Mugla Journal of Science and Technology

I-GROUP PRESENCE QUESTIONNAIRE: PSYCHOMETRICALLY REVISED ENGLISH VERSION

Mehmet İlker BERKMAN*, Dept. of Communication Design, Bahcesehir University, Turkey, ilker.berkman@comm.bau.edu.tr
(https://orcid.org/0000-0002-2340-9373)

Güven ÇATAK, Dept. of Game Design, Bahcesehir University, Turkey, guven.catak@comm.bau.edu.r
(https://orcid.org/0000-0002-4679-8973)

Received: 18.02.2021, Accepted: 26.06.2021 Research Article
*Corresponding author DOI: 10.22531/muglajsci.882271

Abstract

I-group Presence Questionnaire (IPQ), which is used to evaluate the mediated experience of presence -especially for virtual reality applications- is originally developed in German and translated to several other languages. However, there is not any psychometric study for these translations including English version, except the Portuguese and Persian translations. We evaluated English translation of IPQ with 36 participants through 12 VR sessions with an overall of 432 samples. Using a partial least squares based factor analysis approach, the original 14-item set is trimmed into 11-items in order to achieve better psychometric qualities. In addition, a covariance based confirmatory factor analysis is executed to compare models. Several indices, even the conservative Cronbach's alpha indicated that the subscales of 11-item version are reliable, but not the 14-item version. Eliminated items did not lead to a decrease in scales' sensitivity to identify different levels of Realism, Spatial Presence and Involvement for different virtual environments. Although we provided evidence to remove the items which are identically worded and inversely coded that are causing measurement error, we suggest researchers to employ the 14-items but report the results for both 14-item version and 11-item version, until the psychometric qualities of IPQ in English is confirmed with a larger sample of participants.

Keywords: virtual reality, spatial presence, I-group, PLS-CFA

I-GROUP MEVCUDİYET ÖLÇEĞİ: İNGİLİZCE VERSİYONUN PSİKOMETRİK REVİZYONU

Özet

Özellikle sanal gerçeklik uygulamaları için, teknoloji aracılığı ile tecrübe edilen "mevcudiyet" (presence) deneyimini değerlendirmek için kullanılan I-grup Varlık Anketi (IPQ), orijinal olarak Almanca olarak geliştirilmiş ve diğer birçok dile çevrilmiştir. Ancak bu çeviriler için İngilizce versiyonu da dahil olmak üzere, Portekizce ve Farsça çeviriler dışında psikometrik bir çalışma bulunmamaktadır. Bu çalışmada IPQ'nun İngilizce çevirisi, 36 katılımcı ile toplam 432 örneğin derlendiği 12 VR seansı aracılığıyla psikometrik olarak değerlendirilmiştir. PLS tabanlı faktör analizi yaklaşımı kullanılarak, daha iyi psikometrik nitelikler elde etmek için orijinal 14 maddelik set, 11 maddeye indirgenmiştir. Ayrıca, modelleri karşılaştırmak için kovaryansa dayalı doğrulayıcı faktör analizi yapılmıştır. Cornbach's alfa katsayısı başta olmak üzere çeşitli indikatörler 11 maddelik versiyonun alt ölçeklerinin güvenilir olduğunu, ancak 14 maddelik versiyonun olmadığını göstermektedir. Ölçekten çıkarılan maddeler, farklı sanal gerçeklik ortamları için farklı Gerçekçilik, Mekansal Mevcudiyet ve Katılım düzeylerini belirleyen alt ölçeklerin duyarlılığında bir azalmaya yol açmamıştır. Ölçüm hatasına neden olan benzer ifadelerle yazılmış veya ters kodlanmış maddeleri ölçek haricinde tutmak için kanıt sağlanmıştır.

Anahtar Kelimeler: sanal gerçeklik, mekansal mevcudiyet, I-group, PLS-CFA

Cite

Berkman, M.I. and Çatak G. (2021). "I-group Presence Questionnaire: Psychometrically Revised English Version", Mugla Journal of Science and Technology, 7(0), 1-10.

1. Introduction

Presence is one of the most explored mediated experiences by scholars since the early 90s, when virtual reality technologies began to evolve into consumer products.

The concept of presence has been investigated from the perspective of other different disciplines throughout a human experience oriented approach which lead to some still-ongoing debate [1]. Lombard and Jones [2] suggested a framework in order to classify these definitions, based on their similarities and differences.

Investigating their framework, the concept of presence in virtual reality can be characterized "as a technology-mediated phenomenon and focuses on subjective experience of the person that is triggered by external stimuli that leads to an inaccurate perception of technology involvement". Definitions within this scope are mainly interested in spatial presence, social presence, engagement, and realism [3].

Spatial presence refers to the "phenomenal sense of 'being there' including automatic responses to spatial cues and the mental models of mediated spaces that create the illusion of place" [4]. Engagement and concepts closely related to it; attention, involvement, flow, absorption, and (perceived) immersion; are also considered as an important component of presence since they refer to "a strong connection with the content and/or form of an experience" [2]. Realism is twofold: 1) The perceptual realism that the virtual objects are experienced as actual objects in either sensory or non-sensory ways [5], the environment is acting in a way same as the real world in which our perceptual system evolved [6]. 2) The social realism that the narrative in the virtual environment is also "plausible or true to life", "reflecting events that do or could occur in the non-mediated world" [7]. The social realism is also related to social presence, which is described as the "mutual interaction with a perceived entity refers to the degree of initial awareness, allocated attention, the capacity for both content and affective comprehension, and the capacity for both affective and behavioral interdependence with said entity" [8].

Within the context of VR, presence simply refers to the "sense of being physically present with visual, auditory, or force displays generated by a computer" [9]. Being so immersed into the simulation, our perceptual systems lead us to a point that we feel some sense of "being there" [10]. This phenomenon is an illusion that is perceptual but not cognitive [11]. The brain and body automatically reacts to the events and objects in the virtual environment which are identified by the perceptual system, while cognitive system slowly responds with a conclusion of what the person experiences is an illusion.

The efforts of defining presence are accompanied with the efforts in order to evaluate the subjects' feeling of presence. Along with the behavioral, performance-based, and physiological measures, post-test rating scales have been employed to assess presence subjectively [12].

The I-Group Presence Questionnaire (IPQ) [13] is one of the early attempts of quantitative subjective assessment of presence through a 14-item questionnaire, driven by three latent variables, namely Spatial Presence, Involvement and Realism. These dimensions almost cover all aspects of presence in VR, except social presence. Its short structure makes it a useful tool for researchers. The original scale is in German, which we call $\mathrm{IPQ_{DEU}}$ throughout this article, and it has been translated to other languages. Although the psychometric analysis had been executed for $\mathrm{IPQ_{DEU}}$ and versions in Portuguese ($\mathrm{IPQ_P}$ as named by its developers, however we call it $\mathrm{IPQ_{POR}}$ regarding to ISO-632 language codes) and Persian (which we call $\mathrm{IPQ_{FAS}}$, referring to the authentic name of the language: Farsî, as referred in ISO-639 language codes), the English version of IPQ has not been investigated for its psychometric qualities. Even some researchers claim that psychometric qualities verified for some IPQ translations imply that other translations are valid and reliable measures [14], we think that the English version of IPQ should be assessed thoroughly, since it is getting popular among the researchers for its short form compared to other multidimensional measures of presence, as VR technologies are used by the researchers from many different disciplines.

In this study, we would explore IPQ for its validity, reliability and sensitivity, in order to contribute the psychometric evaluation of the IPQ English translation.

2. Related Studies

2.1. Measuring Presence

The methods of measuring presence are classified as objective and subjective methods. Objective measures are behavioral, performance-based and physiological measures of presence [12]. Presence is considered as a subjective phenomenon and subjective measures such as focus groups, continuous subjective presence assessment techniques which employ a slider or counter, psychophysical methods such as magnitude estimation and paired comparisons which produce interval scaled data and qualitative methods such as interviews, think-aloud sessions, ethnographic techniques, experience sampling method and repertory grid analysis hold an important place in measurement of presence [15]. The most popular method was post-test rating scales in form of questionnaires, including IPQ.

Some of the well-known presence scales preceding IPQ are Slater-Usoh-Steed (SUS) [16] questionnaire, Kim and Biocca's [17] questionnaire, the Presence & Realism Questionnaire [18] and Witmer and Singer's Presence Scale [19]. Biocca et al. [20] attempted to measure presence with 18 items in three dimensions and the Independent Television Commission - Sense of Presence Inventory (ITC-SOPI) [21] measures presence with 44 items using four subscales: Sense of physical space, engagement, ecological validity, and negative effects. MEC-SPQ (Project Presence: Measurement, Effects, Conditions - Spatial Presence Questionnaire) [22] has process subscales for attention allocation, spatial situation model, spatial presence; two subscales that refer to spatial presence which are self-location and possible actions; two subscales refer to higher state actions which are higher cognitive involvement and suspension of disbelief; and subscales that address enduring user-related variables: domain specific interest, visual spatial imagery and absorption. The self-location and possible actions dimensions of MEC-SPQ, which refer to spatial presence were published as Spatial Experience Scale (SPES) [23]. Instead of verbally phrased items, Weibel et al. [24] provided a visual alternative as

Pictorial Presence Self-Assessment Manikin. Besides its three subscales of spatial presence, engagement (mental immersion) and perceptual realism which are quite similar to IPQ, the 42-item Temple Presence Inventory (TPI) [25] includes dimensions regarding to social presence, such as parasocial interaction, passive interpersonal presence, active interpersonal presence, social richness and social realism.

2.2. IPQ Studies

The first factor analytic evaluation of IPQ items had revealed the three components: Spatial Presence, Realism and Involvement [26]. The path analysis study showed that “spatial presence is mostly determined by sources of meshed patterns of actions: interaction with the VE, understanding of dynamics, and perception of dramatic meaning”. However, the authors focused on path model rather than measurement model, and they did not offer IPQ as a measurement tool in this study, but provided evidence on multidimensional nature of presence. Through a confirmatory study [13] factor structure was verified, making the IPQDEU available for research community in German language as a psychometrically validated research tool.

The items were selected among the initial 75-items; including items from several studies translated into German, along with the original items. The retained 14 items include G1 general presence item [27], INV1 querying the awareness from real environment and REAL2 item querying the similarity of experience to real environment [19], REAL1 [28], [29] and REAL3 [30] items querying the amount of realism. REAL1 and REAL3 items have the same wording, but the scale anchors are differently worded. Other items were designed by the authors. Majority of the 246 participants of the first study evaluated first-person perspective games (n=191) and used computer monitors (n=224), while just a few of them had used HMD based VR or multiple-projection based VR (n=19). Approximately, 10% percent of the participants were female and the age mean age of the whole group was 25.4 (SD=5.3). The second study involved a group of 296 participants with the mean age of 24.7 (SD=6.2) of a male majority of 88.2%. For 190 first-person view environments, there were 87 third-person views. 227 participants used monitors while 33 used HMDs.

Recent research employing IPQDEU provide evidence that it is sensitive to reveal significantly different scores based on the attributes of the VEs. It is shown that the scores are significantly different for overall general presence as well as Realism and Involvement dimension, between an abstract and a high fidelity representation of a VE, but not for Spatial Presence dimension [31].

The Persian IPQFAS was elaborated with a group of 118 male university students who evaluated their experiences on a driving simulation PC game [32]. The constructs were reformed through an exploratory factor analysis. The reported data does not include the item loading values for SP2. As given at factor analysis results which are latinised on Table 1, the G1 item is highly loaded on Spatial Presence, as observed in IPQDEU and IPQPOR studies, indicating that this item is sharing the variance with the other items of Spatial Presence. However, there is a very small difference between the loadings of the item INV4. Authors provide evidence of internal reliability and test-retest reliability for the overall scale. Their results show that IPQFAS scores are higher for the participants who played the driving game, compared to participants who watched the gameplay as a “passenger”.

Table 1 Factor loadings of items for IPQFAS latinized and translated from Panahi-Shahri et al., 2009 [32]

Involvement	Spatial Presence	Realism	Item ID	Order
		.86	REAL2	4
		.78	REAL1	2
		.76	REAL3	5
		.68	REAL4	12
	.76		SP3	6
	.76		SP4	3
.41	.72		SP5	10
.36	.66	.38	SP1	9
	.62		G1	8
.35	.51	.43	INV4	13
.82			INV3	11
.80			INV1	1
.66		.36	INV2	7

The Portuguese language adaptation called IPQPOR [33] was conducted with a sample of 478 subjects (193 females), who used a Oculus Rift DK1 HMD VR system to engage with the application “Don’t let go”, which is also one of the applications that was evaluated in our study. Testing several models on their dataset using CB-CFA (covariance based confirmatory factor analysis), authors verified that the original model proposed for IPQDEU has revealed better fit indices compared to other solutions. They reported the composite reliability for Involvement as low as 0.31, but reported Cronbach’s alpha values indicate the reliability IPQPOR.

There is also evidence that IPQPOR is sensitive to different VR content and gender, but not sensitive to exposure time of 1 to 7 min [34] and IPQ provided different results for the same content on accessed via different screening methods of HMD, projection and screen [35], but it does not reveal a significant difference between stereoscopic and monoscopic video or audio [36].

3. Methodology

3.1. Participants and Equipment

A group of 36 volunteers (14 female) aged between 19 to 26 (M=21.6, SD=1.66) were recruited for course extra credit. They engaged with 10 different virtual environments in 12 sessions, all in same order, using an Oculus Rift SDK II head mounted display and a pair of headphones. When necessary, they used a game controller, mouse, or a keyboard in order to interact with the virtual environment. None of the participants are native English speakers but they had passed an English qualification exam in order to study in a university program in English language. 4 of them were natively speaking Arabic while others speak Turkish as their mother tongue.

3.2. Stimuli and Procedure

The evaluated VEs were selected from the steampowered.com online game repository based on their user ratings and differences in their control schemes. The list of the applications and their properties are given in Table 2.

Each participant engaged with the VR environments in the given order. They were allowed to attend two VR sessions on the same day, at least with 2 hours of break. In case that they felt any negative effects such as cybersickness symptoms or extremely distressed by the content, they were allowed to quit the session. None of the participants quit any sessions before the exposure of minimum 5 minutes.

After each session, participants were asked to evaluate their experience with the VE they engaged with, using the 14 English translated items acquired from the website of the project [37] provided by the developers, with a warning that it should be "regarded as a non-tested translation by non-native speakers". Although some items are not very clear in this version, we decided to keep them as they were since some studies have already used this translation [e.g. 38-40]. The scale anchors were kept as they were in German version [13]. The scale points were set as 5 points as in the Portuguese version [33] in order to make it simpler and faster for the respondents to decide their answer. Although this may be questionable since decreasing the number of anchor points from 7 to 5 may cause some loss of reliability and precision, recent research show that "improvements in psychometric precision were identified past 6 response options" [41] for personality questionnaires and similarly, the number of response options does not differ much between 3, 5, 7 and 11 point scales for measuring user experience [42].

Table 2 Virtual environments experienced and evaluated by participants

Code	Name	Session	Developer / Publisher	Year
BD	Back to Dinasour Island	1 & 11	CryTech	2015
	VR demo that users located in a pre-historic environment observe several dinosaurs with high quality graphics and sound.			
AFM	Affected : The Manor	2 & 12	Fallen Planet Studios	2017
	VR horror experience that user navigate in the VE using a game controller, trying to find their way out of a haunted hospital with several jump scare actions.			
IE	I Expect You To Die	3	Schell Games	2017
	Puzzle game that users try to escape out of a 1930s spy car in an aeroplane, using the tools and controls in the car via a game controller.			
IM	InMind	4	Luden.io / Nival	2015
	Users experience a journey into brain with gaze-and-wait interactions in order to destroy infected brain cells.			
SE	08:46	5	846 Studios / Kraft, A.	2015
	Users have a narrative driven experience of being in an office at World Trade Center on 9/11, using a controller to move inside the building to escape.			
AFC	Affected : Carnival	6	Fallen Planet Studios	2017
	An extension of Affected series that jump scare actions occur in a less claustrophobic environment which is a carnival.			
EV	Evade	7	N/A	2015
	Fighting simulation that users cannot hit their opponents but try to run away from getting hit by moving their heads.			
PL	Pulse	8	Slick VR	2015
	Skiing downhills with leaning controls.			
DL	Don't let go	9	Skydome Studios	2016
	Scary experience that users engage with several types of embodied experiences such as knives on their virtual hands and spiders on their bodies.			
GL	G2A Land	10	G2A	2015
	Rollercoaster experience (in this study).			

3.3. Data Analysis

Since the PLS based method have "almost no limiting assumptions regarding the model specifications and data" and higher statistical power compared to covariance based methods [43], we embraced PLS (partial least squares) based confirmatory factor analysis approach for assessing the validity and reliability of IPQ English version, following the recent guidelines [44] using the software SmartPLS version 3.3.2 [45]. This analysis was executed on a dataset of 432 samples, which includes the last two VR session in which the same stimuli is evaluated repeatedly. Furthermore, we ran a covariance-based factor analysis using the Onyx software [46].

Using the first session mean score for each factor and the overall IPQ mean score, we ran a paired samples t-test in order to understand the effect of re-evaluating the same environment with IPQ. The sores of the first and the eleventh session in which the BD stimulus was explored paired on subject-basis, as well as the scores of the second and twelfth session of the AFM stimulus.

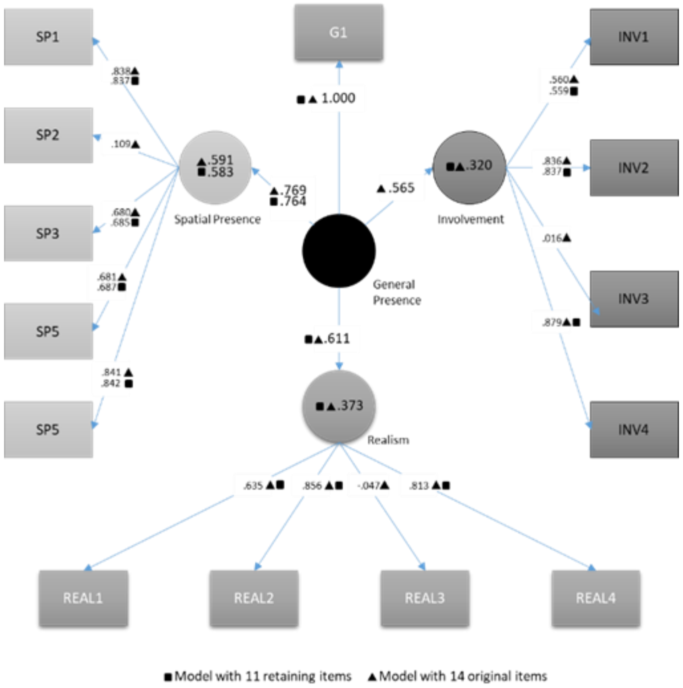

Figure 1 Measurement model

Mean comparisons and correlations are inspected for both 14-item and 11-item versions as an indicator of test-reliability. Furthermore, we think that we also explore the scale's sensitivity to changes in users' level of experience with HMD VR systems, since there are 10 sessions that correspond to 2 hours of VR engagement between the evaluated experiences.

In order to assess sensitivity, we ran a series of ANOVA. Since the aim that these analyses were for identifying the differences on factor scores due to different VEs, analyses were run on a dataset of 360 samples of the first ten sessions, excluding the repeated evaluations of AFC and BD.

We ran a PLS-CFA with the English translations of 14 items based on original German version, employed in the measurement model (Fig. 1). Based on the criterion that an item should load consonantly with a high coefficient on a singer factor, we explored the cross-loadings of each item through the EFA (exploratory factor analysis) result shown on Table 3. Although all of the items had loaded on their intended factor, some of them had very low loadings, below .4. Beginning from the lowest one, we iteratively eliminated low-loaded items until all items have a loading of at least .6, which is considered as high for PLS based analysis. Thus, removing the items REAL3, INV3 and SP03 one after each other and leaving 11-items, we achieved unidimensionality within the model, depicted in Fig. 1. Cross-loadings of final 11-items are also given on Table 3.

Table 3 Cross-loadings of items exploring the original model

		G1 Overall Presence	**INV** Involvement	**REAL** Realism	**SP** Spatial Presence
G1 In the computer generated world I had a sense of "being there" (not at all - very much)	Initial [a]	1.000	.565	.611	.769
	Final [b]	1.000	.565	.611	.764
INV1 How aware were you of the real world surrounding while navigating in the virtual world? (i.e. sounds...) (extremely aware - not aware at all)	Initial	.193	.560	.102	.255
	Final	.193	.559	.102	.241
INV2 I was not aware of my real environment. (fully disagree - fully agree)	Initial	.441	.836	.412	.483
	Final	.441	.837	.412	.483
INV3[c] I still paid attention to the real environment. (fully disagree - fully agree)	Initial	.004	.016	-.085	-.025
	Final	Removed on the second run			
INV4 I was completely captivated by the virtual world. (fully disagree - fully agree)	Initial	.564	.879	.594	.605
	Final	.564	.879	.594	.600
REAL1[c] How real did the virtual world seem to you? (completely real - not real at all)	Initial	.388	.299	.635	.319
	Final	.388	.299	.635	.298
REAL2 How much did your experience in the virtual environment seem consistent with your real world experience? (not consistent - very consistent)	Initial	.544	.514	.856	.516
	Final	.544	.515	.856	.515
REAL3 How real did the virtual world seem to you? (about as real as an imagined world - indistinguishable from the real world)	Initial	-.001	-.003	-.047	.017
	Final	Removed on the first run			
REAL4 The virtual world seemed more realistic than the real world. (fully disagree - fully agree)	Initial	.473	.441	.813	.437
	Final	.473	.442	.813	.447
SP1 Somehow I felt that the virtual world surrounded me. (fully disagree - fully agree)	Initial	.733	.576	.514	.838
	Final	.733	.576	.514	.837
SP2[c] I felt like I was just perceiving pictures. (fully disagree - fully agree)	Initial	.128	.131	.088	.109
	Final	Removed on the third run			
SP3 I did not feel present in the virtual space. (did not feel - felt present)	Initial	.467	.345	.274	.680
	Final	.467	.345	.274	.685
SP4 I had a sense of acting in the virtual space rather than operating something from outside. (fully disagree - fully agree)	Initial	.446	.479	.424	.681
	Final	.446	.479	.424	.687
SP5 I felt present in the virtual space. (fully disagree - fully agree)	Initial	.631	.461	.453	.841
	Final	.631	.462	.453	.842

Initial coefficient value shows cross-loadings of original 14-items [b]. Final coefficient value shows cross-loadings of 11 item version. [c]. Item is reverse coded.

4. Results

Table 4 Final version's inter-correlations between latent variables and √AVE (in bold).

	Gen.Pres.	Involv.	Realism	Spat.Pres.
Gen. Pres.	1.000			
Involv.	.565	.771		
Realism	.611	.552	.774	
Spat. Pres.	.764	.615	.553	.767

Indicating the convergent validity, the AVE (average variance extracted) index should exceed .5 [47]. Higher AVE indicates that latent construct is correctly represented by corresponding items. As shown on Table 4, the initial 14-item version does not match this criterion, but the 11-item final version reveal higher AVE values above .5.

As an indicator of discriminant validity, the square-root of AVE for each latent variable should be higher than its correlation with other variables. As shown in Table V, the final 11-item version also matches with this criterion, namely Fornell-Larcker criterion.

Table 5 Original version's inter-correlations between latent variables and √AVE (in bold)

	Gen.Pres.	Involv.	Realism	Spat.Pres.
Gen. Pres.	**1.000**			
Involv.	.565	**.668**		
Realism	.611	.551	**.671**	
Spat. Pres.	.769	.621	.557	**.685**

On the other hand, our data reveals that original factor structure based on 14 items does not fit to the Fornell-Larcker Criterion, as depicted on Table 5, since the √AVE for Spatial Presence is lower than the correlation between General Presence and Spatial Presence.

As a new criterion applied in variance-based models, we also checked the HTMT (heterotrait-monotrait) ratio as a more robust indicator of discriminant validity [48]. As the "exact threshold level of the HTMT is debatable", Henseler et al. [44] suggest to use HTMT.85 and HTMT.90 as upper threshold for discriminant validity.

Given in Fig. 2, the final 11-item version revealed HTMT ratio's below .85, suggesting that "construct measures represent phenomena of interest that other measures in model do not capture, i.e. discriminant validity [43]. However, HTMT ratio of Spatial Presence to Involvement does not fit HTMT.90 and Involvement to General Presence ratio does not fit HTMT.85 criteria on the initial model based on 14 items.

The reliability of the factor indicated by Cronbach's alpha values were not acceptable for the 14-item version, as given in Table 6. The final values of 11-item version vary between .66 to .76, indicating a value within an acceptable range, although Cronbachs's alpha is a conservative measure of reliability. Composite reliability scores ranging between .89 to .92 (see Table 6), which are more robust measures that the items are weighted based on the construct indicators' individual loadings, indicate a good level of reliability, without any scores above .95 that imply redundancy. Another indicator of reliability, Dillon Goldstein's rho is above .7 for all dimensions except realism, which is .69 that quite close to the threshold value.

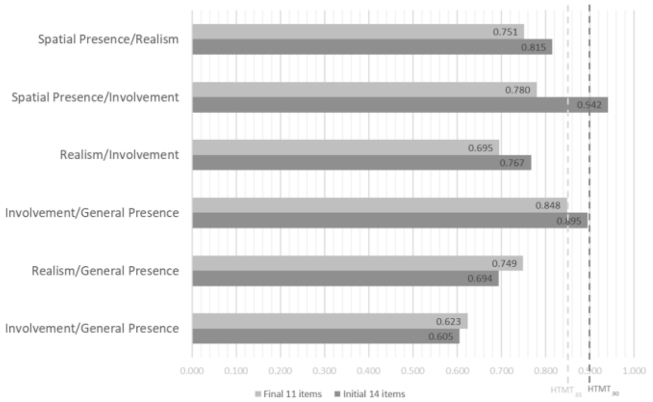

Figure 2 Heterotrait-Monotrait Ratios

When we compared the models through a covariance-based factor analysis, the AIC (Akaike information criterion) and BIC (Bayesian information criterion) revealed that 11-item model lead to better model fit on our data. However, the TLI (Tucker-Lewis index), CFI (comparative fit index), SRMR (standardized root mean square residual) and RMSEA (root mean square error of approximation) indicators do not evince a model fit, according to the common criteria [49]. Values are given in the Table 7.

Table 6 Quality criteria of reflective measurement for (I)nitial and (F)inal item sets

		AVE	Comp. Reliab.	Dillon-Goldstein's ρ	Cronbach's α
Involvement	I	.446	.703	.782	.542
	F	.595	.81	.782	.677
Realism	I	.45	.699	.689	.448
	F	.599	.816	.689	.659
Spatial Presence	I	.47	.789	.795	.658
	F	.588	.85	.804	.766
General Presence	I	1	1	1	1
	F	1	1	1	1
F(9, 350)=6.457; p=0				F(9, 350)=6.182; p=0	

The paired samples t–test reveal that Involvement score is significantly different for the first and second measurements of same stimuli, and correlating moderately at a significant level, as given in Table 8. A similar mean score difference is not observed on Realism and Spatial Presence factors between two versions, although significant correlations were detected at low to moderate levels. The total mean score for IPQ is significantly different between former and later measurement made with both 11-item and 14-item version, yet it is marginally significant with final 11-item version.

In order to assess sensitivity of scales to experiences in different virtual environments, we ran a series of one-way ANOVA's to compare mean differences, followed by post-hoc Bonferroni tests. The last wo experimental sessions were excluded from the analysis, since the stimuli used in first session is identical to session 12, and the stimuli used in the fifth session is identical to session 11. On each of the remaining 10 sessions, participants explored a different virtual environment.

Table 7 Covariance based model fit indicators

	Criterion	Initial 14 items	Final 11 items
AIC	Smaller the better	18509.699	14389.075
BIC	Smaller the better	18635.820	14490.786
TLI	≥ .95 can be 0 > TLI > 1 for acceptance	-0.24	-0.352
CFI	≥ .95 for acceptance	0.0	0.0
SRMR	≤ .08	0.298	0.364
RMSEA	< .06 to .08	0.258	0.325

We observed significant effect on all IPQ dimensions and overall score, for both versions, due to differences between the design and content of the evaluated VEs. As seen on Table 9, the Involvement scores for VE's have a similar magnitude for the highest and least scores. The five games with lowest Involvement scores are in the same order in both versions. The two VE's with the highest involvement scores are also same, there is not a difference between the scores of highest and the second one of the 11 item version means. The ranking of other VEs are different.

The Bonferroni test results reveal similar significant differences between the mean values. The scores for both IM and GL are significantly higher than EV and SE. Other mean differences are not significant. In total, each version detected two significant mean differences.

For Realism, the order of VEs based on score magnitude is same for both versions on the four VEs with lowest scores and three VEs with highest scores (see Table 10). The ranking of remaining three games are not consistent between the two versions. However, it should be noted that Realism score magnitudes of these three VEs are almost similar. As stated previously, a significant difference due to VE differences can be detected using both versions. Bonferroni post-hoc tests reveal that the Realism score for SE is significantly lower than GL, AFM and IM, for both versions. However, the 14-item version also detected a significant difference between AFC and SE. For 3 significant mean differences detected by final 11-item version, the original 14-item detected 4 differences.

Table 8 Mean comparisons of repeated measurements on the same stimuli, initial 14 items and final 11 items.

	Test		Retest		t-test		Pearson	
	M	SD	M	SD	t(71)	Sig.	r	Sig.
				Final 11 items				
Inv.	3.36	.85	3.76	.9	-3.39	.001	.32	.006
Real	3.32	.77	3.34	.82	-.183	.855	.43	.000
Sp.P.	3.71	.64	3.79	.68	-.826	.411	.29	.013
Total	3.53	.58	3.69	.64	-1.99	.05	.4	.001
				Initial 14 items				
Inv.	3.28	.76	3.63	.75	-3.63	.001	.41	.000
Real	3.13	.60	3.25	.65	-1.57	.121	.42	.000
Sp.P.	3.63	.54	3.78	.67	-1.7	.094	.28	.016
Total	3.41	.47	3.61	.55	-3.03	.003	.40	.000

The Spatial Presence dimension mean scores have an identical ranking for the six VEs with lowest scores (see Table 11). However, the ranking of the other four VEs do not match between the 11-item and 14-item versions. The original 14-item version detected that SE virtual environment Spatial Presence score is significantly lower than AFC, BD, GL, DL, IM and AFM environments.

Table 9 Mean scores for Involvement

IPQ Involvement Final			IPQ Involvement Initial		
	M	SD		M	SD
EV	2.92	0.82	EV	2.96	0.65
SE	2.98	0.85	SE	2.99	0.73
BD	3.19	0.96	BD	3.16	0.85
IE	3.25	0.85	IE	3.25	0.68
AFC	3.41	0.82	PL	3.36	0.83
PL	3.44	1.02	AFM	3.40	0.65
DL	3.51	0.83	AFC	3.42	0.68
AFM	3.53	0.71	DL	3.47	0.73
GL	3.75	0.89	GL	3.58	0.78
IM	3.75	0.90	IM	3.66	0.73
F(9, 350)=3.903; p=0			F(9, 350)=3.699; p=0		

The final 11-item version detected the same significant mean differences for SE. Bonferroni tests showed that IE means score is significantly higher than SE and AFC mean score is significantly higher than EV. For the 6 significant mean differences detected by 14-item version, 8 significant mean differences were detected by the psychometrically enhanced 11-item version.

The only difference between the rankings of VEs based the Overall IPQ score are the second and third highest scored items, as given on Table XII. Significant mean differences are observed between the SE significantly lower than BD, AFC, DL, GL, AFM and IM, as well as EV significantly lower than GL, AFM and IM. The same 8 significant mean differences were detected by the 11-item version and initial 14-item version using IPQ Overall score (see Table 12).

Table 10 Mean scores for Realism

IPQ Realism Final			IPQ Realism Initial		
	M	SD		M	
SE	2.58	.82	SE	2.64	SE
PL	2.84	.85	PL0	2.86	PL
EV	2.88	.80	EV	2.92	EV
IE	3.06	.85	IE	2.96	IE
AFC	3.17	.73	BD	3.00	AFC
BD	3.19	.85	DL	3.11	BD
DL	3.19	.90	AFC	3.16	DL
IM	3.29	.95	IM	3.18	IM
AFM	3.45	.68	AFM	3.25	AFM
GL	3.37	.84	GL	3.31	GL
F(9, 350)=3.777; p=0			F(9, 350)=3.406; p=0		

Table 11 Mean scores for Spatial Presence

IPQ Spat Pres. Final			IPQ Spat. Pres. Initial		
	M	SD		M	SD
SE	2.97	.72	SE	2.98	.76
EV	3.24	.77	EV	3.22	.79
PL	3.35	.74	PL	3.35	.74
IE	3.54	.65	IE	3.48	.65
AFC	3.61	.69	AFC	3.59	.68
BD	3.63	.68	BD	3.61	.59
GL	3.65	.67	AFM	3.66	.50
DL	3.69	.65	GL	3.67	.66
IM	3.76	.64	DL	3.67	.66
AFM	3.79	.59	IM	3.72	.68
F(9, 350)=5.32; p=0			F(9, 350)=4.575; p=0		

Since our factor analytic approach is different, the indicators of validity and model fit cannot be compared with the previous studies numerically. The initial 14-items based did not fit the model suggested for IPQ$_{DEU}$ and verified for IPQ$_{POR}$, due to the small magnitudes of loadings on several items.

Table 12 Mean scores for overall IPQ

IPQ Overall Score Initial			IPQ Overall Score Final		
	M	SD		M	SD
SE	2.87	.69	SE	2.89	.62
EV	3.06	.68	EV	3.06	.59
PL	3.26	.76	PL	3.23	.68
IE	3.34	.58	IE	3.28	.49
BD	3.41	.62	BD	3.32	.49
AFC	3.45	.61	AFC	3.43	.56
DL	3.53	.63	DL	3.48	.56
GL	3.65	.63	AFM	3.50	.44
AFM	3.66	.51	GL	3.58	.55
IM	3.67	.61	IM	3.58	.51
F(9, 350)=6.457; p=0			F(9, 350)=6.182; p=0		

5. Discussion

As we eliminated these items, the first one was the REAL3, which has an identical wording with the item REAL1 but different wording on anchors. The opposite anchors of "about as real as an imagined world" and "indistinguishable from the real world" could have been complicated to understand for the participants, as the concepts of "imagined world" and "real world" might have been confusing. Considering that they are not native English speakers, the long wording of the items might have been another issue that has caused the measurement defect.

Table 13 Reliability comparison with previous studies

		Spat. Pres.	Involv.	Realism
Cronbach's α	French	.78	.75	.54
	German IPQ$_{DEU}$	.8	.68	.64
	Portuguese IPQ$_{POR}$	.66	.53	.83
	English 14-items	.65	.54	.45
	English 11-items	.77	.68	.65
Composite Reliability	Portuguese IPQ$_{POR}$	.728	.314	.823
	English 14-items	.789	.699	.703
	English 11-items	.85	.816	.81

We also had to drop the item INV3 worded as "I still paid attention to the real environment" with the anchors "fully disagree - fully agree". As this item requires reverse coding, agreement with this item verifies that participant is focused on the outside world rather than the virtual environment. However, the anchors are worded identical to most of the other items. Similarly, the item removed on the third round of analysis, SP2 "I felt like I was just perceiving pictures" has the same "fully disagree - fully agree" anchors, which also exposed a measurement problem. Although one of the retaining items, "REAL1 - How real did the virtual world seem to you?" has anchors that needs to be reversely coded for analysis, these anchors which are worded as "completely real - not real at all" does not led to a misunderstanding. In addition, the reverse coding of this item might have led to "false answers" on the identically worded REAL3, as the participants had responded the same questionnaire not only once, but 12 times, unlike the other evaluation studies. As a result, participants might have read the questions superficially after a several times, as they were already familiar with the items. This might have also led to more frequent "false answers" on the reversed items.

The reliability indices reveal that 11-item version is the best solution in order to achieve the highest reliability but the 14-item solution also provides results that are comparable to values reported in the IPQ$_{POR}$, as given in Table 13. It should be noted that 14-item English version revealed lowest Cronbach's alpha value for Realism, compared to other values reported by Vasconcelos-Raposo et al. [33] based on their own data and the datasets available online. On the other hand, they reported a very low value of Involvement for the more liberal composite reliability indicator. The 11-item version is also highly consistent with IPQ$_{DEU}$ study in terms of the magnitudes of Cronbach's alpha values. The significant correlations between the evaluations of the same VE by the same participant imply test-retest reliability. Although our analysis through covariance-based factor analysis revealed better results of 11-item measurement model, neither of the two model are not verified according to model-fit indicators.

The Involvement scores are higher for the second engagement with the same stimuli, compared to first engagement. Although it can be assumed that users might lose their interest on their second engagement to the same content; our results are contradictory, showing that users had a higher interest on their second experience with the environment. We also observed higher scores for Realism and Spatial Presence on the second interactions but the difference is very low, which may suggest that users are more interested when they visit a familiar virtual environment, but their perceived spatial presence and judgement of realism is not affected by their prior experiences.

Inspecting the consistency between the order of the score magnitudes of VEs for both versions, we provided evidence that English version is capable of detecting the different "levels" of presence, which could have been due to graphical style, design fidelity or users' personal interest to the content domain. The inspection of these reasons are beyond the scope of this study, but we provide evidence that IPQ applied in English language provide different measurements for different VEs, as it is in IPQ_{POR} and IPQ_{DEU}.

6. Conclusion

We made minor modifications on the original items and the item structure of IPQ when adapting to English, compared to the Portuguese and Persian language adaptations. Our model evaluates the Overall Presence as a second-order latent construct that drives the G1 observable as well as first-order latent constructs of Involvement, Realism and Spatial Presence. Other adaptations employ the G1 variable as an indicator of the Spatial Presence construct.

One of the major limitations of our study is the sample size. While the PLS-CFA method can be executed with small samples, our sample is small for the number of participants, but large enough to execute any factor analytic methods. As the main purpose of the IPQ is evaluating the VE experiences rather than evaluating inter-personal differences, our methodology is concrete. On the other hand, we are aware that sense of presence is affected by personal traits and our sample of 36 participants is not adequate for representing the general population, neither for their demographic variation nor for different personal preferences and attributes. Future studies should focus on the diversity of the participants than the diversity of applications, to investigate the sensitivity of the scale to interpersonal differences.

Our choice of participants can be criticized for being non-native English speakers, for the assessment of an English language scale, but it should be remembered that many researchers employ English language scales in their studies even their subjects are not native English speakers. Future studies may recruit native English speakers to identify possible differences.

We suggest researchers to use the 14-item version of IPQ in English, but report the results for both the 11 and 14 item versions. We also ask them to share their data openly in order to provide a larger sample for further analysis, with a larger group of participants and different immersive technologies. Considering the problems in the reverse coded and identically worded items reported in our study, we also suggest researchers to carefully use the other versions of IPQ, if they are using it in a within-subjects research design of repeated measurements within the same participant group.

Supplementary Material

Data available online at doi.org/10.17632/77tdmnmnr2.

References

[1] Barfield, W. "Musings on presence twenty-five years after 'being there,'" *Presence: Teleoperators and Virtual Environments*. 25(2):148-150, 2016.

[2] Lombard, M. and Jones, M.T. (Lombard, Biocca, Freeman, IJsselsteijn, and Schaevitz,) "Defining presence" in Immersed in *Media: Telepresence Theory, Measurement and Technology,* Springer, 13–34, 2015.

[3] Berkman, M.I. and Akan, E. (Lee) "Presence and Immersion in Virtual Reality" *Encyclopedia of Computer Graphics and Games.* Springer International Publishing, Cham, 1–10, 2019.

[4] Biocca, F., Harms, C. and Burgoon, C. "Toward a More Robust Theory and Measure of Social Presence: Review and Suggested Criteria" in *Presence: Teleoperators and Virtual Environments*, 12(5):456-480, 2003.

[5] Lee, K.M., "Presence, Explicated" *Commun. Theory*, 14(1):27-50, 2004.

[6] Zahorik, P. and Jenison, R.L. "Presence as being-in-the-world" *Presence Teleoperators Virtual Environ.*, 7(1):78-89 1998.

[7] Lombard, M. and Ditton, T., "At the Heart of It All: The Concept of Presence" *J. Comput. Commun.*, 3(2), 1997.

[8] Harms, C. and Biocca, F. "Internal Consistency and Reliability of the Networked Minds Measure of Social Presence" *Seventh Annu. Int. Work. Presence 2004*, 2004.

[9] Sheridan, T.B. "Musings on Telepresence and Virtual Presence" *Presence Teleoperators Virtual Environ.*, 1(1):120–126, 1992.

[10] Biocca, F. and Delaney, B. (Lombard & Levy) "Immersive virtual reality technology" *Communication in the Age of Virtual Real*ity, 57-126, 1995.

[11] Slater, M. "Immersion and the illusion of presence in virtual reality" *British Journal of Psychology,* 109(3):431-433, 2018.

[12] Laarni, J. et al., (Lombard, Biocca, Freeman, IJsselsteijn, and Schaevitz,) "Ways to measure spatial presence: Review and future directions" in *Immersed in Media: Telepresence Theory, Measurement and Technology*, 139-185, Springer, 2015.

[13] Schubert, T., Friedmann, F. and Regenbrecht, H. "The experience of presence: Factor analytic insights" *Presence Teleoperators Virtual Environ.*, 10(3):266–281, 2001.

[14] Elsey, J. W. B., et al., "The impact of virtual reality versus 2D pornography on sexual arousal and presence" *Comput. Human Behav.*, 97:35-43, 2019.

[15] Insko, B. E., (Riva, Davide, & IJsselsteijn) "Measuring Presence: Subjective, Behavioral and Physiological Methods," S*tudies in new technologies and practices in communication.* 109-119, 2003.

[16] Slater, M., Usoh, M. and Steed, A. "Depth of Presence in Virtual Environments," *Presence Teleoperators Virtual Environ.*, 3(2):130-144, 1994.

[17] Kim, T. "Telepresence via Television: Two Dimensions of Telepresence May Have Different Connections to Memory and Persuasion.[1*]," Journal of Computer-Mediated Communication*, 3, 1997.

[18] Barfield, W., Baird, B. E. and Bjorneseth, O. J. "Presence in virtual environments as a function of type of input device and display update rate," *Displays,* 19(2):91-98, 1998.

[19] Witmer, B. G. and Singer, M. J. "Measuring presence in virtual environments: A presence questionnaire," *Presence Teleoperators Virtual Environ.*, 7(3):225-240. 1998.

[20] Biocca, F., Jin, K. and Choi, Y. "Visual touch in virtual environments: An exploratory study of presence, multimodal interfaces, and cross-modal sensory illusions," *Presence Teleoperators Virtual Environ.*, 10(3):247-265, 2001.

[21] Lessiter, J. et al. "A cross-media presence questionnaire: The ITC-sense of presence inventory," *Presence Teleoperators Virtual Environ.*, 10(3):282-297, 2001.

[22] Wirth, W. et al. "A process model of the formation of spatial presence experiences. *Media Psychol.,* 9(3):493-525, 2007.

[23] Hartmann, T. et al., "The spatial presence experience scale (SPES): A short self-report measure for diverse media settings," *J. Media Psychol*., 28(1):1-15, 2016.

[24] Weibel, D. et al., "Measuring spatial presence: Introducing and validating the pictorial presence SAM," *Presence Teleoperators Virtual Environ.*, 24(1):44-61 ,2015.

[25] Lombard, M., Ditton T. B. and Weinstein, L. "Measuring Presence: The Temple Presence Inventory," *Proceedings of Presence 2009 12th Int. Work. Presence*, 2009.

[26] Schubert, T., Friedmann, F. and Regenbrecht, H. "Embodied Presence in Virtual Environments" *Visual Representations and Interpretations*, 269-278 1999.

[27] Slater, M. and Usoh, M. "Representations Systems, Perceptual Position, and Presence in Immersive Virtual Environments" *Presence Teleoperators Virtual Environ.,* 2(3):221-233, 1993.

[28] Hendrix, C. M. "Exploratory Studies on the Sense of Presence in Virtual Environments as a Function of Visual and Auditory Display Parameters" *Master's thesis, Univ. Washingt.,* 1994.

[29] Hendrix, C.M. and Barfield, W. "Presence within virtual environments as a function of visual display parameters" *Presence Teleoperators Virtual Environ.,* 5(3), 274-289:1996.

[30] Carlin, A. S., Hoffman, H. G. and Weghorst, S. "Virtual reality and tactile augmentation in the treatment of spider phobia: A case report" *Behav. Res. Ther.*, 35(2):153-158, 1997.

[31] Schwind, V. et al., N. "Using presence questionnaires in virtual reality" in *Conference on Human Factors in Computing Systems*, 2019.

[32] Panahi-Shahri, M. et al., "Reliability and validity of Igroup Presence Questionnaire (IPQ)" *J. Behav. Sci.*, 3(1):27–34, 2009.

[33] Vasconcelos-Raposo, J. et al., "Adaptation and validation of the Igroup Presence Questionnaire (IPQ) in a Portuguese sample" *Presence Teleoperators Virtual Environ.*, 25(3):191-203, 2016.

[34] Melo, M., Vasconcelos-Raposo, J. and Bessa, M. "Presence and cybersickness in immersive content: Effects of content type, exposure time and gender" *Comput. Graph.*, 71:159-165, 2018.

[35] Rivera, H. et al., "Evaluation of Cybersickness and Sense of Presence in a VR Simulator of Electric-Powered Wheelchairs" in *2nd International Workshop on Assistive Technology (IWAT2019)*, 2019.

[36] Narciso, D. et al., "Immersive 360∘ video user experience: impact of different variables in the sense of presence and cybersickness" *Univers. Access Inf. Soc.*, 18:77–87, 2019.

[37] I-group.org-project, "igroup presence questionnaire *(IPQ) Item Download" igroup.org website. [Online]*. Available: www.igroup.org/pq/ipq/items.php. [Accessed: 15-Mar-2020].

[38] Clifford R. M. S. et al., "Aerial firefighter radio communication performance in a virtual training system: radio communication disruptions simulated in VR for Air Attack Supervision," *Vis. Comput.,* 37:63–76, 2020.

[39] Wallis, G. and Tichon, J. "Predicting the efficacy of simulator-based training using a perceptual judgment task versus questionnaire-based measures of presence," *Presence Teleoperators Virtual Environ.,* 22(1):67-85, 2013.

[40] Archambault, P. S. et al., "Driving performance in a power wheelchair simulator," *Disabil. Rehabil. Assist. Technol.,* 7(3):226-233, 2012.

[41] Simms, L. J. et al., "Does the Number of Response Options Matter? Psychometric Perspectives Using Personality Questionnaire Data" *Psychol. Assess.,* 31(4):557 2019.

[42] Lewis, J.R., "Measuring User Experience With 3, 5, 7, or 11 Points: Does It Matter?" *Hum. Factors,* Online first, 2019.

[43] Hair, J. F., Ringle, C. M. and Sarstedt, M. . "PLS-SEM: Indeed a Silver Bullet" *J. Mark. Theory Pract*., 19[2]: 139–152, 2011.

[44] Henseler, J., Hubona, G. and Ray, J. "Using PLS path modeling in new technology research: updated guidelines" *Ind. Manag. Data Syst*., 116(1):2–20, Feb. 2016.

[45] Ringle, C., Wende, S. and Becker, J.M. "SmartPLS3." *Boenningstedt: SmartPLS GmbH*, 2015.

[46] von Oertzen, T., Brandmaier, A. M. and Tsang, S. "Structural Equation Modeling With Ωnyx," *Struct. Equ. Model*., 22(1):148-161, 2015.

[47] Fornell, C. and Larcker, D. F. "Evaluating structural model with unobserved variables and measurement errors" *J. Mark. Res.*, 18(1):39-50 1981.

[48] Henseler, J., Ringle, C. M. andSarstedt, M. "A new criterion for assessing discriminant validity in variance-based structural equation modeling". *J. Acad. Mark. Sci.*, 43(1):115-135, 2014.

[49] Schreiber, J.B. et al., "Reporting structural equation modeling and confirmatory factor analysis results: A review," *Journal of Educational Research*. 99(6):323-338, 2006.