The development and validation of the Immersion Experience Questionnaire Short Form (IEQ-SF)

Joe Cutting, Myat Aung, Francis Perret, Paul Cairns

Abstract
Digital games are a dominant form of entertainment and of increasing interest to HCI researchers. Unlike productivity applications, the chief outcome of a game is the experience it provides and one of the key aspects of that experience is the feeling of immersion. Immersion can be measured using the Immersion Experience Questionnaire (IEQ) (Jennett et al., 2008), however this scale was developed over 16 years ago using outdated methods and its length (31 items) makes it unsuited for studies which use multiple scales or recruitment platforms which charge by the minute. To address this, we used item response theory and open science practices to develop a new Immersion Experience Questionnaire Short Form (IEQ-SF). We performed a large-scale factor analysis and developed a candidate short-form with 11-items which represent the subfactors of Involvement, Real-World Dissociation and Challenge. This scale was then validated with four studies using a mix of existing data available through Open Science and a new validation study. The short-form was found to function very similarly to the original IEQ. We also provide best practice guidelines for development of short form scales, and through OSF, all analysis scripts and data used during this scale development.

# Introduction

Digital games have become increasingly dominant both as a form of leisure (Unity Technologies, 2022) and a domain of study by HCI researchers. As a research domain, games differ from early HCI studies for which the key goal was the maximising of productivity. Instead, the key feature of digital games is the experience of playing and arguably the most important aspect of that experience is that users are *engaged* – they play because they want to. Game engagement has accordingly attracted considerable academic interest, both in its own right and in terms of concepts such as flow, enjoyment, presence and immersion. Though these terms differ, there is growing evidence that pragmatically they are simply different facets of engagement (Denisova et al., 2016). As a significant facet of game engagement, immersion has been found to be highly valued by players and developers (Brown & Cairns, 2004) and refers to the experience of finding the game so engaging that players do not notice their surroundings and feel “in the game” (Brown & Cairns, 2004; Jennett et al., 2008). Game immersion has been seen as a product of sensory, gameplay and narrative factors (Abeele et al., 2020; Calleja, 2011; Jennett et al., 2008; Mäyrä & Ermi, 2005) and found to be moderated by specific changes to in game design (Denisova & Cairns, 2015; Iacovides et al., 2015; Sanders & Cairns, 2010). Players enjoy more immersive games (Brown & Cairns, 2004) so increasing immersion is a goal for game designers but immersion has also been linked to both positive game effects such as learning (Cheng et al., 2015; Hamari et al., 2016) and negative effects such as addiction (Seah & Cairns, 2008).

Accordingly, considerable research has been done onto how games create immersion and how this can be increased. A key prerequisite for these studies is a robust measure of immersion. Physiological measures such as EEG (Burns & Fairclough, 2015), ECG (Nacke & Lindley, 2008), Eye tracking (Cox et al., 2006) and attention (Cutting & Cairns, 2022) have all been used, but the majority of studies quantify players’ feelings of immersion with a post-game questionnaire. Of these questionnaires the most widely used is the 31-item *Immersion-Experience Questionnaire* (IEQ) (Jennett et al., 2008) with over 2000 citations, translations into other languages such as Bahasa Malaysia (Nordin et al., 2018), and use in numerous player experience studies.

However, in the 16 years since the IEQ was created, the requirements and context for player experience questionnaires has changed; in particular there is now a need for shorter questionnaires that take less time to fill in. The original IEQ has 31 questions and can take around 3 minutes to complete. Many game experience studies are now run via online recruitment platforms such as *Mechanical Turk* (e.g. Bowey et al., 2015; Johanson et al., 2019) or *Prolific* (e.g. Cutting & Iacovides, 2022) which charge researchers for every minute of time taken by participants. In a short study taking ten minutes, filling in a long questionnaire can easily take 25% of the study time and accordingly 25% of the participant fees. A shorter questionnaire thus has the potential to save 10-15% of the whole study cost. This is more of an issue in studies that use a battery of questionnaires to measure changes in experience over time or several different constructs, so that filling in the questionnaire may form a significant part of the study duration which then both increases cost of research and raises concerns that participants will become fatigued or bored and the quality of their responses will suffer (O’Brien et al., 2018). In response to these concerns, researchers have used individual subscales of longer questionnaires such as the Intrinsic Motivation Inventory (Deci & Ryan, 1985), used short questionnaires such as the 6-item NASA-TLX (Hart & Staveland, 1988) and the 10-item System Usability Scale (Lewis, 2018) or created short-forms of already established questionnaires such as the User Engagement Scale (O’Brien et al., 2018). However, none of these approaches are suitable for measuring immersion because the current IEQ requires all subscales to measure immersion and there is, as yet, no appropriate short form immersion questionnaire.

A further consequence of the time that has passed since the creation of the original IEQ is that there have been two main developments in the field, which were not available at the time, but now present an opportunity to create an improved questionnaire. The first of these advances is the availability of tools for using *Item Response Theory* (IRT) (Embretson & Reise, 2013), which has been made easily available through the *mirt* package (Chalmers, 2012) for the *R* statistics system. IRT improves on the Classical Test Theory (CTT) used to develop the original IEQ in that IRT considers that responses to individual test items may be due to both differences between the items and differences between the respondents. The second of these advances is the *Open Science* movement (Munafò et al., 2017; Nosek et al., 2015; Vornhagen et al., 2020), which in response to the replication crisis in psychology and other disciplines (Cockburn et al., 2020; Shrout & Rodgers, 2018; Tackett et al., 2019) has normalised the sharing of datasets. The original development and validation of the IEQ was limited to whatever the data the original authors could collect which was only 244 participants. Now, in part due to the IEQ’s success, many more immersion data sets have been collected and due to the Open Science movement, they are available for use in creating improved questionnaires. As a further consequence of Open Science developments, new questionnaires can be developed using open principles of pre-registration and data sharing thus creating improved transparency and accountability over previous immersion measures.

Therefore, in response to the requirement from researchers for a shorter version of the IEQ alongside developments in Open Science and item response theory, our goal is to develop a modern measure of immersion designed to meet the pragmatic requirements current player research. In this paper, we describe the development of the Immersion Experience Questionnaire Short Form (IEQ-SF) including the selection of question items and an extensive testing and validation process.

# Background

This section explores the current context of both game immersion measurement and questionnaire development techniques in order to determine a rigorous, pragmatic methodology that will create a short form which is appropriate for current needs. As such, it first discusses research on immersion and approaches to measuring game experience including the existing Immersion Experience Questionnaire (IEQ). We then describe advances in scale development since the original IEQ and the implications of the rise of the *Open Science* movement on the creation of questionnaires. Finally, we consider recommendations for short form scale development and lay out the procedure we followed to develop the short form immersion experience questionnaire.

The experience of immersion in video games and other media has attracted considerable interest from researchers. Brown and Cairns (2004) performed a grounded theory analysis of player interviews to find a common experience described as “When you stop thinking about the fact that you’re playing a computer game and you’re just in a computer”. Around the same time, Mäyrä and Ermi (2005) created a “gameplay experience model” in which they considered three different types of immersion; sensory, challenge-based and imaginative immersion which, roughly speaking, are created by the hardware, gameplay and narrative of a game respectively.

Since then there have been a variety of formulations of immersion and game engagement, for example Curran (2013) focused on the identification of the player with the setting or characters of a game and Calleja (2011) extended immersion to the notion of incorporation as a collection of forms of involvement. These different formulations tend to not only consider the experience of immersion but what might be called the sources of immersion, that is, the elements of a game which engender the sense of immersion. Agrawal et al. (2019) recognised these different sources of immersion represented in the research literature. However, they synthesised a definition of immersion as an experiential outcome, much like Brown and Cairns (2004). That is, they say that immersion is a “deep mental involvement in which [people’s] cognitive processes (with or without sensory stimulation) cause a shift in their attentional state such that one may experience disassociation from the awareness of the physical world.”

It is such an experiential outcome that Jennett et al. (2008) operationalised as immersion across a set of studies. They related immersion to real world dissociation (Jennett et al., 2009) and concluded that immersion was a form of selective attention (Jennett, 2010). As part of this they developed the Immersive Experience Questionnaire (IEQ) which had five constituent factors of Cognitive Involvement, Emotional Involvement, Real World Dissociation, Challenge, and Control, the first three factors being the outcome of immersion and the last two aspects of the experience of games specifically that lead to immersion.

Concurrent with these investigations into the nature of immersion were a number of approaches to creating measurement instruments. Jennett et al. (2008)’s initial experiments related immersion to reduced ability to carry out demanding tasks outside of the game, but since then a wide variety of physiological and other measures have been considered including EEG (Burns & Fairclough, 2015), ECG (Nacke & Lindley, 2008), Eye tracking (Cox et al., 2006) and attention (Cutting & Cairns, 2022). These physiological measures show some potential but the need for specialist equipment and real expertise in the use of that equipment (Aung, 2022) limits their widespread use. As a consequence, due to its ready availability and straightforward deployment, the most widely used measure has been Jennett et al’s IEQ. This was created using accepted standards of the time, based on 244 responses of whom only 5.7% were female and using classical test theory based on principal components analysis. The resulting 31 item questionnaire (see Appendix) has been very influential with around 2500 citations and has been employed to measure the effect on immersion of a wide variety of design elements including heads-up displays (Iacovides et al., 2015), camera perspectives (Denisova & Cairns, 2015), diegetic game elements (Pfister & Ghellal, 2018), augmented reality (Zhang & Robb, 2021) and emotional aspects of game play (Robinson et al., 2018). The IEQ has also served as the basis for investigations of immersion in non-game contexts, such as the 24 item *Film IEQ* scale (Rigby et al., 2019) for measuring immersion in non-interactive video experiences.

Since the original IEQ was created in 2008, the field of player experience has blossomed – a recent Google scholar search of the term “Player experience” gives 19,200 results. Immersion is one aspect of the broader concept of player experience and the IEQ is just one of many game experience questionnaires such as the Game Engagement Questionnaire (GEngQ) (Brockmyer et al., 2009) or PENS (Ryan et al., 2006) looking at engagement through a variety of different lenses. Denisova et al. (2016) compared these two questionnaires with the IEQ and found a high degree of overlap. Despite this overlap, there are several issues which prevent these scales being effective short measures of immersion. The GEngQ was developed for use in horror games and only contains a single item directly related to immersion. PENS has some validity issues (Johnson et al., 2018) but the biggest issue with this scale is that it is proprietary, which is in direct contradiction to Open Science goals of transparency of measurement, so we cannot recommend the PENS for any research which aims to be replicated. The PXI (Abeele et al., 2020) is an open validated scale which contains a 3-item immersion construct. However, Perrig et al. (2024) re-validated the scale on 1518 players and found low reliability and validity for the immersion construct. These issues suggest that existing scales, including the IEQ, which were developed using older analysis and data gathering techniques may not live up to modern standards of transparency and validity.

The context of studies has also changed since early work on player experience, rather than lab-based studies which use a single questionnaire, modern studies are increasingly likely to be online, subject participants to a battery of different questionnaires and recruit them via platforms such as *Mechanical Turk* or *Prolific* which charge for every minute of study time. All of these factors mean that researchers are finding that older scales with 30-40 items are too long and so turn to subscales of long questionnaires (O’Brien, 2016) or short-form scales such as the User Experience Scale Short Form (O’Brien et al., 2018) and the miniPXI (Haider et al., 2022). These developments suggest that modern statistical techniques and data availability could be used to create a new short immersion scale which will be both quicker to complete and have higher validity than the existing IEQ.

Another development since the development of the original IEQ is the rise of the Open Science movement. This movement was prompted by the “Reproducibility crisis” finding that a significant proportion of results in leading psychology journals could not be replicated (Open Science Collaboration, 2015). Similar findings and concerns have been raised for other fields such as cancer research (Begley & Ellis, 2012) and artificial intelligence (Hutson, 2018). In response, recommendations have been made for increasing transparency and openness (Nosek et al., 2015) with a view to increasing reproducibility (Munafò et al., 2017). These proposals are known as “Open Science” and have two main focuses; transparency of all aspects of study and pre-registration of hypotheses before they are tested. Transparency consists of reporting all aspects of a study’s design; including all experimental materials, all analytical code, all results and all experimental data. Pre-registration of hypotheses entails publishing a detailed plan of the analysis and hypotheses to be tested before they are carried out. The motivation for this is to ensure against “p-hacking” in which multiple hypotheses are tested to find one which gives a significant outcome which is then published as though it were the only test. The Open Science movement was motivated by a desire to ensure the reproducibility of experimental results and a key criterion for a successful scale is that results should be *reliable,* that is they should be highly likely to be reproduced. As such, it follows that the principles of transparency and pre-registration can also improve questionnaire development. Questionnaire development requires large data sets which previously needed to be collected specifically for each questionnaire, but now Open Science transparency requirements mean that existing studies which use questionnaires are more likely to publish the full data set. Accordingly, for the development of our new scale we have made use of existing published data sets for both exploratory scale analysis and for subsequent validation of the resulting questionnaire. As well as providing additional data for analysis, Open Science practices of transparency and pre-registration also allow us to improve the reproducibility and confidence of our scale. By pre-registering our validation studies, we ensure that neither scale nor analysis were altered so that the scale validated successfully. By publishing our analysis code and data sets we make the scale development more open and provide a resource for future researchers. It is important to note that Open Science critics have raised concerns about privacy, misuse of personal data (Dennis et al., 2019; Ruben & Stosic, 2024) and implications for early career researchers (Allen & Mehler, 2019). Thus, Open Science should not be seen as a universal requirement for all research. However, all data in this study is completely anonymised which we believe obviates any privacy concerns and the additional benefits of transparency and reproducibility create a stronger basis for future work by other researchers.

## Best practice for developing short forms

It is well recognised that a short form of questionnaires can be useful and that many established questionnaires therefore have a short form, e.g. the well-known Eysenck Personality Questionnaire (EPQ) has at least two short forms (Colledani et al., 2019). However, there are not widely recognised procedures for producing a short-form that have kept pace with more recent developments in the creation of questionnaires.

A relatively old but still relevant resource is Smith et al. (2000). This offers systematic guidance on producing short forms as a set of recommendations. The first and probably most important recommendation is to make sure that the starting point is a well validated “long” form. Another recommendation is to ensure that where the long form has multiple factors, those multiple factors are captured in any short form and remain valid and meaningful in relation to the original factors. The final recommendation is to evaluate whether the short form actually delivers the benefits that were intended.

Smith et al mainly consider short-form development in the context of Classical Test Theory (CTT) (Crocker & Algina, 1986) which is based on the assumption that each participant has a “true score” for the underlying concept to be measured. Accordingly, the error rate between the true and recorded score can be minimised by using factor analysis to create an optimum scale. Up until recently, CTT was the most commonly used method for creating both long and short form scales (Prieto et al., 2003) due to the easy availability of factor analysis tools and the comparatively low computing power requirements to use them. However, Smith et al. did also consider *Item Response Theory* (IRT) which considers variability within both items and participants and note that it is superior to CTT in creating smaller more reliable scales. A noteworthy aspect of this is that IRT acknowledges that some items may be more likely to have random answers, some may be more likely to have positive responses and others are better at discriminating differences in the underlying concepts (O’Brien et al., 2018). At the time Smith et al (2000) were writing, IRT required prohibitively high computing power to run, however now both the computing power and tools (Chalmers, 2012) are available so IRT can be considered the recommended technique for short form development.

Another approach to factor analysis that Smith et al. were not in a position to consider is the use of bifactor analysis (Reise, 2012). In this approach, the underlying model is that there is a general factor, ***g***, that represents and captures the core concept represented by the questionnaire, for example IQ or Immersion. There are then distinct subfactors that represent related but distinct variations on that core concept, for example for IQ, there are verbal and mathematical reasoning factors that are correlated with ***g*** but that are also conceptually distinct. A bifactor model assumes that each item loads on ***g*** and one other subfactor. The value of this approach is in recognising whether the concept being considered is meaningfully decomposed into distinct subfactors, like say personality questionnaires, or, like IQ, better interpreted as a single factor with some distinct sources of variation.

Smith et al. therefore remains an important guide for the development of short forms but there is growing recognition that these new methods of factor analysis should be used (Greer & Liu, 2016). More significantly, Greer and Liu also emphasise the importance of both theoretical and empirical bases for the selection of items from the long form. Items can be selected (or not) based on their similarity of wording to other items, possible wording ambiguities, possible unreliable interpretations of wording depending on the context as well as closeness to the concept being measured. This selection should further consider statistical measures, specifically item-total correlation and whether dropping an item brings about any change in reliability as measured by the Cronbach alpha. This latter measure is referred to as the “alpha if dropped” for the item. They suggest a minimum threshold of 0.2 for item-total correlations to indicate that an item is acceptable and that if the alpha-if-dropped is appreciably less than the factor’s overall alpha, then the item is important to the factor.

A more extreme approach to producing a short form is through exploring the space of all possible short forms whether exhaustively or through some optimisation algorithm such as genetic algorithms (Franzen et al., 2022). However, this can only be done in relation to statistical measures, such as reliability and neglects considerations of the conceptual underpinning of scale being produced contrary to the recommendations of Greer and Liu and others such as Cairns (2019 Chapter 17).

As well as being a more robust method, IRT can provide a measure of the information captured by each item (Koğar, 2020). Items that are more informative about the latent traits represented by the model are clearly better than less informative items. However, there are no guidelines as to what constitutes a more informative item in the context of items with multiple responses such a 5-response Likert scale. Practically speaking though, the factor loading of an item on the latent concept does broadly correspond to the informativeness of items and therefore we can use more traditional factor loading thresholds for considering this.

In addition, IRT can also produce individual item fit statistics, specifically Infit and Outfit, that Greer and Liu recommend being considered with good items having acceptable standardised fit measures between 0.5 and 1.5 (1.0 is the ideal).

There are also other considerations in the various recommendations such as considering item difficulty (e.g. Calamia et al., 2011) and whether items function differently between different subpopulations (e.g. Smith et al., 2000). However, these typically require maturity in the existing measures so that, in effect, the proposed items of the short form can be evaluated against an established ground truth. In the field of player experience questionnaires, we are simply not able to produce the required benchmarking because of the subjective nature of any instance of an individual’s particular experience.

Overall, we propose the following steps for the development of a short form questionnaire in the context of existing player experience instruments where it is (currently) impossible to identify a ground truth of the actual experience. These steps build primarily on Smith et al but incorporate the use of IRT and bifactor analysis into the approach:

1. Select a well validated questionnaire of the player experience of interest as the long form (Smith et al., 2000).
2. Confirm the factor structure of the long form using exploratory factor analysis and confirmatory factor analysis including bifactor analysis (Reise, 2012; Smith et al., 2000).
3. Develop a candidate short form based on the wording of items in relation to the concepts represented and also consideration of the statistical properties of the shortened scales, specifically item-total correlations, Cronbach alpha and alpha-if-dropped (Greer & Liu, 2016).
4. Confirm the factor structure of the short form on new data (Smith et al, 2000).
5. Validate the short form by evaluating its performance on new data (Smith et al, 2000).

The remainder of the paper will proceed through these steps for extracting a short-form from the IEQ.

# Short form development process

To develop our short form of the IEQ we followed the 5 steps outlined in the previous section on short form development best practice.



Figure Diagram summarising the derivation of the 11 item IEQ Short Form (Steps 2,3 and 4 of our analysis). Processes are shown in blue rounded rectangles, inputs and outputs are in green. RWD is an abbreviation of the Real-World Dissociation factor.

## **Step 1**: Select a well validated questionnaire of the player experience of interest as the long form

In our case this was the existing 31 item IEQ (see Appendix) which is well validated and widely-used questionnaire.

## **Step 2**: Confirm the factor structure of the long form using exploratory factor analysis and confirmatory factor analysis including bifactor analysis

A key criterion for determining the best subset of questions is that the new shorter scale should include all of the factors which were included in the original scale (Smith et al., 2000). The original factor analysis of the IEQ (Jennett et al., 2008) found one omnibus factor and five subfactors; Cognitive Involvement, Real World Dissociation, Challenge, Emotional Involvement and Control. However, this was on a smaller data set using less sophisticated PCA and scree plot analysis. Also, the relationship between the overall IEQ and the constituent subscales was not clarified as bifactor analysis was not available at that time.

To ensure our short form meets the requirements of this step of the process, we performed a new rigorous investigation of the IEQ factor structure using a large-scale data set (n=3539) of responses to the IEQ questionnaire (Perrett, 2018). The respondents completed the survey in response to playing 669 different games that included *Overwatch, Pokemon Go, The Sims 4, Rocket League and Monster Hunter: World.* This dataset also contained another questionnaire and some demographic items to make a total of 67 items. Based on a ten percent trimmed mean, the mean time to respond to an item was 6.8s.

This dataset was split randomly into an exploratory half and a validation half. The exploratory half was used for the exploratory factor analyses carried out for the initial exploration of the factor structure, and the selection of IEQ-SF items. The validation half was used later for the sole purpose of conducting the confirmatory item response theory (IRT) factor analysis.

### Exploratory analysis

The initial IEQ exploratory factor analysis only found one omnibus factor; however, this did not account for 6 of the 31 questions (Jennett et al., 2008). These omitted questions were from four of the five original questionnaire factors; 2 from Challenge, 1 from Cognitive Involvement, 2 from RWD and 1 from Control. Furthermore, the reliability statistics of the Challenge and Control factors were well below typical thresholds (see ).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Factor** | **Low loading items** | **α (95% CI)** | **Item-total r range** | **α if dropped range** |
| Cognitive Involvement  | 22 | 0.77 (0.73-0.77) | 0.31-0.62 | 0.71-0.76 |
| Emotional Involvement | n/a | 0.66 (0.63-0.68) | 0.35-0.62 | 0.42-0.72 |
| Real World Dissociation | 8, 9 | 0.75 (0.73-0.77) | 0.44-0.68 | 0.70-0.74 |
| Challenge | 18, 20 | 0.23 (0.17-0.29) | 0.37-0.74 | 0.06-0.67 |
| Control | 10 | 0.37 (0.32-0.42) | 0.13-0.51 | 0.22-0.40 |

Table Reliability analysis of the original formulation of the IEQ performed on the exploratory half of our large dataset (n = 1770).

To ensure that our new analysis considered all of the most likely factor structures we carried out five exploratory factor analyses corresponding to between 1 and 5 factors. Structure exploration (Chalmers, 2012) was done using the *mirt* R package. Evaluations of scale consistency and validation analysis used the *agricolae* (De Mendiburu & Simon, 2015) and *psych* (Revelle, 2020) packages. For each of the 5 factor options the items with the greatest factor loadings were chosen in a shortlist, with a nominal exclusion threshold of a loading above 0.3 or below -0.3 for inclusion in the factor, as recommended by (Kline, 2000, 2014).

 Our main criteria for evaluating the factor structures were:

* Completeness: Does the factor structure account for as many of the original set of 31 questions as possible?
* Minimum cross loadings: Cross loading is when one question loads onto more than one factor. It should be minimised to ensure that each factor is separate from the others (Kline, 1994)
* Parsimony: factors offer a useful description of the concept of interest, in this case immersion, but to be effective it is generally better to have fewer factors provided they meaningfully represent the overarching concept.
* Conceptual clarity: Each factor should map onto a single concept which is easy to relate to the experience being measured

### Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 factor | 2 factor | 3 factor | 4 factor | 5 factor |
|  | 1F-F1 | 2F-F1 | 2F-F2 | 3F-F1 | 3F-F2 | 3F-F3 | 4F-F1 | 4F-F2 | 4F-F3 | 4F-F4 | 5F-F1 | 5F-F2 | 5F-F3 | 5F-F4 | 5F-F5 |
| IEQ1 | 0.72 | -0.58 | -0.30 | **-0.50** | **-0.32** | 0.17 | **-0.56** | **-0.35** | **0.11** | **-0.02** | **-0.23** | **-0.35** | **0.09** | **-0.01** | **0.41** |
| IEQ2 | 0.76 | -0.73 | -0.13 | -0.65 | -0.16 | 0.18 | -0.72 | -0.20 | 0.12 | 0.00 | -0.28 | -0.21 | 0.08 | 0.02 | 0.54 |
| IEQ3 | 0.62 | -0.68 | 0.02 | -0.58 | 0.01 | 0.27 | -0.63 | -0.01 | 0.22 | 0.01 | -0.13 | 0.00 | 0.10 | -0.06 | 0.61 |
| IEQ4 | 0.54 | -0.66 | 0.11 | -0.56 | 0.11 | 0.27 | -0.66 | 0.06 | 0.20 | 0.07 | -0.13 | 0.07 | 0.11 | -0.01 | 0.61 |
| IEQ5 | 0.52 | -0.29 | -0.42 | -0.25 | -0.44 | 0.07 | -0.13 | -0.37 | 0.09 | -0.26 | -0.17 | -0.36 | 0.06 | -0.22 | 0.07 |
| IEQ6 | 0.38 | 0.07 | -0.80 | 0.09 | -0.79 | -0.01 | 0.03 | -0.79 | -0.03 | -0.03 | 0.03 | -0.80 | 0.00 | 0.00 | -0.04 |
| IEQ7 | 0.53 | -0.30 | -0.43 | -0.28 | -0.46 | 0.00 | -0.23 | -0.42 | -0.01 | -0.16 | -0.24 | -0.44 | 0.02 | -0.05 | 0.08 |
| IEQ8 | 0.25 | 0.18 | -0.73 | 0.21 | -0.71 | 0.04 | 0.08 | -0.74 | 0.01 | 0.08 | 0.12 | -0.75 | 0.06 | 0.06 | -0.07 |
| IEQ9 | 0.14 | 0.22 | -0.59 | 0.24 | -0.56 | 0.08 | 0.00 | -0.65 | 0.02 | 0.28 | 0.22 | -0.65 | 0.04 | 0.17 | 0.04 |
| IEQ10 | 0.29 | -0.14 | -0.29 | -0.13 | -0.29 | 0.00 | -0.26 | -0.34 | -0.05 | 0.13 | -0.12 | -0.39 | 0.02 | 0.21 | 0.11 |
| IEQ11 | 0.54 | -0.48 | -0.16 | -0.52 | -0.22 | -0.15 | -0.16 | -0.08 | -0.08 | -0.54 | -0.59 | -0.12 | -0.01 | -0.22 | -0.13 |
| IEQ12 | 0.54 | -0.17 | -0.68 | -0.15 | -0.70 | -0.02 | -0.06 | -0.64 | 0.00 | -0.23 | -0.08 | -0.63 | -0.04 | -0.20 | 0.08 |
| IEQ13 | 0.66 | -0.59 | -0.17 | -0.57 | -0.21 | 0.01 | -0.27 | -0.09 | 0.06 | -0.50 | -0.57 | -0.11 | 0.10 | -0.24 | -0.01 |
| IEQ14 | 0.56 | -0.22 | -0.62 | -0.21 | -0.66 | -0.08 | -0.05 | -0.56 | -0.04 | -0.34 | -0.16 | -0.55 | -0.08 | -0.27 | 0.03 |
| IEQ15 | 0.48 | -0.21 | -0.49 | -0.19 | -0.51 | 0.00 | **-0.01** | **-0.42** | **0.04** | **-0.33** | -0.12 | -0.40 | -0.02 | -0.29 | 0.03 |
| IEQ16 | 0.35 | -0.43 | 0.09 | -0.48 | 0.06 | -0.12 | -0.33 | 0.10 | -0.11 | -0.22 | -0.46 | 0.05 | -0.03 | 0.03 | 0.05 |
| IEQ17 | 0.36 | -0.38 | 0.00 | -0.13 | 0.07 | 0.83 | -0.19 | 0.06 | 0.80 | 0.04 | -0.06 | 0.03 | 0.79 | 0.03 | 0.16 |
| IEQ18 | 0.10 | -0.15 | 0.05 | **-0.32** | **0.02** | **-0.47** | **-0.37** | **-0.03** | **-0.51** | **0.08** | **-0.47** | **-0.12** | **-0.31** | **0.39** | **-0.06** |
| IEQ19 | 0.73 | -0.79 | 0.00 | -0.75 | -0.04 | 0.09 | -0.62 | -0.01 | 0.08 | -0.23 | **-0.46** | **-0.03** | **0.06** | **-0.09** | **0.38** |
| IEQ20 | 0.26 | -0.21 | -0.08 | 0.04 | -0.02 | 0.83 | 0.01 | -0.01 | 0.83 | 0.00 | -0.06 | -0.04 | 0.89 | 0.03 | -0.06 |
| IEQ21 | 0.41 | -0.52 | 0.11 | -0.52 | 0.07 | -0.03 | -0.41 | 0.09 | -0.04 | -0.17 | -0.30 | 0.08 | -0.06 | -0.07 | 0.27 |
| IEQ22 | 0.29 | -0.41 | 0.13 | -0.53 | 0.09 | -0.34 | **-0.58** | **0.03** | **-0.40** | **0.07** | **-0.17** | **0.05** | **-0.45** | **0.06** | **0.48** |
| IEQ23 | 0.56 | -0.47 | -0.20 | -0.46 | -0.24 | -0.02 | -0.17 | -0.13 | 0.02 | -0.46 | **-0.35** | **-0.11** | **-0.03** | **-0.36** | **0.07** |
| IEQ24 | 0.53 | -0.58 | 0.03 | -0.55 | -0.02 | 0.04 | -0.28 | 0.09 | 0.08 | -0.43 | -0.51 | 0.07 | 0.10 | -0.19 | 0.04 |
| IEQ25 | 0.34 | -0.34 | -0.03 | -0.21 | -0.02 | 0.37 | -0.35 | -0.07 | 0.32 | 0.14 | 0.28 | -0.01 | 0.12 | -0.12 | 0.63 |
| IEQ26 | 0.33 | -0.25 | -0.15 | -0.06 | -0.12 | 0.54 | -0.01 | -0.08 | 0.56 | -0.12 | 0.15 | -0.04 | 0.40 | -0.26 | 0.23 |
| IEQ27 | 0.47 | **-0.31** | **-0.31** | -0.27 | -0.33 | 0.04 | 0.06 | -0.18 | 0.11 | -0.53 | -0.09 | -0.11 | -0.06 | -0.62 | 0.07 |
| IEQ28 | 0.44 | -0.53 | 0.08 | -0.57 | 0.05 | -0.10 | **-0.36** | **0.11** | **-0.08** | **-0.32** | -0.56 | 0.07 | -0.01 | -0.04 | 0.05 |
| IEQ29 | 0.67 | -0.83 | 0.16 | -0.85 | 0.12 | -0.05 | -0.71 | 0.15 | -0.06 | -0.24 | -0.76 | 0.09 | 0.06 | 0.09 | 0.20 |
| IEQ30 | 0.48 | -0.34 | -0.27 | -0.31 | -0.28 | 0.07 | -0.20 | -0.26 | 0.06 | -0.22 | -0.21 | -0.27 | 0.06 | -0.17 | 0.09 |
| IEQ31 | 0.59 | -0.66 | -0.04 | -0.64 | -0.07 | 0.05 | -0.52 | -0.02 | 0.02 | -0.24 | -0.53 | -0.08 | 0.11 | -0.04 | 0.18 |

Table Multidimensional exploratory factor analysis results. The numbers in the column headers refer firstly to the number of specified factors, and then to the specific factor within each model, e.g. the column 4F-F2 refers to Factor 2 of the 4 Factor model. Items that cross-load on multiple factors are emboldened. Loadings above the threshold of 0.3 are colored red for negative values and green for positive values.

As IRT analysis is an inherently stochastic and black box process, there is a possibility that the solutions it finds are unstable, although repeated runs of the IRT analysis did produce consistent results. Nonetheless, for robustness, we therefore also replicated the analysis in SPSS. The resulting factor structures consistently had more cross-loadings than those produced by *mirt* but if items were allocated to factors based on the strongest loading, the factor structures were essentially identical. Thus, *mirt* actually produced factor structures in agreement with, but clearer than the traditional techniques implemented in SPSS.

### Interpretation

As can be seen from and , in choosing between these models there is a trade-off to be made between maximising the number of original scale items loaded on the model and minimising the number of cross-loadings. We rejected 1 and 2 factor models because they account for a low number of items (24 and 27 respectively). We rejected the 5-factor model due to 5 cross-loadings. Although the 3-factor model does not load 1 item (IEQ10), that item is just below the threshold of loading onto F2. The 3-factor model is also superior to the 4-factor model due to having only two cross-loadings which makes it much easier to conceptualise those three factors as variants of the original IEQ subfactors. Together with the consideration of parsimony, we concluded that the 3-factor model is the optimum choice from our multi-dimensional factor analysis.

|  |  |  |
| --- | --- | --- |
| Factors | Loaded items / 31 | Cross loadings |
| 1 | 25 | 0 |
| 2 | 27 | 1 |
| 3 | 30 | 2 |
| 4 | 30 | 5 |
| 5 | 30 | 5 |

Table Multi-factor analysis loaded items and cross-loadings

Inspecting the words used in the items of each factor and resolving the issues represented in Table 2, the factors were defined as:

* F1: *Involvement*, where item 1 was taken to load most strongly on this factor
* F2: *Real-world Dissociation* (RWD), where item 10 assigned to this factor as nearly loading at the 0.3 threshold
* F3: *Challenge*, where item 17 was taken to load most strongly on this factor

It is notable that the first two factors align strongly with Agrawal et al. (2019)’s definition which focuses on involvement leading to dissociation. The extra Challenge factor may arise as the game specific experience of immersion as opposed to immersion from watching a film or reading a book.

A reliability analysis on these factors showed that the Cronbach alpha of each factor was: Involvement, 0.82, RWD 0.78, Challenge 0.36. Further summaries of the item-total correlations and alpha-if-dropped are given in . The low alpha of Challenge was concerning. Shorter scales tend to have a lower reliability (Rammstedt & Beierlein, 2014) but further consideration shows that item 18, which already cross-loaded in the 3F model, was problematic. Its alpha-if-dropped raised the reliability to 0.68 and its correlation with the Challenge factor was only 0.13.

|  |  |  |  |
| --- | --- | --- | --- |
| **Factor** | **α (95% CI)** | **Item-total r range** | **α if dropped range** |
| Involvement | 0.83 (0.82-0.84) | 0.36-0.68 | 0.81-0.83 |
| RWD | 0.78 (0.77-0.80) | 0.38-0.72 | 0.74-0.79 |
| Challenge | 0.36 (0.31-0.41) | 0.13-0.71 | 0.14-0.69 |
| Challenge w/o 18 | 0.68 (0.66-0.71) | 0.60-0.78 | 0.55-0.70 |

Table Reliability analysis for first split half

However, the loading of item 18 was quite high on the Challenge factor and conceptually it does fit with identifying Challenge: “There were times in the game in which I just wanted to give up”.

We therefore decided to retain item 18 in our confirmatory analysis. With classical test theory approaches such as Principal Component Analysis, weak items can undermine the effectiveness of the method. However, with IRT, because it models each item separately, it is more robust to including weak or unrepresentative items (Embretson & Reise, 2013).

### Confirmatory factor analysis

To confirm that the previous factor analysis was not an artefact of particular data or over-analysis we performed a confirmatory factor analysis on the other half of the large set from Perrett (2018).

We performed two different types of analysis on this half of the data; a confirmatory factor analysis using IRT multi-factor analysis for the 3 factors identified and a confirmatory bi-factor analysis (Reise, 2012). The standard 3 factor confirmatory model showed that all items loaded above 0.3 (or below -0.3) on their respective factors, with the exception of item 10, which loaded again just below this threshold at 0.281. This strongly suggested that item 10 is a weak item overall in the IEQ.

Reliability analysis on this second subset of data gave very similar results () to the first subset with item 18 again obviously weakening the Challenge factor. An examination of infit and outfit statistics (Greer & Liu, 2016) did not raise any concerns as all infit and outfit values were in the range 0.55 to 1.1. The reliability of the Challenge factor (excluding item 18) is lower than that of the other two factors but this difference is most likely an artefact of it simply being shorter than the other two.

|  |  |  |  |
| --- | --- | --- | --- |
| **Factor** | **α (95% CI)** | **Item-total r range** | **α if dropped range** |
| Involvement | 0.82 (0.81-0.83) | 0.36-0.66 | 0.81-0.83 |
| RWD | 0.76 (0.74-0.78) | 0.35-0.71 | 0.72-0.76 |
| Challenge | 0.32 (0.27-0.37) | 0.17-0.70 | 0.10-0.63 |
| Challenge w/o 18 | 0.63 (0.60-0.66) | 0.55-0.76 | 0.48-0.67 |

Table Reliability analysis for second split half

The bi-factor analysis fits a model where each item loads on a putative single common dimension (g), in this case immersion, and a second factor which is the specific aspect of immersion proposed by our previous analysis. The bi-factor analysis produces a “communality value” (h2) being the degree to which the variance of each question item is accounted for by the model as a whole. A good threshold for h2 is to be above 0.1 because this corresponds to a single factor loading of 0.3 or above (0.32 ≈ 0.1)

The bi-factor analysis suggests a good model fit, see . The h2 values are all above 0.1 suggesting the model can account for an appreciable level of variance in every item. Furthermore, most items load on the general factor, g, suggesting that there is good reason to think of the questionnaire as pertaining to a single underlying concept. Where items do not load strongly on g, that is below 0.3, they have a stronger loading on their specific factor. The exception is item 10, which was noted as weak in the exploratory analysis.

This suggests that, overall, the model does capture both the general concept of immersion and that the separate factors of Involvement, RWD and Challenge do add further specific aspects of immersion that can be differentiated from the general factor.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | g | Involvement | RWD | Challenge | h2 |
| IEQ1 | 0.72 | 0.01 |  |  | 0.52 |
| IEQ2 | 0.79 | 0.11 |  |  | 0.63 |
| IEQ3 | 0.65 | 0.22 |  |  | 0.47 |
| IEQ4 | 0.62 | 0.19 |  |  | 0.42 |
| IEQ5 | 0.47 |  | 0.32 |  | 0.33 |
| IEQ6 | -0.17 |  | -0.77 |  | 0.63 |
| IEQ7 | 0.41 |  | 0.42 |  | 0.34 |
| IEQ8 | 0.03 |  | -0.73 |  | 0.53 |
| IEQ9 | 0.09 |  | -0.63 |  | 0.41 |
| IEQ10 | -0.22 |  | -0.24 |  | 0.11 |
| IEQ11 | 0.49 | 0.27 |  |  | 0.31 |
| IEQ12 | 0.41 |  | 0.62 |  | 0.55 |
| IEQ13 | 0.58 | 0.25 |  |  | 0.40 |
| IEQ14 | 0.46 |  | 0.52 |  | 0.48 |
| IEQ15 | 0.39 |  | 0.37 |  | 0.28 |
| IEQ16 | 0.30 | 0.35 |  |  | 0.21 |
| IEQ17 | 0.35 |  |  | 0.75 | 0.69 |
| IEQ18 | -0.09 |  |  | 0.38 | 0.16 |
| IEQ19 | 0.63 | 0.4 |  |  | 0.56 |
| IEQ20 | -0.17 |  |  | -0.77 | 0.62 |
| IEQ21 | 0.32 | 0.36 |  |  | 0.23 |
| IEQ22 | 0.27 | 0.32 |  |  | 0.17 |
| IEQ23 | 0.46 | 0.19 |  |  | 0.25 |
| IEQ24 | 0.45 | 0.33 |  |  | 0.31 |
| IEQ25 | 0.33 |  |  | 0.24 | 0.16 |
| IEQ26 | 0.36 |  |  | 0.47 | 0.35 |
| IEQ27 | 0.44 |  | 0.19 |  | 0.23 |
| IEQ28 | 0.41 | 0.44 |  |  | 0.36 |
| IEQ29 | 0.52 | 0.73 |  |  | 0.80 |
| IEQ30 | 0.42 | -0.01 |  |  | 0.17 |
| IEQ31 | 0.47 | 0.59 |  |  | 0.57 |

Table Multidimensional item response theory confirmatory bifactor analysis results for a 3-factor defined structure. h2 communality coefficients are reported with respect to within defined factors combined with the general IEQ construct. Blank values correspond to a value of 0 as items were only assigned one factor in the model definition.

## **Step 3**: Develop a candidate short-form

Having determined that the underlying structure behind the full IEQ questionnaire has 3 distinct factors that could be aggregated into a single unified factor g, we then needed to determine a candidate short form scale which could be validated to ensure it meets quality requirements such as validity and reproducibility. When creating a short form scale there is often a trade-off between keeping the questionnaire as short as possible and quality. Ideally our short form should be no longer than existing short forms such as miniPXI (11 items) and UES-SF (12 items) whilst still maintaining quality. We therefore aimed to select four items per factor.

To determine which questions to include in the short form, we chose those questions for each factor which had the highest loading onto that factor in the exploratory analysis and had high h2 values in the confirmatory bifactor analysis. We then considered the items themselves in relation to whether they had similar wordings or any ambiguity as to how the items could be interpreted in a variety of games. For the Involvement factor, items 1, 2, 3, 13, 19, 29 and 31 had suitably high loadings and h2 values. However, item 31 is about wanting to play the game again and some games do not lend themselves to replay, it was also likely to be close in meaning to item 29 (that asks about enjoyment). Item 13 was retained because it refers to the game as an experience which we felt was a unique conceptual facet of involvement in a game. The remaining three items were determined by statistical analysis rather than conceptual analysis. Thus, item 19 was retained because it had high loadings. That left items 1, 2 and 3 which were all good candidates distinct from the others and each had different key terms: item 1 refers to attention, item 2 to focus and item 3 to effort. We therefore looked more carefully at the statistical relationship between each item and the potential subscale. By considering each item in turn with items 13, 19 and 29, item 2 gave the highest Cronbach alpha to the short scale and had the highest correlation with all of the other items considered. We also considered pairs of items to potentially give a short form of five items but no pairing gave a markedly improved Cronbach alpha, We, therefore selected item 2 as the fourth item to complete the short form of the scale.

For the RWD factor, based on factor loadings and h2 values, items 6, 7, 8, 9, 12, 14 and 15 were considered. All items do seem well linked to the concept of real-world dissociation but item 8 has been previously found to have ambiguity with participants being unsure if the surroundings referred to in-game or real-world surroundings. Item 9 was dropped as being similar in content to item 6 and with slightly less strong loadings. Finally, item 14 was not selected because of the comparison that it asked participants to make rather than asking them directly about the experience as felt.

For the Challenge factor, with only five items to choose from, only item 18 was dropped because of the issues with it in the previous analysis. The reliability analysis of the short forms is given in .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Short Factor** | **Items from IEQ** | **α (95% CI)** | **Item-total r range** | **α if dropped range** |
| Involvement | 2, 13, 19, 29 | 0.67 (0.65-0.70) | 0.51-0.67 | 0.59-0.67 |
| RWD | 6, 7, 12, 15 | 0.67 (0.64-0.69) | 0.49-0.68 | 0.53-0.63 |
| Challenge | 17, 20, 25, 26 | 0.63 (0.60-0.66) | 0.25-0.41 | 0.48-0.67 |
| Challenge v2 | 17, 20, 26 | 0.67 (0.64-0.70) | 0.47-0.69 | 0.5-0.72 |

Table The items from the full IEQ chosen to constitute the Short Form together with the Cronbach alpha and alpha-if-dropped range for the Short Form factors.

In the initial four-item version of the Challenge factor, item 25 had a low correlation of only 0.25 with the scale overall and if dropped, the Cronbach alpha of the short scale increased. The phrasing of the item is about wanting to “win” the game which does not necessarily apply to some games such as resource management games. Given these issues, we therefore chose to drop this item. In the three-item version of Challenge, item 26 did have an increased alpha-if-dropped but correlates well with the scale and the other two items. Its phrasing also does capture a sense of challenge. Further, removing it would leave only two items in the shortened Challenge scale which would be undesirably short (Cairns, 2013).

This then gives the 11-item short form shown in .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **IEQ-SF Item** | **IEQ Item** | **Content** | **Factor** | **Reverse scored?** |
| 1 | 2 | I felt focused on the game. | Involvement | No |
| 2 | 13 | The game was something that I wasexperiencing, rather than just doing. | Involvement | No |
| 3 | 19 | I felt motivated when playing the game. | Involvement | No |
| 4 | 29 | I enjoyed playing the game. | Involvement | No |
| 5 | 6 | I felt consciously aware of beingin the real world whilst playing. | Real world dissociation | Yes |
| 6 | 7 | I forgot about my everyday concerns. | Real world dissociation | No |
| 7 | 12 | I felt that I was separatedfrom the real-world environment. | Real world dissociation | No |
| 8 | 15 | I found myself so involved thatI was unaware I was using controls | Real world dissociation | No |
| 9 | 17 | I found the game challenging. | Challenge | No |
| 10 | 20 | I found the game easy. | Challenge | Yes |
| 11 | 26 | I felt in suspense about whether ornot I would do well in the game. | Challenge | No |

Table The 11 item IEQ Short Form scale.

### A preliminary sensitivity analysis

It is important to understand whether the short form is able to detect differences in immersion in a way that is similar to the full IEQ. The real test of this is that it performs well on new data as will be demonstrated in the subsequent validation studies. However, as a preliminary test, it is useful to see that, in principle, the short form had the sensitivity to detect potentially small differences in immersion picked up by the full IEQ.

To do this, we ran a bootstrap simulation based on the second half of the data from (Perrett, 2018). A power calculation showed that for a t-test with significance at 0.05, two samples of 175 should detect small effect sizes, Cohen’s d = 0.3, with a power of 0.8 (which is typically assumed). We therefore generated pairs of samples by bootstrapping with resampling from the data. Pairs whose difference in IEQ scores corresponded to an effect size ranging from d = 0.30 to d = 0.35 were retained until there were 1000 such pairs. The candidate short form scores were then generated for each pair and a t-test was used to evaluate whether the two samples in a pair were significantly different. Of the 1000 pairs, 879 of the pairs were significantly different based on the short form alone. This corresponds to a power of 0.879. This gives good confidence that the short form has similar power to the full IEQ in detecting small effect sizes, at least in simulation. Furthermore, because of the bootstrapping process, the means of the samples were very close to 3.5, that is, very much not at the extreme ends of immersion and where detecting small differences might be most challenging. Overall, then, this preliminary evaluation provides some confidence that the short form has good sensitivity for small effect sizes.

## **Steps 4 and 5:** Confirm and validate the Short Form on new data.

As recommended by Smith et al. (2000), the short form needs to be validated in two ways. First, the factor structure should be confirmed on previously unseen data and secondly that the short form should reproduce the findings of studies that used the long form. Fortunately, due to the adoption of Open Science practices, we could access an existing dataset that used the full IEQ. This formed the basis for the first confirmation and validation, which was pre-registered ahead of the analysis, and we then replicated this study, also as a pre-registration, using only the short form to show that it would work correctly independently of the other IEQ items. We also validated the short form on two additional game immersion data sets.

In both studies, to address the first goal of confirming the factor structure, our criteria were that all items load satisfactorily on their corresponding factors and that the reliability and fit statistics remained in accordance with the criteria used previously.

To address the second goal that the IEQ-SF behaves like the full IEQ, there are no explicit criteria in the literature for what it means in practice to reproduce the findings from using the long form. Smith et al. (2000) warn against correlating the short form with all or part of the long form from which it was extracted so this was not done. As typically the IEQ is used to assess differences in players’ immersion in experimental settings, ideally the IEQ-SF should be able to produce effect sizes similar to those seen by the IEQ. In these validation studies we operationalise “statistically similar” as the situation where the effect size as measured by the IEQ-SF lies within the 95% confidence interval of the effect size as measure by the IEQ and that conversely, the IEQ effect size lies within the 95% confidence interval of the IEQ-SF effect size. Meeting these criteria would mean that the IEQ-SF has been validated in an experimental setting as producing a similar outcome to the IEQ.

### Validation Study 1

The first validation study was done on an existing Open Science data set from Cutting et al. (2020). The data was from experiment 2 in this paper, which was a between-participants study with two conditions, each with 80 participants, making 160 in total (77 male, 81 female and 2 non-binary, age range 18–40, mean age = 25.8). In each condition participants played either an engaging or a less engaging version of the self-paced puzzle game *Two Dots* and then filled in the full IEQ. In the original study, a t-test analysis found a significant difference in immersion between conditions with a large effect size. We were able to use this data to validate the new 11-item short form IEQ because the short form is a subset of the longer form and so the full IEQ data contains all the items for new form. To validate our new short form, we repeated the original analysis which compared immersion between conditions but used our new short form IEQ instead of the original full-length questionnaire.

This validation study employed another Open Science technique: “pre-registered analysis of existing data” (Munafò et al., 2017). This involves pre-registering the study analysis and hypotheses to be used in the analysis of an existing data set, by specifying both analysis and hypotheses before performing them, this ensures that neither of them can be changed post-analysis and that the results can be interpreted as data independent confirmatory research without the possibility of “p-hacking” (Simmons et al., 2016). Pre-registering our validation is a more severe test of the new short form because it ensures that neither the short form nor analysis can be changed after the results are known.

#### Design and Materials

Cutting et al. (2020)’s study had a between participants design with 2 conditions. In both conditions participants played a game for 5 minutes and then filled in the IEQ to record their game experience. The game chosen was a clone of *Two Dots,* a commercially successful self-paced mobile puzzle game. *Two Dots* features a grid of different coloured dots and players must draw lines between dots of the same colour. The connected dots then disappear and new dots drop down from the top of the screen. The aim of the game is to remove a pre-set number of dots of each colour within a move limit.

There were two different versions of the game and the independent variable between conditions was the version of the game that the participant played. One version was a straight clone of the first ten levels of *Two Dots. Two Dots* is known to be engaging which is likely to lead to high levels of immersion so this version is known as the *high immersion* game (see Figure 1). The other version of the game has the same grid of dots which can be joined and drop down; however, all the dots are the same colour and there is no move limit. These changes make this version of the game considerably less engaging which is likely to reduce immersion, hence this version is known as the *low immersion* game (See ).



Figure The high immersion version of Two Dots (Cutting et al., 2020). Players have to join the circles (known as dots) which are the same colour in order to meet the targets at the top of the screen. Cutting et al put pictures into the dots to study attentional processes during the game, but they are irrelevant to the game task.



Figure The low immersion version of Two Dots. This is the same as the high immersion version above but all the dots are the same and there is no move limit.

#### Pre-registered hypotheses and analysis

The original study (Cutting et al., 2020) which generated the data had a main hypothesis that immersion would be significantly different between high and low immersion game conditions. In the new IEQ-SF the three factors are equally as important as the unified immersion metric so our hypotheses for this study included both comparisons between conditions of the overall immersion metric as the individual factors.

The main hypothesis for the original experiment was that the immersion score would be significantly different between the low and high immersion condition. However, as the new IEQ-SF places a stronger emphasis on the 3 individual factors our main hypothesis for this study was that there would be a significant difference between conditions in both the overall immersion score and the individual immersion factor scores as measured by the IEQ-SF. The high immersion game was hypothesised to have a significantly higher scores for all measures than the low immersion game.
(NB We pre-registered the hypotheses that the individual factors would be different, but due to an error we did not pre-register the hypothesis on the overall immersion score comparison).

#### Results

This study aimed to both confirm the factor structure of the IEQ-SF and validate it against the original IEQ. Therefore, we first performed a confirmatory factor analysis and then analysed how well the IEQ-SF reproduced the original IEQ findings.

##### Confirmatory factor analysis

The confirmatory factor analysis gave good support for the 3-factor structure of the IEQ-SF. All loadings in the 3-factor analysis were above 0.45 with h2 values all above 0.4 with the exception of 3 items whose h2 values were above 0.2. The corresponding bifactor analysis showed good support for the g factor of Immersion with all but item IEQ6 (item 5 of IEQ-SF) loading strongly on g and all items having h2 values above 0.35. The reliability statistics shown in are very similar to those found when developing the short form. Overall, then this confirmatory model meets our acceptability criteria used previously.

All but two items gave infit and outfit statistics from the bifactor model in the desired range of 0.5 to 1.5. The exceptional items were item 5 (IEQ6) where infit was 0.47 and outfit 0.48 and item 9 (IEQ17) where outfit was 0.49 (and infit 0.79). These are close to the nominal threshold of 0.5 but this does suggest item 5 may not be closely fitting the proposed short form model. The item asks players about their awareness of the real world whilst playing so it may be that in the experimental context of this study that players were very aware that they are also participants and this makes this item respond differently.

|  |  |  |  |
| --- | --- | --- | --- |
| **Short Factor** | **α (95% CI)** | **Item-total r range** | **α if dropped range** |
| Involvement | 0.72 (0.65-0.79) | 0.68-0.82 | 0.60-0.74 |
| RWD | 0.69 (0.60-0.76) | 0.63-0.81 | 0.52-0.71 |
| Challenge | 0.71 (0.62-0.78) | 0.72-0.89 | 0.41-0.77 |
| IEQ-SF | 0.82 (0.78-0.86) | 0.35-0.75 | 0.79-0.83 |

Table The reliability statistics for the IEQ-SF and its factors in Validation Study 1

##### Reproduction of experimental findings

The results of the study were analysed according to the pre-registration of this analysis. T-tests were used to examine significance in each factor and the whole IEQ-SF.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Factor | Low Immersion ConditionMean (SD) | HighImmersion ConditionMean (SD) | *p* | *t(158)* | Effect Size (d) | Effect Size CIs |
| Involvement | 3.15 (0.75) | 4.03 (0.56) | <0.001 | -8.38 | *d* = *-*1.325 | [-1.670, -0.980] |
| RWD | 2.83 (0.85) | 3.33 (0.74) | <0.001 | -3.977 | *d* = *-*0.629 | [-0.949, -0.309] |
| Challenge | 2.47 (0.97) | 3.20 (0.76) | <0.001 | -5.269 | *d* = *-*0.833 | [-1.158, -0.508] |
| IEQ-SF | 2.85 (0.63) | 3.55 (0.53) | <0.001 | -7.594 | *d* = -1.200  | [-1.540, -0.861] |

Table Analysis of IEQ-SF factors and unified immersions core. The high immersion game was significantly higher in all measures than the low immersion game

Our main hypothesis was supported in that the overall immersion score and all three subfactors of the IEQ-SF were a significantly greater in the high immersion game than the low immersion game. (See ). For comparison, the original study from Cutting et al found a significant difference between conditions in immersion as measured by the full IEQ (*t(158)* = *-*7.854, *p <* .001) with a substantial effect size (*d* = *-*1.242, [*-*1.583, *-*0.9]). Thus, the IEQ-SF is producing an effect size and confidence interval that is in statistically similar to the full IEQ.

#### Discussion

This validation analysis of the new IEQ-SF used existing data from a previously published study which used the original IEQ. The original publication found a significant difference in IEQ between conditions, with a strong effect size (d=*-*1.242) and a statistically similar result (d=-1.200) was also found with the IEQ-SF which provides strong validation that the new IEQ-SF works well as a replacement for the IEQ in experimental contexts. The effect sizes observed with the IEQ-SF factors range from roughly moderate (*d* = -0.629) to very strong (*d* = *-*1.325), which are also comparable to the originally reported effect size. These results support the claim that the IEQ-SF is a viable alternative to the full IEQ, with minimal loss to power.

In addition, Cronbach *a* coefficients for each factor support our goal that the IEQ-SF would be adequately reliable in a real-world use despite the reduced number of items.

However, this validation may be limited by using IEQ-SF items derived from a full IEQ dataset that had been collected for another purpose. It is possible that a case could be made that the process of completing the full IEQ that may have been transferred some quality to these IEQ-SF scores that would otherwise not exist if they only completed the IEQ-SF. Perhaps some items in the full IEQ can inform participants about others, either by a cueing effect or by task familiarisation. To overcome this challenge to the validity of the IEQ-SF, it is important that it is validated independently from the full set of IEQ items (Smith et al., 2000).

### Validation Study 2

To address the issues of Validation Study 1, we performed an additional validation study which collected new data and which only asked participants to complete the IEQ-SF without the additional items from the original IEQ but otherwise was a replication of the same Study 2 from Cutting et al. (2020)ie Validation Study 1. This study took advantage of another Open Science practice, namely the sharing of experimental source code. By modifying the same source code as used in the original study we were able to replicate the study exactly, apart from the change in questionnaire used.

#### Study design

As a replication of Study 2 from Cutting et al. (2020) the study was identical in all ways to the original study except that, after playing the game, participants completed the IEQ-SF rather than the full IEQ. As such, this study was also a between-participants study with two conditions. The main experimental manipulation was the level of immersion experienced in different versions of the game *Two Dots.* The full design and materials are the same as those described previously in section *Validation Study 1*. The study was approved through the ethical approval procedures of the ANONYMISED FOR REVIEW.

Participants

We aimed to have 160 participants as this was the number of participants in the original experiment. Participants were recruited via the Prolific online recruitment platform and paid ANONYMISED FOR REVIEW.

In total, 206 participants were recruited for the main experiment, of which 7 participants were omitted for colour blindness, 1 participant was omitted for not completing the tutorial, 14 participants were omitted for taking a break during the game, 6 were omitted for failing the attention check dummy question, 12 participants were omitted due to faulty behaviour of the study (bugs and glitches, etc), and 2 participants’ data were lost due to technical issues. The final remaining sample was *N* = 164 participants, with 83 participants in the low immersion condition and 81 participants in the high immersion condition. Participants were 83 male, 80 female and 1 non-binary, the age range was 18–41, with the mean age = 27.1.

#### Procedure

Participants were allocated to the high or low immersion conditions in turn, so the first participant did the high immersion condition, the second the low immersion and so on. After completing a consent procedure, they were taken through a short tutorial on how to play the game. They then played the version of *Two Dots* which was appropriate to the experimental condition. During both games in the original study, participants were also shown 60 different icons from the *Webdings* typeface. Immediately after the game, they were tested on how many of these icons they recognised. Our replication kept these elements in the study, but as the image recognition data was not relevant to this validation, we did not analyse it further. After completing the icon test, participants completed the IEQ-SF which also included a dummy question to check participants were answering the questions in good faith.

#### Results

As with Validation Study 1, we first performed a confirmatory factor analysis and then analysed whether the new IEQ-SF replicated the same hypothesis tests as the original study with similar effect sizes

##### Confirmatory factor analysis

For this study, the confirmatory factor analysis also showed good support for the 3-factor model. All factor loadings, but one, were above 0.5, the exception being item 11 where the loading was 0.36, and correspondingly all h2 values were above 0.3 except for item 11 with a value of 0.130. The corresponding bifactor analysis showed good support for the g factor of Immersion with all but items 5 and 10 loading strongly on g (loadings of 0.11 and 0.21 respectively) and all items having h2 values above 0.3. The reliability statistics shown in are very comparable to those found when developing the short form.

|  |  |  |  |
| --- | --- | --- | --- |
| **Short Factor** | **α (95% CI)** | **Item-total r range** | **α if dropped range** |
| Involvement | 0.77 (0.70-0.82) | 0.70-0.85 | 0.64-0.77 |
| RWD | 0.75 (0.68-0.81) | 0.68-0.86 | 0.59-0.75 |
| Challenge | 0.68 (0.58-0.75) | 0.69-0.85 | 0.42-0.82 |
| IEQ-SF | 0.82 (0.77-0.86) | 0.40-0.69 | 0.79-0.82 |

Table The reliability statistics for the IEQ-SF and its factors in Validation Study 2

For all items in the bifactor analysis, the infit value was between 0.5 and 1.5 though for items 3 and 7, the outfit values were 0.43 and 0.41 respectively. Across these analyses, there does not seem to be any item that is a consistent source of concern and the issues seen in Validation Study 1 with item 5 were not seen in this study despite the replication of the study design.

##### Reproduction of experimental findings

As in Validation 1, our analysis took into account that in the new IEQ-SF, the 3 constituent factors of Involvement, Real-World Dissociation and Challenge are as important as the combined overall immersion score.
Thus, we tested four hypotheses for this study, namely that in the high immersion condition there would be a significant difference between all 3 of these factors and also the overall immersion score. These results are shown in .

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Factor | Low Immersion ConditionMean (SD) | HighImmersion ConditionMean (SD) | *p* | *t(162)* | Effect Size (d) | Effect Size CIs |
| Involvement | 3.38 (0.67) | 4.08 (0.34) | <0.001  | -6.22  | *-*0.97  | [-1.29, -0.64]  |
| RWD | 2.70 (0.88) | 3.07 (0.9) | 0.007 | -2.708 | -0.423  | [-0.735, -0.111] |
| Challenge | 2.22 (0.89) | 3.14 (0.69) | <0.001 | -7.357 | -1.149  | [-1.482, -0.816] |
| IEQ-SF | 2.82 (0.63) | 3.46 (0.56) | <0.001 | 6.874 | -1.073  | [-1.403, -0.744] |

Table Analysis of IEQ-SF factors and unified immersions core. The high immersion game was significantly higher in all measures than the low immersion game

#### Discussion

This second validation of the IEQ-SF replicated an existing IEQ study with new data provided by a new set of participants, but using the IEQ-SF in place of the original IEQ. The original study found a significant difference in IEQ between conditions, with a strong effect size (d=*-*1.242) and we found a similarly large effect between conditions (d=-1.073). In particular, it should be noted that the 95% confidence interval for this effect size ([-1.403, -0.744) encompasses the effect size of the original study and vice versa. Thus, the two studies give statistically similar effect sizes for the experimental manipulation on immersion as measured by the IEQ-SF. The effect sizes observed with the factors of the IEQ-SF range from roughly moderate (*d* = -0.423) to very strong (*d* = *-*1.149). These effects are also very similar to those in Validation Study 1. Again, consideration of the 95% CI of the effect sizes show that the two studies have statistically similar effect sizes for each individual factor. The results of this second validation therefore support the claim that the IEQ-SF is a viable alternative to the full IEQ, with minimal loss of power.

This replication of the original study with new data also provides a strong validation of the original study results. This also provides support for view that the Open Science practice of sharing code can lead to effective replications and that these replications can be used to develop new measures which can then be effectively validated against existing measures.

### Further validation on two additional datasets

The previous validation studies provide confidence in the use of the IEQ-SF in experimental studies. To add further weight and to vary the context for validating the IEQ-SF, we were able to take advantage once more of Open Science to also consider two further publicly available datasets that used the IEQ, namely studies 6 and 7 from Denisova (2016). These help to broaden the validation as the games were quite different from those used in Validation Studies 1 & 2. Denisova’s Study 6 (n=120, 82 male, 38 female, age range 18-50, mean age =24.2) used an isometric action shooter game and Study 7 (n=42, 41 male, 19 female, age range 18-50, mean age=25.6) a third-person object collection game. In both studies, the dependent variable was immersion as measured by the IEQ and the independent variables were whether the game adapted difficulty in response to player performance (known as *Adaptation)* and the information that players had about this adaptation (known as *Information)*.

The original analysis of the full IEQ and the new analysis of short form and its constituent factors are summarised in for Study 6 and for Study 7.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Effect** | **IEQ** | **Short Form** | **Involvement** | **RWD** | **Challenge** |
| *Adaptation* | 0.164\*\*\*(0.058-0.280) | 0.074\*\*(0.009-0.162) | 0.080\*\* | 0.038\* | 0.008 |
| *Information* | 0.081\*\*\*(0.006-0.173) | 0.067\*(0.001-0.153) | 0.084\*\* | 0.086\*\* | 0.005 |
| *Adaptation* interaction with *Information* | <0.01(0-0.018) | <0.01(0.-0.018) | 0.012 | <0.01 | 0.005 |

Table Effect sizes (partial eta squared) and 95% CI of the full IEQ, the short form and the three factors of the short form for (Denisova, 2016), Study 6. NB all values are effect sizes. Asterisks indicate significance at: \* p<0.05; \*\* p<0.01; \*\*\* p<0.001

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Effect** | **IEQ** | **Short Form** | **Involvement** | **RWD** | **Challenge** |
| *Adaptation* | 0.145\*\*(0.018-0.301) | 0.152\*\*(0.021-0.310) | 0.097\* | 0.019 | 0.110\* |
| *Information* | 0.119\*\*(0.009-0.266) | 0.182\*\*\*(0.035-0.347) | 0.078\* | 0.061 | 0.099\* |
| *Adaptation* interaction with *Information* | 0.018(0.-0.085) | 0.060(0-0.174) | 0.016 | 0.027 | 0.027 |

Table Effect sizes of the full IEQ, the short form and the three factors of the short form for (Denisova et al., 2016), Study 7. NB all values are effect sizes. Asterisks indicate significance at: \* p<0.05; \*\* p<0.01; \*\*\* p<0.001

As can be seen, in both studies, the short form replicated the pattern of significant effects seen with the IEQ as whole. This provides good support that the IEQ-SF is a viable alternative to full IEQ. Although there are differences between the full IEQ and short form in their estimates of effect sizes, which in some cases is considerable, all of the short form effect sizes are within the 95% CI of the full IEQ’s effect size and vice versa with the exception of Adaptation effect in Study 6. For that, the effect size from the IEQ is 0.164 which is outside the Short Form’s 95% CI though it is close to the upper bound of 0.162. Thus, in this case it would seem that the short form has significantly underestimated the original effect size but it is still plausible that a “true” effect sits within the 95% CI of both measures. In terms of the experimental outcome though, the effect is still significant. Overall then, the results from using the IEQ-SF are almost all statistically similar to using the full IEQ. This does suggest that the short form can vary from the effects of the full IEQ but, with sufficient power, we can have some confidence that it can be interpreted to have the same meaning as the IEQ.

It is also worth noting differences in patterns of significance between the three factors of the short form. In Study 6, only the Involvement and RWD factors show significant effects due to the two independent variables whereas in Study 7, it is Involvement and Challenge that show significant effects. This adds support to the previous factor analysis that the factors do constitute distinct facets of immersion that, whilst related, give insights into the different ways in which people experience immersion.

# Overall discussion

Our aim was to develop a new short form of the IEQ which had considerably fewer items than the original questionnaire and also took advantage of advances in scale development and Open Science practices to create a more reliable and valid measurement scale.

As is best practice, we based the new short form on the well validated IEQ. We then used a large data set (n=3539) to confirm that this long form was best represented by 3 factors. Qualitative analysis of conceptual relevance determined that those 3 factors corresponded to the concepts of Involvement, Real World Dissociation and Challenge. We then used a combination of quantitative analysis of reliability and qualitative analysis of the question wording to determine the 11 items that comprised the most appropriate short form. The IEQ-SF omits the Control subfactor, that is, the sense of being in control of what is happening in the game. Although the experience of control may moderate immersion, our factor analysis on a much larger and wider selection of games found that for most games this is unlikely to be a direct link, instead the experience of control impacts the subfactor of RWD which is included in the IEQ-SF.

This short form was validated over four datasets that had not been used in the development of the short form. The first validation used a reanalysis of existing data from published study that used the IEQ and the second collected new data to replicate the same study but using only the short form IEQ. Both studies found that the new IEQ-SF had very similar results to the original IEQ with statistically similar effect sizes and little loss of power. The third and fourth validations similarly used published studies that used the IEQ but in very different games from the previous two validations. Again, the short form gave substantially the same pattern of results as the full IEQ.

These validations collectively provide good evidence that the new 11 item IEQ-SF is a conceptually well-founded and validated effective replacement to the original IEQ whilst satisfying our requirements to be considerably shorter.

The IEQ-SF is thus suitable for use in all cases where the IEQ was previously used, but is particularly suitable for use in studies which need participants to fill in several scales, where the reduced size of the IEQ-SF could help prevent the study becoming too long and participants from suffering questionnaire fatigue. For example, Safar (2024) made use of the IEQ-SF when investigating the effect of game immersion on stress reduction which required repeated measures of stress as well as measuring immersion. The IEQ-SF is particularly suited for use cases such as described by Alexandrovsky et al. (2020) who highlight the value of scales which can be used *within* VR environments where short questionnaires are particular important, especially for repeated measures used to assess to changes in immersion over time or when one participant is assessing multiple study conditions.

As previously mentioned (in section ), Perrett (2018)’s large-scale IEQ study found a mean time of 6.8 seconds to complete each item of the IEQ. As the IEQ-SF has 20 fewer items than the IEQ it may result in a saving of around 2 minutes in completion time. An example use case is Cutting et al. (2020), which used the IEQ for online studies on several hundred participants via the Prolific platform, but would have reduced the time and thus cost of the studies by around 15% if the IEQ-SF had been used instead. We also recommend its use in situations where studies would otherwise choose an unvalidated subset of IEQ items. This includes existing studies on VR (e.g. Hudson et al. (2019) used 4 items of the IEQ and Li and Cesar (2023) used 5 items from the IEQ) and Education (e.g. Kim et al. (2022) used 15 items based on the IEQ and Nicolaidou et al. (2023) used 7 items from the IEQ). The IEQ-SF factor structure was created from a data set containing a wide variety of game genres and the validation encompassed both self-paced casual puzzle games and fast moving action games which together could be seen as representing the majority of games commonly played (Unity Technologies, 2022). However, there may be other game genres, for example, interactive fiction or idle games (Cutting et al., 2019) where the new IEQ-SF may not be quite as appropriate. Indeed, the Film IEQ scale (Rigby et al., 2019) which adapted the IEQ to film found that although film immersion also contains the same Involvement and RWD factors, the Challenge factor is less relevant. Thus, researchers applying the IEQ-SF to new genres of game should be aware that not all of the 3 factors may be equally salient for that particular game experience, but between them they are likely to account for a substantial proportion of the experience of immersion.

Our scale development process made use of a number of Open Science practices which resulted in a better conceptualised and validated scale than would otherwise have been the case. Our initial factor analysis made use of a large shared dataset (Perrett, 2018) of IEQ responses which had been collected for another study. Our first validation study for the scale was based on a reanalysis of another public dataset of IEQ responses to a hypothesis test study (Cutting et al., 2020). The second validation study collected new data, but made use of the publicly shared experiment source code from the same hypothesis test study. This shared source code enabled us to replicate the original complex game interaction and change the questionnaire to our new IEQ-SF, but without having to create the software from scratch.

## Summary of contribution

In summary, we make three main contributions to the field of game experience and scale development. The primary contribution is the creation of a new short form immersion questionnaire; the IEQ-SF. This new scale is 11 items in length, around a third of the length of the original IEQ. Not only is the IEQ-SF shorter than the original IEQ, it has considerably stronger foundations and validation. The original IEQ was developed with 244 participants, whereas the IEQ-SF was developed with 3539 participants and then later validated over four different studies with 486 participants in total. In addition, the IEQ-SF was created using item-response theory which creates a more reliable scale than the classical test theory used for the IEQ. This improved development and validation resulted in a new 3 factor structure for the IEQ-SF, with immersion consisting of the 3 factors of Involvement, Real-world Dissociation and Challenge.

A secondary contribution to the field of scale development is the use of Open Science datasets for scale development and validation. Open Science data sharing is usually motivated by a desire to aid replication and reproducibility. To the best of our knowledge, it has not previously been repurposed for scale development and validation. Reusing data in this way allows for much larger and more varied datasets in scale creation, giving a much stronger basis for scale development and allowing the scale to be validated in a wider variety of contexts. In our case, this provided a large-scale dataset for scale development and further validation on 3 different games without the collection of new data.

Our third main contribution is to aid future scale development by including a detailed model for developing short form scales which includes theoretical justification, methods and analysis scripts. Here we again follow the lead of Open Science in making materials and data available to serve as a foundation for future work and replication. But in addition to this, our paper synthesises recent guidance into a step-by-step best-practice guide to creating short forms (in section ). We also make our full R analysis scripts and validation game source code available for reuse. Together these elements contribute a strong basis for future short form scale development both within the area of game experience and also in the broader field of HCI.

## Limitations

A limitation of our scale development process is that although the feeling of immersion is common in a wide variety of game, two of the validation studies of the new scale used variants of same self-paced puzzle game. However, the other two validation studies used diverse action games and the data set used for the initial factor analysis used a wide variety of different types of digital games with a much larger sample size (n=3539) than the one used to create the original IEQ (n=244). This gives the new IEQ-SF much stronger statistical foundations than the original IEQ scale. It may be possible that there is some loss of power when the new IEQ-SF is used with other types of games that we did not validate against, but this seems unlikely as the underlying 3-factors of *Involvement, Real-World Dissociation* and *Challenge* are common to almost all digital games. Indeed, the new IEQ-SF removes some elements of original IEQ (e.g. the *Control* subfactor) which are more specific to action games, and focuses more on the experience of immersion which is common to almost all games.

The new IEQ-SF reduces the original 31 items of the IEQ to 11, and common concern about such short forms is that there may be some loss of power, however our four validation studies showed that this is minimal and that the new IEQ-SF can be used in place of the original IEQ.

## Use of IEQ-SF

All items of the new Short Form IEQ scale are shown in Table 8 together with their associated factors. To use the new scale all items should be presented to the participant with a 5-point Likert scale for them to give their answer. Each Likert scale is scored from 1 to 5 points where 1 corresponds to “Strongly Disagree” and 5 to “Strongly Agree” except for items 5 and 10 which are reverse scored (see Table 7). The score for each of the three factors is the mean score of the 3 or 4 items associated with that factor. The overall immersion score is the mean score of all 11 items.

Our development and validation process indicates that the IEQ-SF is a good substitute for the full IEQ in both high and low immersion situations. It is possible that reducing the number of items from 31 to 11 has resulted in a small loss of power, so if a study requires very high power, then it may be advisable to still use the original IEQ scale. However, for most other uses the new IEQ-SF can be considered a shorter and improved measure of immersion in games and can be used in place of the original long scale.

# Conclusions

In this paper we have presented the motivation, development and validation of the 11 item Short Form Immersion Experience Questionnaire (IEQ-SF). This addresses the need for a shorter version of the original 31 item Immersion Experience Questionnaire. To develop this new scale, we made use of new developments such as Item Response Theory and Open Science. Open Science data sharing practices allowed us to use pre-collected data for factor analysis (n=3539), evaluation (n=343) and validation (n=322) of the new scale. We also performed validation on new data (n=164). The resulting 11 item scale shows good validity and reliability but in contrast, to the original single factor IEQ the new IEQ-SF has 3 subfactors of Involvement, Real-world Dissociation and Challenge which make up an omni-factor of immersion.
The final questionnaire can be used in place of the original IEQ with very little loss of power. We envisage that it will be a useful measure for game experience research, particularly in studies where participants are paid by the minute or are required to fill in a battery of questionnaires. These criteria describe increasing numbers of investigations, so we see the IEQ-SF becoming a key tool in the future of game research.

# Supplementary materials

The OSF repository at <https://osf.io/a3bzk/?view_only=ce0975a91b7345f4b9ce80a7759707f5> contains all datasets used for creation and validation of the scale. It also contains all R analysis scripts, pre-registration and all source code needed to run the experiment.

# Acknowledgements

ANONYMISED FOR REVIEW

# Competing interests statement

The authors have no competing interests.

13409 words excluding abstract and references

Abeele, V. V., Spiel, K., Nacke, L., Johnson, D., & Gerling, K. (2020). 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*, *135*, 102370.

Agrawal, S., Simon, A., Bech, S., Bærentsen, K., & Forchhammer, S. (2019). Defining immersion: Literature review and implications for research on immersive audiovisual experiences. *Journal of Audio Engineering Society*, *68*(6), 404-417.

Alexandrovsky, D., Putze, S., Bonfert, M., Höffner, S., Michelmann, P., Wenig, D., Malaka, R., & Smeddinck, J. D. (2020). Examining design choices of questionnaires in VR user studies. Proceedings of the 2020 CHI conference on human factors in computing systems,

Allen, C., & Mehler, D. M. (2019). Open science challenges, benefits and tips in early career and beyond. *PLoS biology*, *17*(5), e3000246.

Aung, M. T. (2022). *Investigating the non-disruptive measurement of immersive player experience* University of York].

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531-533.

Bowey, J. T., Birk, M. V., & Mandryk, R. L. (2015). Manipulating leaderboards to induce player experience. Proceedings of the 2015 annual symposium on computer-human interaction in play,

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.

Brown, E., & Cairns, P. (2004). A grounded investigation of game immersion. CHI'04 extended abstracts on Human factors in computing systems,

Burns, C. G., & Fairclough, S. H. (2015). Use of auditory event-related potentials to measure immersion during a computer game. *International Journal of Human-Computer Studies*, *73*, 107-114.

Cairns, P. (2013). A commentary on short questionnaires for assessing usability. *Interacting with Computers*, *25*(4), 312-316.

Cairns, P. (2019). *Doing better statistics in human-computer interaction*. Cambridge University Press.

Calamia, M., Markon, K., Denburg, N. L., & Tranel, D. (2011). Developing a short form of Benton's judgment of line orientation test: an item response theory approach. *The Clinical Neuropsychologist*, *25*(4), 670-684.

Calleja, G. (2011). *In-game: From immersion to incorporation*. MIT Press.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software*, *48*, 1-29.

Cheng, M. T., She, H. C., & Annetta, L. A. (2015). Game immersion experience: its hierarchical structure and impact on game‐based science learning. *Journal of computer assisted learning*, *31*(3), 232-253.

Cockburn, A., Dragicevic, P., Besançon, L., & Gutwin, C. (2020). Threats of a replication crisis in empirical computer science. *Communications of the ACM*, *63*(8), 70-79.

Colledani, D., Anselmi, P., & Robusto, E. (2019). Development of a new abbreviated form of the Eysenck Personality Questionnaire-Revised with multidimensional item response theory. *Personality and Individual Differences*, *149*, 108-117.

Cox, A. L., Cairns, P., Berthouze, N., & Jennett, C. (2006). The use of eyetracking for measuring immersion. Workshop on What have eye movements told us so far, and what is next,

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. ERIC.

Curran, N. (2013). *The psychology of immersion and development of a quantitative measure of immersive response in games* National University of Ireland]. Cork, Ireland.

Cutting, J., & Cairns, P. (2022). Investigating game attention using the distraction recognition paradigm. *Behaviour & Information Technology*, 1-21.

Cutting, J., Cairns, P., & Kuhn, G. (2020). Nothing else matters: Video games create sustained attentional selection away from task-irrelevant features. *Attention, Perception, & Psychophysics*, *82*(8), 3907-3919.

Cutting, J., Gundry, D., & Cairns, P. (2019). Busy doing nothing? What do players do in idle games? *International Journal of Human-Computer Studies*, *122*, 133-144.

Cutting, J., & Iacovides, I. (2022). Learning by doing: Intrinsic Integration directs attention to increase learning in games.

De Mendiburu, F., & Simon, R. (2015). *Agricolae-Ten years of an open source statistical tool for experiments in breeding, agriculture and biology* (2167-9843).

Deci, E. L., & Ryan, R. M. (1985). Conceptualizations of intrinsic motivation and self-determination. In *Intrinsic motivation and self-determination in human behavior* (pp. 11-40). Springer.

Denisova, A. (2016). *Adaptive technologies in digital games: the influence of perception of adaptivity on immersion* University of York].

Denisova, A., & Cairns, P. (2015). First Person vs. Third Person Perspective in Digital Games: Do Player Preferences Affect Immersion? Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems,

Denisova, A., Nordin, A. I., & Cairns, P. (2016). The convergence of player experience questionnaires. Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play,

Dennis, S., Garrett, P., Yim, H., Hamm, J., Osth, A. F., Sreekumar, V., & Stone, B. (2019). Privacy versus open science. *Behavior research methods*, *51*, 1839-1848.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Franzen, P., Arens, A. K., Greiff, S., van der Westhuizen, L., Fischbach, A., Wollschläger, R., & Niepel, C. (2022). Developing and Validating a Short-Form Questionnaire for the Assessment of Seven Facets of Conscientiousness in Large-Scale Assessments. *Journal of Personality Assessment*, *104*(6), 759-773.

Greer, F., & Liu, J. (2016). Creating short forms and screening measures.

Haider, A., Harteveld, C., Johnson, D., Birk, M. V., Mandryk, R. L., Seif El-Nasr, M., Nacke, L. E., Gerling, K., & Vanden Abeele, V. (2022). miniPXI: Development and Validation of an Eleven-Item Measure of the Player Experience Inventory. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CHI PLAY), 1-26.

Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., & Edwards, T. (2016). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior*, *54*, 170-179.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). Elsevier.

Hudson, S., Matson-Barkat, S., Pallamin, N., & Jegou, G. (2019). With or without you? Interaction and immersion in a virtual reality experience. *Journal of business research*, *100*, 459-468.

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. In: American Association for the Advancement of Science.

Iacovides, I., Cox, A., Kennedy, R., Cairns, P., & Jennett, C. (2015). Removing the HUD: the impact of non-diegetic game elements and expertise on player involvement. Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play,

Jennett, C. (2010). *Is game immersion just another form of selective attention? An empirical investigation of real world dissociation in computer game immersion* UCL (University College London)].

Jennett, C., Cox, A. L., & Cairns, P. (2009). Investigating computer game immersion and the component real world dissociation. CHI'09 Extended Abstracts on Human Factors in Computing Systems,

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.

Johanson, C., Gutwin, C., Bowey, J. T., & Mandryk, R. L. (2019). Press pause when you play: comparing spaced practice intervals for skill development in games. Proceedings of the Annual Symposium on Computer-Human Interaction in Play,

Johnson, D., Gardner, M. J., & Perry, R. (2018). Validation of two game experience scales: the player experience of need satisfaction (PENS) and game experience questionnaire (GEQ). *International Journal of Human-Computer Studies*, *118*, 38-46.

Kim, J., Kim, K., & Kim, W. (2022). Impact of immersive virtual reality content using 360-degree videos in undergraduate education. *IEEE Transactions on Learning Technologies*, *15*(1), 137-149.

Kline, P. (2000). *A psychometric primer*. Free Association Books.

Kline, P. (2014). *An easy guide to factor analysis*. Routledge.

Koğar, H. (2020). Development of a Short Form: Methods, Examinations, and Recommendations. *Journal of Measurement and Evaluation in Education and Psychology*, *11*(3), 302-310.

Lewis, J. R. (2018). The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction*, *34*(7), 577-590.

Li, J., & Cesar, P. (2023). Social virtual reality (VR) applications and user experiences. In *Immersive video technologies* (pp. 609-648). Elsevier.

Mäyrä, F., & Ermi, L. (2005). Fundamental components of the gameplay experience: analysing immersion.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1-9.

Nacke, L., & Lindley, C. A. (2008). Flow and immersion in first-person shooters: measuring the player's gameplay experience. Proceedings of the 2008 Conference on Future Play: Research, Play, Share,

Nicolaidou, I., Pissas, P., & Boglou, D. (2023). Comparing immersive virtual reality to mobile applications in foreign language learning in higher education: A quasi-experiment. *Interactive Learning Environments*, *31*(4), 2001-2015.

Nordin, A. I., Hudson, M., Cairns, P., Rasid, N., & Baharin, H. (2018). Translating and testing the Immersive Experience Questionnaire (IEQ). *Advanced Science Letters*, *24*(3), 1545-1549.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., & Christensen, G. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422-1425.

O’Brien, H. (2016). Translating theory into methodological practice. In *Why engagement matters* (pp. 27-52). Springer.

O’Brien, H. L., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies*, *112*, 28-39.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Perrett, F. (2018). *The Challenge with Uncertainty and Immersion: Investigating the Relationship between Immersion and Uncertainty* University of York]. York, UK.

Perrig, S. A., Scharowski, N., Brühlmann, F., von Felten, N., Opwis, K., & Aeschbach, L. F. (2024). Independent Validation of the Player Experience Inventory: Findings from a Large Set of Video Game Players. Proceedings of the CHI Conference on Human Factors in Computing Systems,

Pfister, L., & Ghellal, S. (2018). Exploring the influence of non-diegetic and diegetic elements on the immersion of 2d games. Proceedings of the 30th Australian Conference on Computer-Human Interaction,

Prieto, L., Alonso, J., & Lamarca, R. (2003). Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health and quality of life outcomes*, *1*, 1-13.

Rammstedt, B., & Beierlein, C. (2014). Can’t we make it any shorter? *Journal of Individual Differences*.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate behavioral research*, *47*(5), 667-696.

Revelle, W. (2020). How to: Use the psych package for factor analysis and data reduction. *Northwestern University, Department of Psychology: Evanston, IL, USA*.

Rigby, J. M., Brumby, D. P., Gould, S. J., & Cox, A. L. (2019). Development of a Questionnaire to Measure Immersion in Video Media: The Film IEQ. Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video,

Robinson, R., Murray, J., & Isbister, K. (2018). " You're Giving Me Mixed Signals!" A Comparative Analysis of Methods that Capture Players' Emotional Response to Games. Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems,

Ruben, M. A., & Stosic, M. D. (2024). Preserving privacy in the era of openness: Commentary on open science requirements for identifiable data in psychological science journals. *American psychologist*, *79*(3), 463.

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.

Safar, S. N. (2024). *Investigating the Effectiveness of Playing Digital Games for Relieving Psychological Stress* University of York].

Sanders, T., & Cairns, P. (2010). Time perception, immersion and music in videogames. Proceedings of the 24th BCS Interaction Specialist Group Conference,

Seah, M.-l., & Cairns, P. (2008). From immersion to addiction in videogames. Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction-Volume 1,

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual review of psychology*, *69*, 487-510.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2016). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological assessment*, *12*(1), 102.

Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science. *Annual review of clinical psychology*, *15*, 579-604.

Unity Technologies. (2022). *Unity Gaming Report 2022*. <https://create.unity.com/gaming-report-2022>

Vornhagen, J. B., Tyack, A., & Mekler, E. D. (2020). Statistical significance testing at chi play: Challenges and opportunities for more transparency. Proceedings of the Annual Symposium on Computer-Human Interaction in Play,

Zhang, B., & Robb, N. (2021). Immersion Experiences in a Tablet-Based Markerless Augmented Reality Working Memory Game: Randomized Controlled Trial and User Experience Study. *JMIR Serious Games*, *9*(4), e27036.

# Appendix: Immersive Experience Questionnaire (IEQ)

The following lists the items of the IEQ. These are presented to participants as a seven-point Likert items with responses ranging from Strongly Disagree to Strongly Agree. To score, these are converted to a score from 1 to 7 (with the items indicated with an ‘r’ reverse scored) and the mean of all items used as the IEQ score (or subset of items for the relevant factor scores). Also, to clarify the relationship between the IEQ and the IEQ-SF, the table indicates to which factors each item belongs in both the original IEQ formulation and the Short Form formulation.

The factor names are abbreviated to fit in the table:
Inv = Involvement, Inv: E = Emotional Involvement, Inv: C = Cognitive Involvement, RWD = Real World Dissociation, Ch = Challenge, Ctrl = Controls

|  |  |  |  |
| --- | --- | --- | --- |
| **Item key** | **Item** | **IEQ Factor** | **SF Factor** |
| IEQ1 | The game had my full attention | Inv: C |  |
| IEQ2 | I felt focused on the game | Inv: C | Inv |
| IEQ3 | I put effort into playing the game | Inv: C |  |
| IEQ4 | I tried my best | Inv: C |  |
| IEQ5 | I lost track of time | RWD |  |
| IEQ6 (r) | I felt consciously aware of being in the real world whilst playing | RWD | RWD |
| IEQ7 (r) | I forgot about my everyday concerns | RWD | RWD |
| IEQ8 (r) | I was very much aware of myself in my surroundings | RWD |  |
| IEQ9 | I noticed events taking place around me | RWD |  |
| IEQ10 (r) | I felt the urge to stop playing and see what was happening around me | Ctrl |  |
| IEQ11 | I felt like I was interacting with the game environment | Ctrl |  |
| IEQ12 | I felt that I was separated from the real-world environment | RWD | RWD |
| IEQ13 | The game was something that I was experiencing, rather than just doing | Inv: E | Inv |
| IEQ14 | The sense of being in the game environment was stronger than the sense of being in the real world | RWD |  |
| IEQ15 | I found myself so involved that I was unaware I was using controls | Ctrl | RWD |
| IEQ16 | I moved through the game according to my own will | Ctrl |  |
| IEQ17 | I found the game challenging | Ch | Ch |
| IEQ18 (r) | There were times in the game in which I just wanted to give up | Ch |  |
| IEQ19 | I felt motivated when playing the game | Inv: C | Inv |
| IEQ20 (r) | I found the game easy | Ch | Ch |
| IEQ21 | I felt that I was making progress towards the end of the game | Inv: C |  |
| IEQ22 | I performed well in the game | Inv: C |  |
| IEQ23 | I felt emotionally attached to the game | Inv: E |  |
| IEQ24 | I was interested in seeing how the game's events would progress | Inv: E |  |
| IEQ25 | I wanted to "win" the game | Inv: C |  |
| IEQ26 | I felt in suspense about whether or not I would do well in the game | Ch | Ch |
| IEQ27 | I found myself so involved that I wanted to speak to the game directly | Inv: E |  |
| IEQ28 | I enjoyed the graphics and the imagery | Ctrl |  |
| IEQ29 | I enjoy playing the game | Inv: C | Inv |
| IEQ30 | When I stop playing, I am disappointed that the game is over | Inv: E |  |
| IEQ31 | I would like to play the game again | Inv: E |  |