0. In addition, the distributions for mean scores of the GP, SP, and INV are shifted towards the positive anchor of the Likert-type scale, with mean scores slightly higher than zero at 0.9, 0.8, and 0.3, respectively. In contrast, the distribution for mean scores of the REAL sub-scale is shifted towards the negative anchor, with a mean of −0.3 (lower than zero).

Table 6 summarises normality test results for sub-scales average scores. Table 7 shows the skewness and kurtosis values and kernel density estimation statistics for the distributions for average scores distributions for the sub-scales, while Figure 6 displays the distributions and their kernel density estimate curves for mean scores of all sub-scales.

Out of the 243 studies, only 162 reported scores for each IPQ sub-scale. Similar to the ranking scale for the whole IPQ score, we assigned the range for the ranking classes of sub-scale average scores based on the percentiles at 50th, 75th, 90th, and 95th. We found that GP has higher rating scores than the other sub-scales, followed by SP. As a result, these sub-scales have higher threshold values than those for INV and REAL. The top 5% of scores for SP are above 1.99, while the threshold for "Exceptional" class of GP is 2.10. In addition, 50% of scores for GP and SP are at or above 1.00 and 0.72, respectively. The thresholds for "Exceptional" and "Low" classes for INV are 1.28 and 0.26, respectively. Among the sub-scales, the average score on REAL is below zero. The range of scores for this sub-scale is between −1.76 and 1.35, while the range between "Low" and "Exceptional" classes for this sub-scale is between −0.37 and 0.95. Overall, Figure 7 and Table 8 present the distribution statistics for these sub-scales from 162 reported studies.

Table 6. Results of Normality Tests for Mean Scores of Sub-Scales from 162 Studies

Sub-scale	Test	Statistic	p-value
General Presence	Shapiro–Wilk	$W = 0.99$	$p = 0.10$
	Cramer–von Mises	$W\text{-}Sq = 0.10$	$p = 0.12$
	Anderson–Darling	$A\text{-}Sq = 0.58$	$p = 0.13$
Spatial Presence	Shapiro–Wilk	$W = 0.99$	$p = 0.50$
	Cramer–von Mises	$W\text{-}Sq = 0.09$	$p = 0.14$
	Anderson–Darling	$A\text{-}Sq = 0.51$	$p = 0.20$
Involvement	Shapiro–Wilk	$W = 0.99$	$p = 0.68$
	Cramer–von Mises	$W\text{-}Sq = 0.05$	$p > 0.25$
	Anderson–Darling	$A\text{-}Sq = 0.31$	$p > 0.25$
Experienced Realism	Shapiro–Wilk	$W = 0.98$	$p = 0.006$
	Cramer–von Mises	$W\text{-}Sq = 0.18$	$p = 0.01$
	Anderson–Darling	$A\text{-}Sq = 1.14$	$p = 0.006$
Overall	Shapiro–Wilk	$W = 0.99$	$p = 0.23$
	Cramer–von Mises	$W\text{-}Sq = 0.11$	$p = 0.09$
	Anderson–Darling	$A\text{-}Sq = 0.57$	$p = 0.14$

Table 7. Descriptive Characterisations of Mean Values for Each Sub-Scale with Kernel Density Estimates with Asymptotic Mean Integrated Square Errors (AMISE) and Bandwidths of Their Distributions

Sub-scale	Kernel Density Estimation		Skewness	Kurtosis
	Bandwidth	AMISE		
Spatial Presence	0.29	0.01	0.10	−0.36
Involvement	0.25	0.01	0.23	−0.05
Experienced Realism	0.18	0.01	0.47	0.54
General Presence	0.33	0.01	−0.20	−0.58
Overall Score	0.21	0.01	0.19	−0.13

Overall Score represents statistics of scores from studies reporting scores for each sub-scale.

We also conducted Spearman's correlation tests to investigate the relationship among the IPQ sub-scales and between the sub-scales' average scores and the overall average presence scores of the questionnaire. Table 9 shows the results of the tests. It can be seen that there were significant strong correlations between the overall and the sub-scales scores: SP ($r_s = 0.88$, $p < 0.001$), INV ($r_s = 0.83$, $p < 0.001$), REAL ($r_s = 0.66$, $p < 0.001$), GP ($r_s = 0.88$, $p < 0.001$).

In addition, we found that GP had a significantly strong correlation with SP ($r_s = 0.77$, $p < 0.001$) and INV ($r_s = 0.64$, $p < 0.001$) sub-scales and a significantly moderate correlation with REAL ($r_s = 0.44$, $p < 0.001$). As described by the IPQ authors [56], GP correlated more strongly with SP than with INV and REAL. Note that these relations were observed at the sample level, meaning that samples with high average levels on one scale also had high average levels on other scales. In contrast, typical analyses within a sample level established such relations on the individual level.

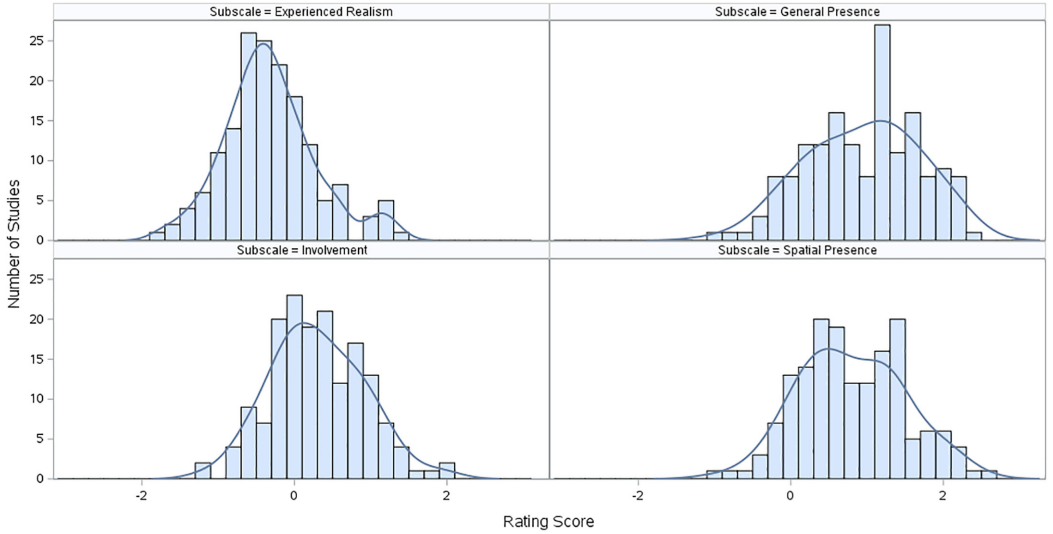

Fig. 6. Distribution with kernel density estimation curves of aggregated mean values of rating scores for sub-scales of the questionnaire: (Bottom Right) SP (bandwidth = 0.29, *AMISE* = 0.01), (Bottom Left) INV (bandwidth = 0.25, *AMISE* = 0.01), (Top Left) REAL (bandwidth = 0.18, *AMISE* = 0.01), and (Top Right) GP (bandwidth = 0.33, *AMISE* = 0.01).

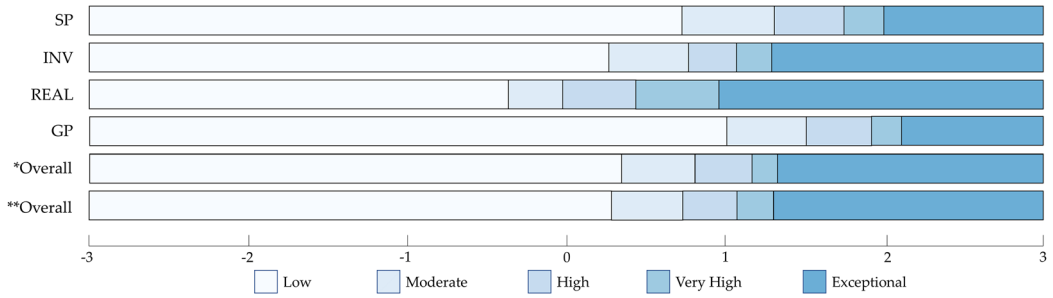

Fig. 7. Ranking scale for each sub-scale of the IPQ. *Overall score from 162 studies reporting sub-scales scores, and **Overall score from all 243 studies. (SP: Spatial Presence, INV: Involvement, REAL: Experienced Realism, GP: General Presence)

Table 8. Descriptive Statistics for Each Sub-Scale in the IPQ from 162 Collected Studies

Sub-scale	Mean	SD	P_{50th}	P_{75th}	P_{90th}	P_{95th}
Spatial Presence	0.8	0.7	0.72	1.30	1.74	1.99
Involvement	0.3	0.6	0.26	0.76	1.06	1.28
Experienced Realism	−0.3	0.6	−0.37	−0.03	0.43	0.95
General Presence	0.9	0.7	1.00	1.50	1.91	2.10
Overall Score	0.4	0.6	0.34	0.80	1.16	1.32

Overall Score represents statistics of scores from studies reporting scores for each sub-scale.

Table 9. Results of Spearman's Correlation Tests between Sub-Scales: Spatial Presence (SP), Involvement (INV), Experienced Realism (REAL), and General Presence (GP) in the IPQ from 162 Studies

	SP		INV		REAL		GP		Overall Score	
	r_s	p	r_s	p	r_s	p	r_s	p	r_s	p
SP	1.00	< 0.001	0.65	< 0.001	0.44	< 0.001	0.77	< 0.001	0.88	< 0.001
INV	0.65	< 0.001	1.00	< 0.001	0.47	< 0.001	0.64	< 0.001	0.83	< 0.001
REAL	0.44	< 0.001	0.47	< 0.001	1.00	< 0.001	0.44	< 0.001	0.66	< 0.001
GP	0.77	< 0.001	0.64	< 0.001	0.44	< 0.001	1.00	< 0.001	0.88	< 0.001
Overall Score	0.88	< 0.001	0.83	< 0.001	0.66	< 0.001	0.88	< 0.001	1	

Overall Score represents mean scores from studies reporting scores for each sub-scale.

6 Discussion

We presented our work on analysing IPQ presence measurements from past studies with the aim of developing IPQ ranking scales that in the future can be used for aiding the interpretation of IPQ presence measurements. However, during the development of our initial IPQ presence distribution curves and the IPQ ranking scales, we had to overcome several limitations in the available IPQ data. In the following, we will provide a critical discussion of the identified limitations and consequently our approach.

6.1 Usage of Data Distributions and Ranking Classification Scales

As mentioned before, IPQ scores are usually calculated as part of comparative or A/B studies. However, with VR increasingly leaving the research labs, A/B studies are not always desired or feasible. In particular, in industry environments or field studies, it is often not possible to introduce a second condition for comparison.

Our approach is addressing this gap. From the distributions of the aggregated scores, users of the IPQ can identify where their scores lie within those distributions and, thus, how their VR experiences compare to existing ones. However, since each user study is conducted with its specific research questions, requirements, and apparatus, we also provide ranking scales for each visual display modality and IPQ sub-scale to cater for more fine-grained comparisons across studies.

There might be effects of visual display modalities on how users perceive the level of presence in virtual environments. 3D HMDs generally provide a higher level of presence as the differences between this display technology and others, including monoscopic 3D and projection displays, are significant. However, the rating scores on presence between 3D Monoscopic and projection displays are not significantly different. Here, further investigations are necessary on the relationship between immersive characteristics and presence. One neglected aspect might be the grouping of all projective systems into one category.

In addition, the range of scores for each sub-scale is slightly different to some degree. When the score from only one of the sub-scales is presented, its class in the sub-scale ranking scale should be considered as a measure of the capability of generating that aspect of presence only. Among the sub-scales, REAL received the lowest scores, which are in line with the original findings by Schubert et al. [56]. A generally sufficient fidelity was provided by the studies we analysed; increased realism (e.g., higher resolution displays) does not necessarily lead to proportionally higher gains in presence. On the other hand, the scores on the other sub-scales lean toward the highest point. In particular,

we found that the relationship between GP and SP is strong, indicating that the spatial aspect of the virtual environment makes the strongest contribution towards feeling present within it.

In general, we would like to establish a classification scheme for IPQ scores for all kinds of studies. There are different factors which contribute to the sense of presence. This leads to differences in the rating scores between different VR application domains. We combine scores from all domains and with this, the classification derived from the scores can be used to classify the scores from these domains. Our classification provides users of the IPQ with an overview of their collected scores in comparison with data from previous studies. We argue that each examined technique should be compared with its similar domain technique.

With the currently available data and classifications, researchers are able to compare the confidence intervals of their sample to the overall distribution of the IPQ. All measurements suffer from error variance. Every presence measurement has a confidence interval. If an observed sample is small, its confidence interval may, in fact, span two or three categories of our classification. In that case, assigning one category would be meaningless or assigning a very precise absolute number of presence would not be possible.

There are three main purposes for the classifications and data distributions. First of all, they provide a means to compare scores with the pool of available data in the community. From those comparisons, the users of the IPQ can figure out where their scores are located in relation to others. They, then, can correlate the impact of their conditions on the scores and enhance their system in order to induce higher levels of presence, if desired. Secondly, the classifications and data distributions of IPQ scores can lead to better, simplified, and comparable reporting on the presence scores. Finally, the classifications and data distributions present an alternative method to interpret the meaning of the rating scores. The users of the questionnaire will no longer have to presume that their scores are high enough, low, or equal to the middle point of the Likert-type scale (or equal to 0). Instead, they have the option to report their scores more objectively.

Based on our classifications and data distribution, researchers can already draw useful conclusions without creating separate distributions. They can link the properties of our data set with the properties of their own system. Imagine, for instance, a system using a "hypnotic" soundtrack that increased presence. The researcher tests this using a desktop setup. A comparison with our data set shows that the control condition is squat in the middle of the distribution, while the soundtrack version is in the 95th percentile, both with small confidence intervals. The researcher can then conclude that their manipulation lifts their simple desktop setup to the level of more sophisticated systems that use much more advanced technology. This is exactly the kind of inference our classification and data set would contribute.

Overall, if possible, we recommend using the general ranking scale to compare and judge the level of presence of implemented systems in comparison with previous studies regardless of the difference between technologies and sensory feedback. The sense of presence is a complex, multi-dimensional construct and deserves to be measured in that way. While sub-scales might inform specific design and development aspects they are not necessarily indicative of the achieved overall presence.

6.2 Guidelines for Future Administration and Reporting of the IPQ

From our analysis, it is clear that IPQ usage has been inconsistent since the questionnaire was first introduced. Many studies omit certain sub-scales, use different scales for Likert responses, or report their results in a non-standard way, all of which can make comparisons between studies difficult. To encourage more standardised reporting and make future meta-analysis more viable,

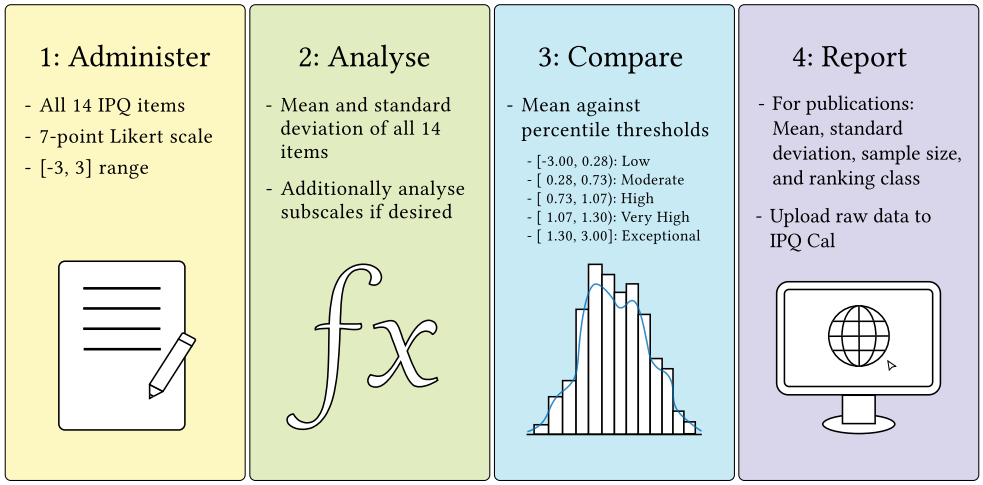

Fig. 8. Our recommended procedure for administering and reporting the results of the IPQ.

we have devised the following guidelines for administering and reporting the results of the IPQ. This procedure is summarised in Figure 8.

Administer:

(1) Present all 14 items of the IPQ using a 7-point Likert-type scale. Each should have the range $[-3, 3]$ and use the original scale anchors.

Analyse:

(1) Multiply the scores from items SP2, SP3, and INV3 by -1 to invert them.
(2) Calculate the aggregated mean and standard deviation for all 14 items for each study condition.
(3) If desired, also calculate the aggregated mean and standard deviation for each sub-scale.

Compare:

(1) Refer to Figure 4, Table 4, or IPQ Cal for the percentile thresholds for each ranking class.
(2) Optionally refer to Table 5 to find thresholds for each visual display modality or Table 8 for each sub-scale's thresholds.
(3) Identify where your score lies on this distribution to easily compare your study results to existing work.

Report:

(1) When disseminating your work, report the mean, standard deviation, sample size, and ranking class of each study condition's aggregated presence score.
(2) Also report the mean and standard deviation for each sub-scale, as well as the visual modality used.
(3) Upload raw study data to IPQ Cal so that the percentile thresholds can be adjusted based on the new data.

To encourage the adoption of these guidelines, we have developed an online tool, *IPQ Cal*, for analysing and reporting the results of the questionnaire. Researchers can upload their raw IPQ study data and the website will produce a statistical analysis as well as a graphical and textual summary of the results for use in publications. The website also provides an overview of previous IPQ data, not only as aggregated in this article but also accumulated from previous submissions, so that the results of new studies can be easily compared and classified in relation to existing research. IPQ Cal is available from the following link: https://hci.otago.ac.nz/ipqcal/.

6.3 Limitations

While we believe that the presented classification scheme is of very practical and academic value for current work investigating presence in virtual environments, there are some important limitations in our work. First, our approach is different from traditional meta-analyses. While mean values for the score are commonly reported in studies, not all papers report more detailed statistics, such as standard deviations. As such, although the number of participants taking part in each aggregated study is reported in the selected studies, we cannot infer general distributions from the aggregated scores without knowing more statistical details. As a result, our method treats all scores from aggregated studies at the same weight even though there might be a significant difference in the number of participants between the studies.

Second, the ranking class thresholds are assigned based on percentile values. These are computed with a method that is free from the distribution of the aggregated scores. Nevertheless, the threshold values are subject to change if new studies are added to the collected data. For this reason, the ranking scale is suitable to be used as a contemporary classification. In order to increase the capability of the scale, we urge future users of the IPQ to contribute their scores through our online tool to enable supplementing and revising the scale to match with studies conducted in the years to come.

As previously discussed, a meta-analysis of IPQ scores was not possible as usually only aggregated scores are presented, not scores for individual questionnaire items or raw data. Instead, we base our method on the limited study data that is available and, with this, get indicative presence score distributions and baselines even in the absence of raw data. As addressed in our outlook, those baseline data form a potential and promising starting point for future raw data collection. If more raw data becomes available in the future, we could compare our current method based on parametric data with some alternative non-parametric meta-analytic techniques. A more comprehensive modelling using random effects-based meta-analytic techniques would be possible with further scores from future studies.

At the moment, we only use means for developing the classification. Other inferential statistical tests and ordinal data, e.g., medians, p-values, and F-values, are not involved. Because of the lack of reporting standard deviations from previous studies, we do not use these values here. Additionally, presenting or discussing different statistics and the issues of statistical evaluations of the IPQ are beyond the scope of this article.

There is another constraint of the classification development, namely the overlap between estimations of the confidence interval range for each percentile. The lower 95% CI of P_{95th} is located in the confidence interval range for P_{90th} (95% CI [0.93, 1.20]). Table 10 shows the estimates for each percentile with its confidence intervals. Although there are overlaps in the percentiles' confidence intervals, the estimated values of our selected percentiles are not in the confidence interval range of the other percentiles. This shows that the current classification using the estimated values can be applied to classify the IPQ scores. However, for better and more reliable classification, we would urge users of the IPQ to report, distribute, and contribute their scores using the guidelines in Section 6.2.

Table 10. Percentiles and Their Estimated Values and Confidence Intervals from 243 User Studies Reporting on Mean Presence Scores Measured Using the IPQ

Percentile	Value	95% CI	
		Lower	Upper
P_{50th}	0.28	0.22	0.40
P_{75th}	0.73	0.66	0.84
P_{90th}	1.07	0.93	1.20
P_{95th}	1.30	1.17	1.61

6.4 Future Work

As indicated throughout the article, besides inconsistencies in applying the IPQ, the current quality and detail of the reported data do not allow for a traditional meta-analysis of IPQ presence measurements. However, we hope that this work sheds light on this problem and helps with reporting IPQ presence measurement in future work. Together with the general trend towards open data, we argue that the quality of the reported data will increase in the future, consequently allowing for future work to run a meta-analysis on presence scores. Furthermore, the data from the current studies using the IPQ does not allow us to infer any adjective rating (e.g., what IPQ scores actually represent an awful feeling of presence). Similar to the SUS, this would require the addition of a separate adjective rating scale and the collection of enough data to allow for a robust analysis (SUS used data from 964 participants) [4]. Thus, future work could either adapt the IPQ for all future studies by adding an adjective rating scale or individual entities run enough studies using IPQ to create sufficient data from their studies and decide to add an adjective rating scale to all of their studies. A candidate could be the recently established "Virtual Experience Research Accelerator."[4]

Finally, future studies [13] can also look at linking other types of measurements towards reported presence measurements. This includes possible triangulation of different measurements, including questionnaires, observations, and physiological measurements [13]. The idea would be to correlate presence scores with other measurements to further support baselines or an adjective rating scale (e.g., with a score of X we increasingly see elevated physiological measurements or behavioural changes).

7 Conclusion

In this work, we proposed a first rating scale for interpreting presence scores based on analysing 243 published studies using the IPQ. Although there are more than 15 different presence questionnaires developed over more than two decades, there is no available scale to rank or compare presence absolutely or across studies. The IPQ as a subjective presence measurement has shown effectiveness and sensitivity to evaluating the sense of presence in virtual environments and is considered one of the most reliable and effective questionnaires to evaluate presence in virtual environments [58]. It is based on sound theory, it is validated, and it is convenient to use due to the relatively small number of questions.

Unfortunately, most studies using IPQ or similar questionnaires for measuring presence lack a discussion about their scores. There is no discussion of the absolute scores and how presence was induced in their studies [54]. Sometimes, in the absence of any baseline measures to date, presence

[4] https://sreal.ucf.edu/vera/

scores were also compared with the mid-point of the Likert-type scale in order to determine high/low presence measures. For example, Devigne et al. reported that "Scores were on average above 3 for INV and above 4 for General Impression (G) and SP [15]. For the first time, our work allows the interpretation of IPQ scores relative to earlier studies reported in the literature. This includes scores for the entire IPQ as well as for individual display modalities and sub-scales. We furthermore made our data available through a web service that also can assist in reporting future IPQ studies, specifically targeting the inconsistencies in current reporting. As such, we see this as a first step towards more standardised reporting. It will also allow other researchers to improve the quality of presence measurements and to more easily relate their own study to others.

Author Contributions

Tanh Tran performed writing—original draft, conceptualisation, investigation and data curation, Tobias Langlotz performed writing—review & editing, conceptualisation, funding acquisition, and supervision, Jacob Young performed writing—review & editing and visualization, Thomas Schubert performed writing—review & editing, methodology, Holger Regenbrecht performed writing—review & editing, conceptualisation, funding acquisition, and supervision.

References

[1] Dmitry Alexandrovsky, Susanne Putze, Michael Bonfert, Sebastian Höffner, Pitt Michelmann, Dirk Wenig, Rainer Malaka, and Jan David Smeddinck. 2020. Examining design choices of questionnaires in VR user studies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, New York, NY, 1–21. DOI: https://doi.org/10.1145/3313831.3376260

[2] Ivan Alsina-Jurnet and José Gutiérrez-Maldonado. 2010. Influence of personality and individual abilities on the sense of presence experienced in anxiety triggering virtual environments. *International Journal of Human-Computer Studies* 68, 10 (2010), 788–801. DOI: https://doi.org/10.1016/j.ijhcs.2010.07.001

[3] R.M. Baños, C. Botella, A. Garcia-Palacios, H. Villa, C. Perpiña, and M. Alcañiz. 2000. Presence and reality judgment in virtual environments: A unitary construct? *CyberPsychology & Behavior* 3, 3 (2000), 327–335. DOI: https://doi.org/10.1089/10949310050078760

[4] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of Usability Studies* 4, 3 (May 2009), 114–123.

[5] Woodrow Barfield and Suzanne Weghorst. 1993. The sense of presence within virtual environments: A conceptual framework. *Advances in Human Factors Ergonomics* 19 (1993), 699–699.

[6] Lutz Bornmann, Loet Leydesdorff, and Rüdiger Mutz. 2013. The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. *Journal of Informetrics* 7, 1 (2013), 158–165. DOI: https://doi.org/10.1016/j.joi.2012.10.001

[7] Lutz Bornmann and Rüdiger Mutz. 2011. Further steps towards an ideal method of measuring citation performance: The avoidance of citation (ratio) averages in field-normalization. *Journal of informetrics* 1, 5 (2011), 228–230.

[8] Karl-Erik Bystrom, Woodrow Barfield, and Claudia Hendrix. 1999. A conceptual model of the sense of presence in virtual environments. *Presence: Teleoperators and Virtual Environments* 8, 2 (1999), 241–244. DOI: https://doi.org/10.1162/105474699566107

[9] E. Chapoulie, R. Guerchouche, P. Petit, G. Chaurasia, P. Robert, and G. Drettakis. 2014. Reminiscence therapy using image-based rendering in VR. In *Proceedings of the IEEE Virtual Reality (VR)*. IEEE, 45–50.

[10] Vuthea Chheang, Patrick Saalfeld, Fabian Joeres, Christian Boedecker, Tobias Huber, Florentine Huettl, Hauke Lang, Bernhard Preim, and Christian Hansen. 2021. A collaborative virtual reality environment for liver surgery planning. *Computers & Graphics* 99 (2021), 234–246. DOI: https://doi.org/10.1016/j.cag.2021.07.009

[11] Jacob Clarkson. 2018. *The Effects of Augmented Virtuality on Presence, Workload, and Input Performance*. Ph.D. Dissertation. University of Cape Town.

[12] R. M. S. Clifford, T. McKenzie, S. Lukosch, R. W. Lindeman, and S. Hoermann. 2020. The effects of multi-sensory aerial firefighting training in virtual reality on situational awareness, workload, and presence. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 93–100.

[13] J. Collins, H. Regenbrecht, T. Langlotz, Y. Said Can, C. Ersoy, and R. Butson. 2019. Measuring cognitive load and insight: A methodology exemplified in a virtual reality learning context. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 351–362. DOI: https://doi.org/10.1109/ISMAR.2019.00033

[14] Angela De La O, Kristine C. Jordan, Karen Ortiz, Laurie J. Moyer-Mileur, Greg Stoddard, Mike Friedrichs, Rachel Cox, Emily C. Carlson, Elizabeth Heap, and Nicole L. Mihalopoulos. 2009. Do parents accurately perceive their child's weight status? *Journal of Pediatric Health Care* 23, 4 (2009), 216–221. DOI: https://doi.org/10.1016/j.pedhc.2007.12.014

[15] L. Devigne, M. Babel, F. Nouviale, V. K. Narayanan, F. Pasteau, and P. Gallien. 2017. Design of an immersive simulator for assisted power wheelchair driving. In *Proceedings of the International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 995–1000.

[16] H. Q. Dinh, N. Walker, L. F. Hodges, Chang Song, and A. Kobayashi. 1999. Evaluating the importance of multi-sensory input on memory and the sense of presence in virtual environments. In *Proceedings of the IEEE Virtual Reality (Cat. No. 99CB36316)*. IEEE, 222–228. DOI: https://doi.org/10.1109/VR.1999.756955

[17] Dongsik Cho, Jihye Park, G. J. Kim, Sangwoo Hong, Sungho Han, and Seungyong Lee. 2003. The dichotomy of presence elements: the where and what. In *Proceedings of the IEEE Virtual Reality*. IEEE, 273–274. DOI: https://doi.org/10.1109/VR.2003.1191155

[18] Ioannis Doumanis, Daphne Economou, and Lemonia Argyriou. 2021. Measuring and comparing QoE of hybrid VR applications under increased network load. In *Proceedings of the 7th International Conference of the Immersive Learning Research Network (iLRN)*. IEEE, 1–7. https://doi.org/10.23919/iLRN52045.2021.9459316

[19] Bonita Falkner, Samuel S. Gidding, Ronald Portman, and Bernard Rosner. 2008. Blood pressure variability and classification of prehypertension and hypertension in adolescence. *Pediatrics* 122, 2 (2008), 238–242. DOI: https://doi.org/10.1542/peds.2007-2776

[20] Anna Felnhofer, Oswald D. Kothgassner, Mareike Schmidt, Anna-Katharina Heinzle, Leon Beutl, Helmut Hlavacs, and Ilse Kryspin-Exner. 2015. Is virtual reality emotionally arousing? Investigating five emotion inducing virtual park scenarios. *International Journal of Human-Computer Studies* 82 (2015), 48–56. DOI: https://doi.org/10.1016/j.ijhcs.2015.05.004

[21] M. Gerhard, David J. Moore, and Dave J. Hobbs. 2001. Continuous presence in collaborative virtual environments: Towards a hybrid avatar-agent model for user representation. In *Proceedings of the 3rd International Workshop on Intelligent Virtual Agents (IVA '01)*. Springer-Verlag, 137–155. Retrieved from http://dl.acm.org/citation.cfm?id=648034.744242

[22] Arthur M. Glenberg. 1997. What memory is for. *Behavioral and Brain Sciences* 20, 1 (1997), 1–55. DOI: https://doi.org/10.1017/S0140525X97000010

[23] Dwi Hartanto, Isabel L. Kampmann, Nexhmedin Morina, Paul G. M. Emmelkamp, Mark A. Neerincx, and Willem-Paul Brinkman. 2014. Controlling social stress in virtual reality environments. *PLoS ONE* 9, 3 (2014), 1–17. DOI: https://doi.org/10.1371/journal.pone.0092804

[24] Tilo Hartmann, Werner Wirth, Holger Schramm, Christoph Klimmt, Peter Vorderer, André Gysbers, Saskia Böcking, Niklas Ravaja, Jari Laarni, Timo Saari, Feliz Ribeiro Gouveia and Ana Sacau. 2015. The spatial presence experience scale (SPES). *Journal of Media Psychology* 28, 1 (2015), 1–15.

[25] K. A. Hernandez-Ossa, B. Longo, E. Montenegro-Couto, M. A. Romero-Laiseca, A. Frizera-Neto, and T. Bastos-Filho. 2017. Development and pilot test of a virtual reality system for electric powered wheelchair simulation. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2355–2360.

[26] Philipp Hock, Johannes Kraus, Marcel Walch, Nina Lang, and Martin Baumann. 2016. Elaborating feedback strategies for maintaining automation in highly automated driving. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (Automotive'UI 16)*. ACM, New York, NY, 105–112. DOI: https://doi.org/10.1145/3003715.3005414

[27] M. F. Huque. 1988. Experiences with meta-analysis in NDA submissions. In *Proceedings of the Biopharmaceutical Section of the American Statistical Association*, Vol. 2. American Statistical Organization, 28–33.

[28] Rob J. Hyndman and Yanan Fan. 1996. Sample quantiles in statistical packages. *The American Statistician* 50, 4 (1996), 361–365. Retrieved from http://www.jstor.org/stable/2684934

[29] H. Y. Kang, G. Lee, D. S. Kang, O. Kwon, J. Y. Cho, H. Choi, and J. H. Han. 2019. Jumping further: Forward jumps in a gravity-reduced immersive virtual environment. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 699–707.

[30] Taeyong Kim and Frank Biocca. 1997. Telepresence via television: Two dimensions of telepresence may have different connections to memory and persuasion.[1]. *Journal of Computer-Mediated Communication* 3, 2 (1997), Article JCMC325. https://doi.org/10.1111/j.1083-6101.1997.tb00073.x

[31] Michael Krauss, Rainer Scheuchenpflug, Walter Piechulla, and Alf Zimmer. 2001. Measurement of presence in virtual environments. *Experimentelle Psychologie im Spannungsfeld von Grundlagenforschung und Anwendung Proceedings* 43 (2001), 358–362.

[32] Sonja Th Kwee-Meier, Alexander Mertens, and Sabina Jeschke. 2019. Recommendations for the design of digital escape route signage from an age-differentiated experimental study. *Fire Safety Journal* 110 (2019), 102888. DOI: https://doi.org/10.1016/j.firesaf.2019.102888

[33] Jane Lessiter, Jonathan Freeman, Edmund Keogh, and Jules Davidoff. 2001. A cross-media presence questionnaire: The ITC-sense of presence inventory. *Presence: Teleoperators and Virtual Environments* 10, 3 (2001), 282–297. DOI: https://doi.org/10.1162/105474601300343612

[34] Stefan Liszio, Katharina Emmerich, and Maic Masuch. 2017. The influence of social entities in virtual reality games on player experience and immersion. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*. ACM, 1–10.

[35] Matthew Lombard and Theresa Ditton. 1997. At the heart of it all: The concept of presence. *Journal of Computer-Mediated Communication* 3, 2 (1997), Article JCMC321.

[36] Matthew Lombard, Theresa B. Ditton, Daliza Crane, Bill Davis, Gisela Gil-Egui, Karl Horvath, Jessica Rossman, and S. Park. 2000. Measuring presence: A literature-based approach to the development of a standardized paper-and-pencil instrument. In *Proceedings of the 3rd International Workshop on Presence, Delft, the Netherlands*, Vol. 240. MIT Press, 2–4.

[37] Matthew Lombard, Theresa B. Ditton, and Lisa Weinstein. 2009. Measuring presence: the temple presence inventory. In *Proceedings of the 12th Annual International Workshop on Presence*. MIT Press, 1–15.

[38] Mark McGill, Daniel Boland, Roderick Murray-Smith, and Stephen Brewster. 2015. A dose of reality: Overcoming usability challenges in VR head-mounted displays. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, 2143–2152. DOI: https://doi.org/10.1145/2702123.2702382

[39] Joanne E. McKenzie, Elaine M. Beller, and Andrew B. Forbes. 2016. Introduction to systematic reviews and meta-analysis. *Respirology* 21, 4 (2016), 626–637. DOI: https://doi.org/10.1111/resp.12783

[40] Jantsje M. Mol, Eline C. M. van der Heijden, and Jan J. M. Potters. 2020. (Not) alone in the world: Cheating in the presence of a virtual observer. *Experimental Economics* 23 (2020), 961–978. DOI: https://doi.org/10.1007/s10683-020-09644-0

[41] Nexhmedin Morina, Willem-Paul Brinkman, Dwi Hartanto, and Paul MG Emmelkamp. 2014. Sense of presence and anxiety during virtual social interactions between a human and virtual humans. *PeerJ* 2 (2014), e337.

[42] Jörg Müller, Tobias Langlotz, and Holger Regenbrecht. 2016. PanoVC: Pervasive telepresence using mobile phones. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 1–10.

[43] Sarah Nichols, Clovissa Haldane, and John R Wilson. 2000. Measurement of presence and its consequences in virtual environments. *International Journal of Human-Computer Studies* 52, 3 (2000), 471–491. DOI: https://doi.org/10.1006/ijhc.1999.0343

[44] K. L. Nowak and F. Biocca. 2003. The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence* 12, 5 (2003), 481–494.

[45] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* 372 (2021), 9 pages. DOI: https://doi.org/10.1136/bmj.n71

[46] John Porter III, Matthew Boyer, and Andrew Robb. 2018. Guidelines on successfully porting non-immersive games to virtual reality: A case study in minecraft. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM, New York, NY, 405–415.

[47] Matthew Price, Natasha Mehta, Erin B. Tone, and Page L. Anderson. 2011a. Does engagement with exposure yield better outcomes? Components of presence as a predictor of treatment response for virtual reality exposure therapy for social phobia. *Journal of Anxiety Disorders* 25, 6 (2011), 763–770. DOI: https://doi.org/10.1016/j.janxdis.2011.03.004

[48] Matthew Price, Natasha Mehta, Erin B. Tone, and Page L. Anderson. 2011b. Does engagement with exposure yield better outcomes? Components of presence as a predictor of treatment response for virtual reality exposure therapy for social phobia. *Journal of Anxiety Disorders* 25, 6 (2011), 763–770. DOI: https://doi.org/10.1016/j.janxdis.2011.03.004

[49] Holger Regenbrecht, Simon Hoermann, Graham McGregor, Brian Dixon, Elizabeth Franz, Claudia Ott, Leigh Hale, Thomas Schubert, and Julia Hoermann. 2012. Visual manipulations for motor rehabilitation. *Computers & Graphics* 36, 7 (2012), 819–834. DOI: https://doi.org/10.1016/j.cag.2012.04.012

[50] Holger Regenbrecht and Thomas Schubert. 2002. Real and illusory interactions enhance presence in virtual environments. *Presence: Teleoperators & Virtual Environments* 11, 4 (2002), 425–434. DOI: https://doi.org/10.1162/105474602760204318

[51] Holger Regenbrecht, Thomas Schubert, Cristina Botella, and Rosa Baños. 2017. Mixed reality experience questionnaire (mreq)-reference. *The Information Science Discussion Paper Series* 1, 2017/01 (2017), 2 pages.

[52] R. T. Reinhard, M. Kleer, and K. Dreßler. 2019. The impact of individual simulator experiences on usability and driving behavior in a moving base driving simulator. *Transportation Research Part F: Traffic Psychology and Behaviour* 61 (2019), 131–140. DOI: https://doi.org/10.1016/j.trf.2018.01.004

[53] Bernhard E. Riecke, Aleksander Väljamäe, and Jörg Schulte-Pelkum. 2009. Moving sounds enhance the visually-induced self-motion illusion (circular vection) in virtual reality. *ACM Transactions on Applied Perception* 6, 2 (March 2009), Article 7, 27 pages. DOI: https://doi.org/10.1145/1498700.1498701

[54] Enrico Ronchi, Max Kinateder, Mathias Müller, Michael Jost, Markus Nehfischer, Paul Pauli, and Andreas Mühlberger. 2015. Evacuation travel paths in virtual reality experiments for tunnel safety analysis. *Fire Safety Journal* 71 (2015), 257–267. DOI: https://doi.org/10.1016/j.firesaf.2014.11.005

[55] Rufat Rzayev, Sven Mayer, Christian Krauter, and Niels Henze. 2019. Notification in vr: The effect of notification placement, task and environment. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM, 199–211.

[56] Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. 2001. The experience of presence: Factor analytic insights. *Presence: Teleoperators and Virtual Environments* 10, 3 (2001), 266–281. DOI: https://doi.org/10.1162/105474601300343603

[57] Thomas W. Schubert. 2003. The sense of presence in virtual environments:. *Zeitschrift für Medienpsychologie* 15, 2 (2003), 69–71. DOI: https://doi.org/10.1026//1617-6383.15.2.69

[58] Valentin Schwind, Pascal Knierim, Nico Haas, and Niels Henze. 2019. Using presence questionnaires in virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, Article 360, 12 pages. DOI: https://doi.org/10.1145/3290605.3300590

[59] Thomas B. Sheridan. 1992. Musings on Telepresence and Virtual Presence. *Presence: Teleoperators and Virtual Environments* 1, 1 (1992), 120–126. DOI: https://doi.org/10.1162/pres.1992.1.1.120

[60] Mel Slater, Martin Usoh, and Anthony Steed. 1994. Depth of presence in virtual environments. *Presence: Teleoperators and Virtual Environments* 3, 2 (1994), 130–144. DOI: https://doi.org/10.1162/pres.1994.3.2.130

[61] Mel Slater and Sylvia Wilbur. 1997. A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators and Virtual Environments* 6, 6 (1997), 603–616. DOI: https://doi.org/10.1162/pres.1997.6.6.603

[62] Jonathan Steuer. 1992. Defining virtual reality: Dimensions determining telepresence. *Journal of Communication* 42, 4 (1992), 73–93. DOI: https://doi.org/10.1111/j.1460-2466.1992.tb00812.x

[63] Clara Suied, George Drettakis, Olivier Warusfel, and Isabelle Viaud-Delmon. 2013. Auditory-visual virtual reality as a diagnostic and therapeutic tool for cynophobia. *Cyberpsychology, Behavior, and Social Networking* 16, 2 (2013), 145–152. DOI: https://doi.org/10.1089/cyber.2012.1568

[64] Marcel Takac, James Collett, Kristopher J. Blom, Russell Conduit, Imogen Rehm, and Alexander De Foe. 2019. Public speaking anxiety decreases within repeated virtual reality training sessions. *PLoS ONE* 14, 5 (2019), e0216288. DOI: https://doi.org/10.1371/journal.pone.0216288

[65] Wenjing Tang, Gun A. Lee, Mark Billinghurst, and Robert W. Lindeman. 2018. User virtual costume visualisation in an augmented virtuality immersive cinematic environment. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction (OzCHI '18)*. ACM, New York, NY, 219–223. DOI: https://doi.org/10.1145/3292147.3292188

[66] Karin Tanja-Dijkstra, Sabine Pahl, Mathew P. White, Jackie Andrade, Cheng Qian, Malcolm Bruce, Jon May, and David R. Moles. 2014. Improving dental experiences by using virtual reality distraction: A simulation study. *PLoS ONE* 9, 3 (2014), 1–10. DOI: https://doi.org/10.1371/journal.pone.0091276

[67] Gordon Tao and Philippe S. Archambault. 2016. Powered wheelchair simulator development: Implementing combined navigation-reaching tasks with a 3D hand motion controller. *Journal of NeuroEngineering and Rehabilitation* 13, 1 (2016), 3. DOI: https://doi.org/10.1186/s12984-016-0112-2

[68] Jennifer G. Tichon and Guy M. Wallis. 2010. Stress training and simulator complexity: Why sometimes more is less. *Behaviour & Information Technology* 29, 5 (2010), 459–466. DOI: https://doi.org/10.1080/01449290903420184

[69] Tanh Quang Tran, Tobias Langlotz, and Holger Regenbrecht. 2024. A survey on measuring presence in mixed reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, New York, NY, Article 543, 38 pages. DOI: https://doi.org/10.1145/3613904.3642383

[70] Martin Usoh, Ernest Catena, Sima Arman, and Mel Slater. 2000. Using presence questionnaires in reality. *Presence: Teleoperators and Virtual Environments* 9, 5 (2000), 497–503. DOI: https://doi.org/10.1162/105474600566989

[71] Jacinto Vasconcelos-Raposo, Maximino Bessa, Miguel Melo, Luis Barbosa, Rui Rodrigues, Carla Maria Teixeira, Luciana Cabral, and António Augusto Sousa. 2016. Adaptation and validation of the igroup presence questionnaire (IPQ) in a Portuguese sample. *Presence: Teleoperators and Virtual Environments* 25, 3 (2016), 191–203.

[72] Waloddi Weibull. 1939. *The Phenomenon of Rupture in Solids*. Generalstabens litografiska anstalts förlag, Stockholm.

[73] Connor B. Weir and Arif Jan. 2023. *BMI Classification Percentile and Cut Off Points*. StatPearls Publishing, Treasure Island, FL. Retrieved from http://europepmc.org/books/NBK541070

[74] Rick Wicklin. 2017. Sample Quantiles: A Comparison of 9 Definitions. SAS. Retrieved from https://blogs.sas.com/content/iml/2017/05/24/definitions-sample-quantiles.html

[75] Bob G Witmer and Michael F Singer. 1994. *Measuring Presence in Virtual Environments*. Technical Report. Army Research Inst for the Behavioral and Social Sciences, Alexandria, VA.
[76] Bob G. Witmer and Michael J. Singer. 1998. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments* 7, 3 (1998), 225–240. DOI: https://doi.org/10.1162/105474698565686
[77] Ruud Zaalberg and Cees J. H. Midden. 2013. Living behind dikes: Mimicking flooding experiences. *Risk Analysis* 33, 5 (2013), 866–876. DOI: https://doi.org/10.1111/j.1539-6924.2012.01868.x
[78] İpek Memikoğlu and Halime Demirkan. 2020. Exploring staircases as architectural cues in virtual vertical navigation. *International Journal of Human-Computer Studies* 138 (2020), Article 102397. DOI: https://doi.org/10.1016/j.ijhcs.2020.102397

Received 10 June 2024; revised 10 June 2024; accepted 15 July 2024