Systematic Review and Validation of the Game Experience Questionnaire (GEQ) – Implications for Citation and Reporting Practice

Effie L.-C. Law[1], **Florian Brühlmann**[2], **Elisa D. Mekler**[2]
[1]Department of Informatics, University of Leicester; [2]Faculty of Psychology, University of Basel
lcl9@leicester.ac.uk, florian.bruehlmann@unibas.ch, elisa.mekler@unibas.ch

ABSTRACT
Despite lacking a formal peer-reviewed publication, the Game Experience Questionnaire (GEQ) is widely applied in games research, which might risk the proliferation of erroneous study implications. This concern motivated us to conduct a systematic literature review of 73 publications, analysing how and why the GEQ and its variants have been employed in current research. Besides inconsistent reporting of psychometric properties, we found that misleading citation practices with regards to the source, rationale and number of items reported were prevalent, which in part seem to stem from confusion over the "manuscript in preparation" status. Additionally, we present the results of a validation study (N = 633), which found no evidence for the originally postulated 7-factor structure of the GEQ. Based on these findings, we discuss the challenges inherent to the "manuscript in preparation" status and provide recommendations for authors, researchers, educators, and reviewers on how to improve reporting, citation and publication practices.

CCS Concepts
•**Human-Centered Computing** → **HCI design and evaluation methods;** *Empirical studies in HCI;*

Author Keywords
Game Experience Questionnaire; Player Experience; Referencing.

INTRODUCTION
In the last decade games research has become increasingly prominent in HCI, and concomitantly various self-report measuring instruments have been developed to evaluate gameplay experiences. Among others, the Game Experience Questionnaire (GEQ) [54] (including its variants) has widely been applied by games researchers and practitioners to a broad scope of game genres, user groups, gaming environments, and purposes [48, 50]. These range from an individual gamer playing a console game with a joystick [18], over a co-located social game on a multi-touch tabletop for older adults [43] or massive online battle arena (MOBA) games for hardcore gamers [31], to immersive virtual learning environments for students [29].

While the GEQ appears to be a versatile tool, ironically its psychometric properties are yet to be established [6, 30, 31, 50]. Oftentimes, the rationale for employing the GEQ is simply because it has already been used in many other studies. Provocatively speaking, if the GEQ were invalid, its uncritical use might lead to erroneous conclusions and implications. A caveat we want to highlight is that the prevalence of a tool does *not* necessarily imply its validity. This concern has motivated us to look into the basic *why* and *how* questions regarding the uses of the GEQ in games research.

What is the history of the GEQ? The original version comprising 42 items across 7 factors (i.e., Challenge, Competence, Flow, Immersion, Tension, as well as Positive and Negative Affect) was documented as a deliverable of a European research project FUGA ("The Fun of Gaming") by Karolien Poels, Yvonne De Kort and Wijnand IJsselsteijn, and dated 2007 [54]. However, the deliverable was not publicly accessible until some years later; the exact timing is not known. Meanwhile, a 10-page publicly accessible online document dated 2013 [25] was published by the original authors of the GEQ, where the GEQ was described as a 33-item module, which nevertheless retained the original 7-factor structure. However, no explanation was given for the change in that document or elsewhere (NB: a query on this matter posed to one of the original GEQ authors did not yield any response). A handful of attempts were undertaken to verify the factor structure of the GEQ [6, 30, 31], which all reported inconsistent results with the collapse of the existing factors and emergence of new ones. The 7-factor structure seems not replicable by any research groups other than the originators of the GEQ.

Consequently, the confusing history of the GEQ motivated us to conduct a systematic analysis of 73 publications and a validation study with 633 participants. We focus our analytic and empirical work on the core GEQ module only, due to it being considered the most problematic of the FUGA deliverables [6, 31, 50]. Results thereof enable us to infer a clear implication that despite its popularity the GEQ needs to be applied with caution and conscientiousness, especially given its empirically unstable 7-factor structure. Indeed, some researchers justified

CHI PLAY '18, October 28–31, 2018, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-5624-4/18/10…$15.00
DOI: https://doi.org/10.1145/3242671.3242683

their decision to not use the GEQ, based on their observation that "the GEQ and its validation has not been published in its entirety except for overview articles" ([8], p.1071). Other researchers seem more lenient, arguing that the wide adoption of the GEQ without its being formally published could be an example to query the necessity as well as utility of the peer review process (e.g., [50]). Nonetheless, we share the principle upheld by the former while being deeply cautious as well as sceptical about the latter.

Our review revealed a disarray of the references cited. Examples include mis-citing a questionnaire with a similar name 'Game Engagement Questionnaire' [3]; using the 33-item GEQ while referencing the document with the 42-item version [42]; citing GEQ as a manuscript under preparation by papers as recent as 2017 [57, 58]. This resonates loudly and clearly with the criticism of Marshall and Linehan [45] on HCI's "failure of scholarship". As this could be an entrenched problem going beyond the GEQ, we discuss it from a broader perspective – not only with regards to research but also education in HCI and games research.

The main contribution of this work is threefold:

- We identify problems with the provision and application of the GEQ by systematically analysing why it has been chosen and how it has been administered. Our findings provide insights for improving the practice of publicizing and publishing a tool on the provider (i.e., originator / author) side, as well as quality-checking the tool on the consumer (i.e., user) side.
- We provide empirical evidence to corroborate the recommendation that the original factor structure and items of the GEQ be revised. While there is a strong need of a tool like the GEQ, such a tool should be robust to allow valid implications and conclusions to be drawn.
- We lend further support to the criticism of poor research practice of citing prior work [45, 46]. We identify an even more basic problem with regards to proper citations, especially manuscripts in preparation, thereby inferring implications for different roles – author, researcher, reviewer, and educators.

The rest of the paper is structured as follows. First, we describe the process and results of our systematic literature review. Second, we present the design and results of the online survey. We reflect on the insights gained in the Discussion and conclude this work with its implications for future research along this line of inquiry.

SYSTEMATIC LITERATURE REVIEW

Method

The selection of the publications was done according to an adapted QUOROM procedure, which has previously been employed to review research on user experience [2] and game enjoyment [48].

Source selection:
Instead of limiting our search to a pre-defined set of venues, we searched the Scopus database, as it covers most publication venues relevant to HCI and games research (e.g., CHI, CHI PLAY, International Journal of Human-Computer Studies).

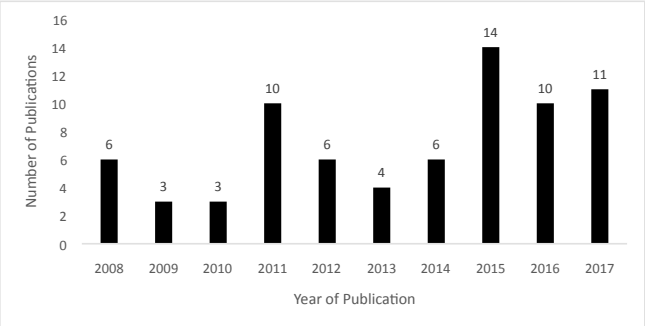

Figure 1. Frequency distribution of the publications from 2008 to 2017.

Search procedure:
The database was searched with the combined terms *game*, *experience* and *questionnaire*. Additionally, at least one of the original authors of the GEQ – IJsselsteijn, de Kort and Poels – had to be cited in the references.

Screening criteria:
For the search query we set the time frame from 2007 to 2017, as this covered over 10 years from the time when the original GEQ report was dated. Only papers that were written in English and published in journals or conference proceedings were considered.

Selection criteria for inclusion in the final analysis:
Altogether 515 papers were identified, although only 147 explicitly referred to the GEQ, with the earliest having been published in 2008. Upon closer analysis, 74 were excluded due to one of the following reasons: (i) Variants of the GEQ (In-game GEQ, Post GEQ, Social Presence in Gaming Questionnaire (SPGQ), KidsGEQ) rather than the core GEQ were employed (e.g., [20]); (ii) The GEQ dimensions were only used to inform the development of the authors' own instrument (e.g., [39]); (ii) The GEQ was claimed to be used but no results were reported (e.g., [36]); (iii) The GEQ was cited but another instrument was actually used (e.g., [64]); (iv) The paper was a duplicate (i.e., same study, same sample), though with a different title (e.g., [17]). In this case, only the earliest publication was included. Eventually, we analysed 73 papers that used the GEQ in its full or partial form. The final list of 73 papers is available as supplementary material. Figure 1 shows the distribution over the span of 10 years.

Papers were coded with regards to whether the full 7-factor questionnaire, only individual dimensions or items were used; which source was cited; the rationale for employing the GEQ; the aim of the study; what game(s) was studied; and whether any psychometric properties of the GEQ were reported (i.e., Cronbach's α, factor loadings). All coding spreadsheets are included as supplementary material.

Data analysis and results

In the following we report on the analysis of the 73 reviewed papers. Results are structured into general observations with

respect to the rationales for employing the GEQ and its application, followed by a more in-depth look at the actual items reported, the cited references, as well as the psychometric properties

Rationale for Using the GEQ

Only 31 papers (42.5%) provided a clear rationale for why the GEQ was chosen, which we categorised into four groups as follows:

- claiming that the GEQ was validated (n = 13; seven of which were published in and before 2012) – despite the lack of such a peer-reviewed formal publication [12];
- that the GEQ is popular and widely used in previous studies with different game genres (n = 10);
- the multidimensional structure of the GEQ (n = 8);
- that the GEQ was theoretically and empirically grounded with quality items (n = 6).

There were also some pragmatic reasons which were mentioned only once. For instance, the GEQ was free of charge [14] or quantitative and inclusive [10]. Note that some papers provided several rationales.

Application of the GEQ

The reviewed studies varied considerably with regards to their aims and the games they examined. The GEQ was commonly employed to evaluate new games or playful systems (e.g., [53]), but also used to assess the player experience of popular and commercially successful games (e.g., [31, 57]). Other studies included the GEQ to triangulate with physiological measures and/or in-game behavioral data (e.g., [51]).

Notably, the majority of reviewed papers (n = 47, 64%) did not report the number of items of the GEQ, or simply provided the overall number of items used in the study (e.g., [47] report 50 items, which encompass both the GEQ and the SPGQ), making it difficult to infer which version of the questionnaire had been employed. However, even if the number of GEQ items was stated, we observed inconsistent reporting. In particular, when the GEQ was reportedly administered in its full and unmodified form with the 7-factor structure, the number of reported items varied considerably (33, 34, 35, 36, 38 or 42).

We also looked into whether and how the GEQ was modified in the studies. Some included only a subset of the GEQ components or even a subset of items of a specific component, with the reported number of items ranging from 5 (e.g., only the Flow component with 5 items [32]) to 25 items (i.e., 5 subcomponents each with 5 items [63]). However, justification for the selection was seldom provided. One example is given by Johnson et al. [31], where the GEQ Competence component was dropped as it overlapped with the Player Experience of Needs Satisfaction dimension of the same name [62].

In addition, 14 papers mentioned modifying the GEQ in some way, but how and why the changes were made was rarely reported. Some exceptions are: in [34], "[...] there are still some items in the GEQ that cannot be applied for board games. Unlike most digital games, the board game Mastermind offers its players neither a storyline nor an imaginary world to immerse in" (p.253-254). Also, in [40], "[...] in 4 items, the word 'game' changed into 'application' " (p. 61). In two cases, some components/items were dropped as a result of factor analysis, in [31] "In total, seven items were dropped, and a final 6-factor solution (which explained 50.4% of the variance) was chosen as best reflecting the underlying structure." (p.2267) and in [33] where 5 items were dropped.

Concomitantly, we checked whether modifications were applied to the wording and labeling of the scale. The GEQ originally features a 5-point Likert scale with the leftmost descriptor: "0 = Not at all" and the rightmost descriptor: "4 = Extremely" [25, 54]. Out of 73 papers, 40 did not report on the scale used. For the remaining 33 only six adhered to the original format, 18 made a minor modification in numbering from 0-4 to 1-5 and six changed the descriptors in terms of 'agreement' (e.g., completely (dis)agree, strongly (dis)agree). Three peculiar ones include a 6-point scale [29], a mix of 5- and 6-point scales [38], and a pictorial (smiley) scale for elderly participants [19].

Confusing and untraceable references

The aforementioned inconsistencies with regards to item numbers motivated us to identify the sources the reviewed papers referenced.

The originators of the GEQ provided information on how to cite the published versions of 2007 and 2013 in the APA format on the front page of the respective reports [25, 54]. We label the two citations as GEQv07 and GEQv13, respectively.

- **GEQv07**: *Poels, K., de Kort, Y. A. W., & IJsselsteijn, W. A. (2007). D3.3: Game Experience Questionnaire: development of a self-report measure to assess the psychological impact of digital games. Eindhoven: Technische Universiteit Eindhoven.*
- **GEQv13**: *IJsselsteijn, W. A., de Kort, Y. A. W., & Poels, K. (2013). The Game Experience Questionnaire. Eindhoven: Technische Universiteit Eindhoven*

Note, however, that GEQv07 was only publicly accessible several years after 2007. The exact timing of release is not known; the educated guess is after 2011, based on informal personal communication between one of the authors of the present paper and one of the originators of the GEQ during CHI 2011 about the issue of access to the report, as well as based on the confusing citation made by Norman in his review of the GEQ [50], suggesting that he might not have had access to GEQv07 then.

Curiously, we identified 15 different references to the GEQ (see Table 1) in the papers reviewed, including 4 different references labeled as "Manuscript in Preparation". We take a lenient assumption that given that GEQv07 and GEQv13 were both publicly accessible in 2013, the majority of publications should have cited either of these two official references. However, GEQv07 was referenced only once by a paper published in 2015 [42], albeit with a slight variation, and none of the 73 papers cited GEQv13. Strictly speaking, only 1% of the references were adequate.

Source Label	N (%)	Description
ACE [26]	8 (11.0%)	4-page conference paper describing the challenge of measuring user experiences in digital games, focusing on the constructs Flow and Immersion. No items of the GEQ were presented. An overview article.
Citation-08 [27]	25 (34.2%)	Corresponding to the [citation] provided by Google Scholar
Engage [3]	2 (2.7%)	Brockmyer's et al. Game Engagement Questionnaire mis-cited
Fun & Games [16]	1 (1.4%)	12-page conference paper where 4 dimensions of the core GEQ were used; no items were shown.
Future Play [55]	1 (1.4%)	7-page conference paper on identifying factors of digital game experience. No items of the GEQ were given.
GEQv07 [54]	1 (1.4%)	46-page project deliverable dated as 2007 describing comprehensively the development of the GEQ, including empirical studies, lists of items, and usage guidelines.
KidsGEQ [56]	1 (1.4%)	1-page work-in-progress poster describing a kid version of the GEQ. No items provided.
M_Beh [28]	7 (9.6%)	2-page conference paper presenting an overview of the development of the GEQ. No items were presented.
No date provided	1 (1.4%)	Only the authors and the title of the GEQ were cited.
Prep-08	6 (8.2%)	The GEQ was cited as Manuscript in preparation in 2008.
Prep-13	3 (4.1%)	The GEQ was cited as Manuscript in preparation in 2013.
Prep-15	2 (2.7%)	The GEQ was cited as Manuscript in preparation in 2015.
Prep-nd	15 (20.6%)	The GEQ was cited as Manuscript in preparation without a publication year.
SPGQ [13]	2 (2.7%)	9-page conference paper documenting the development of the SPGQ.
URL	1 (1.4%)	The link to the FUGA homepage: http://www.gamexplab.nl without specifying the location of the document.

Table 1. Cited sources referenced in the studies using the GEQ. Note that numbers do not add up to 100%, because 3 publications referenced 2 sources.

Although ACE and M_Beh are two overview articles presenting no actual GEQ items, they were referenced 15 times (20%) altogether, with 13 mentions post-2013 (see Table 2). The four different "Manuscript in Preparation" references (Prep_08, Prep_13, Prep_15 and Prep_nd) amount to 26 instances (34%); all featuring the same authorship but slightly different titles. However, they differed with regards to the year of preparation, with fifteen of them not even providing any date. While it is more acceptable for the papers published pre-2013 to cite Prep_08, it is perplexing that papers published as recently as 2015, 2016 and even 2017 referred to the GEQ as still under preparation, up to a decade after it was first released. Similarly bizarre is that according to two 2017 papers [57, 58] the GEQ still had the status of under preparation in 2015. This confusion likely could have been avoided if the authors had conducted a quick search or lightweight check of the literature.

Other oddities include: (i) mis-citing another questionnaire with the same acronym "Game Engagement Questionnaire" [4, 5]; (ii) citing the KidsGEQ [44], a poster containing only seven example items, for a study with university students; (iii) citing the Future Play [41] and Fun & Games [10] papers,

Source	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	**Total**
ACE [26]	-	-	-	-	1	-	1	2	2	2	8
Citation-08 [27]	-	2	1	4	3	4	2	2	4	3	25
Engage [3]	-	-	-	-	-	-	-	2	-	-	2
Fun & Games [16]	-	-	-	-	-	-	1	-	-	-	1
Future Play [55]	-	-	-	-	-	-	-	-	-	1	1
GEQv07 [54]	-	-	-	-	-	-	-	1	-	-	1
KidsGEQ [56]	-	-	-	-	-	-	1	-	-	-	1
M_Beh [28]	-	-	-	1	-	-	-	3	2	1	7
No date provided	-	-	-	-	-	-	-	-	1	-	1
Prep-08	4	-	1	-	-	-	-	1	-	-	6
Prep-13	-	-	-	-	-	-	-	2	-	1	3
Prep-15	-	-	-	-	-	-	-	-	-	2	2
Prep-nd	2	1	-	5	3	-	-	1	2	1	15
SPGQ [13]	-	-	1	-	-	-	-	-	1	-	2
URL	-	-	-	-	-	-	1	-	-	-	1
Total	6	3	3	10	7	4	6	14	12	11	**76**

Table 2. Cited sources referenced by publication year of the reviewed papers. Note that numbers do not add up to 73, because 3 publications cited 2 references.

which do not include any GEQ items; (iv) citing the Social Presence in Gaming Questionnaire (SPGQ) of which items are different from the core GEQ [43, 60]; (v) citing the URL of the homepage of the research lab of which the GEQ originators were members, but without specifying the location of the document [63]; (vi) providing the authors and title of the GEQ without publication year or status [21]. These add up to 8 instances (or more than 10%) of the reviewed papers.

To further complicate matters, most papers did not state where and when the related document and the GEQ items were obtained. It remains unclear whether they were downloaded, for which a web address and retrieval date should have been reported, or sent by the GEQ originators or other researchers, for which a note of dated personal communication should have been included. Without this basic information, it is, for instance, difficult to tell if Prep_08 is the same as GEQv13; the former may refer to a Word document[1] of which the content is identical to the latter[2] (in pdf) except for the cover page.

It was particularly challenging to disambiguate "Citation-08", which was the most frequent reference with 25 instances (34.2%). As noted, no official document of the GEQ was published in 2008. The Word document is neither dated as 2008 nor labeled as "Manuscript in Preparation". However, it is returned as the first hit when searching Google Scholar with the search key "Game Experience Questionnaire" (as of 28th March 2018). Presumably, the two qualifiers – year and status – might be inserted by the Google Scholar search engine. The second hit with the same search key is the 2008: "[CITATION] The Game Experience Questionnaire: Development of a self-report measure to assess player experiences of digital games". No link to any document is provided. In attempting to trace the source of confusion, we identified one plausible cause – a paper published in 2008 [16] by the research group to which the GEQ originators belonged made the following reference:

IJsselsteijn, W.A., de Kort, Y.A.W. & Poels, K.: The Game Experience Questionnaire: Development of a self-report measure to assess the psychological impact of digital games (2008) (in preparation)

The phrase in the sub-heading "the psychological impact" was replaced by "player experiences" in Citation-08, which also dropped the status of "in preparation" while retaining the year of publication. For Prep_08, different combinations and word orders were used for the title. Later in GEQv13, the whole subheading was removed.

The hitherto analysis raises the question about the reliability of the references used in the papers reviewed. We estimated this reliability by collapsing the categories listed in Table 2 into two major groups: Item-Source and No-Item-Source. The former consists of Citation-08, GEQv07, and the four "Manuscript in Preparation" references, which could in principle be sources of the GEQ items, whereas the latter comprises the other 8 references, which did not explicitly present any GEQ items.

[1] http://www.academia.edu/download/44652437/Game_Experience_Questionnaire_English.doc

[2] https://pure.tue.nl/ws/files/21666907/Game_Experience_Questionnaire_English.pdf

As discussed above, the papers of 2014-2017 should have referenced GEQv07 or GEQv13, which were openly accessible in 2013, and the status of Manuscript in Preparation should have been dropped; otherwise, the references should be considered questionable. Arguing along this line, the references of those papers published in and before 2013 using one of the possible item sources are considered as acceptable. Table 3 summarises the distribution.

The acceptable references include Item-Source-Subtotal for 2008-2013 and only one instance of GEQv07 in 2014-2017 (i.e. 30 + 1), resulting in 41% reliability (i.e. 31/76), which is rather low. Nevertheless, it is reasonable to assume that most, if not all of the authors of the reviewed papers, did have access to a source of the core GEQ items for their work, irrespective of the unreliable way of citing it.

Cited Item Sources

As shown in Table 4, 8 of 15 papers which employed the full 33-item version of the GEQ (F33) referenced sources that did not list any GEQ items (4 ACE, 3 M_Beh, and 1 Future Play); with 12 of these papers having been published post-2013, suggesting that they could have referenced GEQv13. It is unclear how authors accessed the 33-item version of the GEQ, when they cited these sources. As described above, in GEQv07 the core module still had 42 items of which 6 items were optional; the reduction from 42 to 33 items (GEQv13) happened at some point and for some unknown reason. Intuitively, the 9 papers of F34-F42 should have been based on GEQv07. The variety of papers using anywhere from 34 to 42 items highlights the flexibility of the GEQ, but also suggests a lack of guidelines on item inclusion/exclusion. Furthermore, it is worrisome to note that the majority of reviewed papers (47 out of 76 instances, 62%) only reported whether they used the full GEQ (FU) or certain subcomponents (SU), but not the number of items.

Overall, except for one paper citing GEQv07 [42] and 2 papers [34, 38] presenting the complete list of items used in the appendix, there is no sure way to infer (apart from clarifying it with the authors) which version of the GEQ and which items were actually used. These ambiguities make it impossible to perform a statistical meta-analysis. This can be seen as one of the undesirable consequences of the lack of a formal publication of the GEQ.

Psychometric properties

We observed a high variation and inconsistency in the way the statistical results were reported across the reviewed studies. Besides many papers not reporting the number of items used, some only reported selected descriptive statistics (e.g., only the GEQ sub-components which showed significant correlations or differences between experimental conditions). In addition, only 17 papers (23%) reported Cronbach's α reliability scores: Flow and Competence consistently showed a relatively high value from 0.7 up to 0.94; Positive Affect and Immersion had a broad range of 0.49 to 0.85. Negative Affect, Challenge and Tension tended to have low values from 0.3 to 0.74. Overall, 12 papers reported low levels of internal consistency, with Challenge (n = 7) and Negative Affect (n = 5) being the most frequent. Even the GEQ authors themselves reported low values for Negative Affect in one publication (.53, [65]).

			Prep_						
	Citation_08	GEQv07	08	13	15	nd	Item-Source Subtotal	No-Item Source*	**Total**
2008–2013	14	-	5	-	-	11	30	3	33
2014–2017	11	1	1	3	2	5	23	20	43

Table 3. Distribution of papers citing a GEQ reference before and after GEQv07 and GEQv13 were made available sometime in 2013. Note that "Item-Source Subtotal" is the sum of the six columns on the left, which in principle present all GEQ items. *The sums are from the 8 categories of Table 2: ACE, M_Beh, Engage, Fun & Games, Future Play, KidsGEQ, SPGQ, and URL, which do not contain any GEQ items. Note that numbers do not add up to 73, because 3 publications cited 2 references.

	F33	F34-42	FU	SK	SU	**Total**
Citation-08	3	5	12	2	3	25
GEQv07	1	-	-	-	-	1
Prep_08	-	1	4	1	-	6
Prep_13	1	-	-	-	2	3
Prep_15	-	-	-	1	1	2
Prep_nd	2	-	10	-	4	16
No-Item-Source	8	3	6	1	5	23
Total	15	9	32	5	15	76

Table 4. Distribution of the references used with the full core GEQ or its subscales. Note: F33 = Full GEQ with 33 items; F34-42 = with 34-42 items; FU = Full GEQ with Unknown number of items; SK = GEQ Sub-components with a Known number of items, ranging from 5 to 25; SU = GEQ Sub-components with Unknown number of items

.

Only six papers presented factor analysis results, but to a varying degree of detail. Nonetheless, none of them could confirm the original 7-factor structure of GEQ. Johnson et al. [33], for instance, state: "*In total, five items were dropped, and a final 6-factor solution (which explained 48.4% of the variance) was chosen as best reflecting the underlying structure. In contrast to the original factor structure, positive and negative affect items were found to load on a single factor which was renamed enjoyment with negative affect items reversed*" (p. 118).

Overall, the incompleteness and often lack of statistical results is another major concern that we identified through the review.

VALIDATION OF THE GEQ

We conducted an online survey to empirically investigate the psychometric properties of the GEQ. While the original GEQ report lists 42 items [54], we chose the 33-items version, as this was the most commonly known deployed variant of the GEQ in the literature review (see Table 4). Our online survey is comparable to the original study described by Poels et al. [54], who recruited half of their 380 participants online to play their game of choice and fill in the GEQ immediately afterwards (see also pp.17 - 18 in [54]).

Procedure

Participants were recruited via Amazon Mechanical Turk. Only participants with a HIT Approval Rate greater than 95% and more than 100 approved HITs were recruited. After providing consent and basic demographic information, they were asked to briefly describe a recent experience – no longer than 24 hours ago, – they had playing a digital game. The survey was strictly limited to the last 24 hours to make sure that participants actually remembered how they felt after the experience. Additionally, they were asked to report what game they had played, what genre it belonged to, on what platform they had played the game on, how many hours ago the experience had taken place, and whether they played it with other people (either online or offline). Next, they were asked to rate their experience on the 33 GEQ items (5-point Likert scale, 0 = Not at all, 4 = Extremely). Finally, participants were asked to indicate whether they answered the questions seriously, which served as a self-reported measure of data quality. They also had the option to comment on the study and were given a completion code to receive $0.60 compensation. Note that in the original GEQ study [54], participants either received a small financial compensation or could partake in a raffle for a PlayStation 3. The survey took 9 minutes ($SD = 6.2$ minutes) to complete on average.

Participants

A total of 640 US participants completed the survey, but 7 participants were excluded because they completed the survey unrealistically fast (i.e., in less than 3 minutes). Hence, a total of 633 participants ($M_{age} = 33.78, SD = 10.57$, $range = 18-80$; 46.9% women, 2.4% non-binary or not specified) were included in the analysis. On average, participants had been playing games for 19.48 years ($SD = 8.87$, $range = 1-43$).

Participants could freely choose a game they played in the last 24 hours. Most participants had played a casual game (17%), followed by strategy games (16%), action-adventure (9%), sport (8%), first-person shooter (8%) and role playing games (8%). About a third of participants played the game on a PC or laptop (31%), another third on a mobile phone (29%), and a third on various consoles or hand-held devices.

Results

We first conducted a confirmatory factor analysis to verify the reliability and original 7-factor structure of the GEQ, followed by an exploratory factor analysis to pinpoint potential problems with the questionnaire. The R script of all statistical analyses is included as supplementary material.

Confirmatory factor analysis (CFA)

To test the multidimensional factor structure of the GEQ, we first conducted a CFA with the initial 7-factor model (i.e., Challenge, Competence, Flow, Immersion, Positive Affect, Negative Affect, Tension). All items were specified to load on their designated factor, and the loading of the first item was constrained to one. As multivariate normality was not given (Mardia tests [37]: $\chi^2_s = 14670.75, p < .001$; $Z_k =$

$57.88, p < .001$), we used a robust maximum likelihood estimation method with Huber-White standard errors and a Yuan-Bentler based scaled test statistic [61]. Results of the CFA suggest that the proposed model does not acceptably fit the data [$\chi^2_{443} = 1582.046$, $p < .001$, $\chi^2/df = 3.57$, $CFI = .879$, $SRMR = .082$, $RMSEA = .068$, $PCLOSE < .001$]. Consideration of other model fit indices (see Table 6) further indicates problems with the originally postulated factor structure of the GEQ, with only GFI and AGFI indices being within an acceptable range [49, 67].

Fit index	GEQ-33	Perfect fit criteria	Source
$\chi^2(df)$	1582.046, p <.001	Low χ^2 value and $p > .05$	[22]
χ^2/df	3.57	$\chi^2/df < 3$	[67, 69]
RMSEA (robust)	.068	RMSEA <.05	[24, 66]
SRMR (robust)	.082	$SRMR \leq 0.5$	[7]
GFI	.959	$.95 \leq GFI \leq 1$	[49, 67]
AGFI	.949	$.85 \leq AGFI \leq 1$	[67]
CFI (robust)	.879	$.95 \leq CFI \leq 1$	[24]
IFI (scaled)	.833	$.95 \leq IFI \leq 1$	[49]

Table 6. Evaluation of model fit indices

As shown in Table 7, the reliability of the individual GEQ sub-components was not satisfactory for Negative Affect (ω < .7) and barely satisfactory for Challenge. This was also reflected in the low internal consistencies measured with Cronbach's α. The average extracted variance (AVE) was lower than the mean shared variance (MSV) for Negative Affect, Challenge, Tension, Competence, and Immersion, indicating poor discriminant validity. Discriminant validity was acceptable for Flow and Positive Affect. The AVE surpassed the threshold of .50 for Flow, Competence, Tension, and Positive Affect, but not for Challenge and Negative Affect. Immersion approached the recommended threshold. Together with the results from the CFA this indicates several problems with regards to items sharing substantial variance with factors other than their proposed factor.

	ω	95% CI	Cronbach's α	AVE	MSV
Immersion	.85	[.83,.87]	.85	.48	.54
Flow	.86	[.84,.88]	.86	.54	.51
Competence	.86	[.83,.86]	.85	.54	.60
Tension	.82	[.79,.85]	.82	.55	.75
Challenge	.71	[.67,.74]	.57	.38	.42
Positive affect	.91	[.89,.92]	.91	.66	.60
Negative affect	.69	[.64,.74]	.68	.31	.75

Table 7. GEQ-33 reliability analysis. Reliability coefficient ω with bias-corrected and accelerated bootstrap (1000 iterations) 95%-confidence intervals as implemented in [35].

	Component	MR2	MR1	MR5	MR3	MR4	MR6	MR7	h2
15 I was good at it	Competence	-.030	.060	-.026	**.798**	-.006	-.111	.030	.739
02 I felt skillful	Competence	-.043	.058	.063	**.701**	-.019	.240	-.051	.451
17 I felt successful	Competence	-.073	.096	.035	**.604**	.102	.011	-.092	.697
21 I was fast at reaching the game's targets	Competence	.114	.042	.122	**.593**	.005	-.090	.067	.399
10 I felt competent	Competence	.016	.170	.105	**.491**	.076	-.042	.115	.650
19 I felt that I could explore things	Immersion	-.025	-.074	**.751**	.004	.030	-.065	.032	.697
03 I was interested in the game's story	Immersion	.002	.050	**.722**	-.037	-.045	-.049	-.017	.668
18 I felt imaginative	Immersion	-.037	-.027	**.680**	.023	.068	.080	-.107	.426
27 I found it impressive	Immersion	.017	.177	**.613**	.025	.015	.111	.005	.546
30 It felt like a rich experience	Immersion	.037	.152	**.489**	.070	.136	.088	.115	.518
12 It was aesthetically pleasing	Immersion	-.050	.269	**.337**	.094	.009	.018	.071	.521
31 I lost connection with the outside world	Flow	.069	.047	-.010	-.070	**.850**	-.077	-.023	.529
13 I forgot everything around me	Flow	-.086	-.128	.065	.110	**.724**	.090	-.042	.631
25 I lost track of time	Flow	.068	.086	.033	-.037	**.696**	-.023	.004	.598
05 I was fully occupied with the game	Flow	.018	.069	.101	.150	**.451**	.140	**.315**	.685
28 I was deeply concentrated in the game	Flow	-.008	.066	.159	.171	**.345**	.240	.282	.639
24 I felt irritable	Tension	**.806**	.013	-.012	-.014	.054	-.011	-.009	.595
22 I felt annoyed	Tension	**.800**	-.004	-.007	-.071	.004	.048	-.015	.693
29 I felt frustrated	Tension	**.656**	.089	-.126	-.131	.035	.275	-.013	.780
23 I felt pressured	Tension	**.393**	-.111	-.044	.104	.088	**.339**	.046	.597
32 I felt time pressure	Challenge	**.338**	-.040	-.007	.119	.000	.253	.041	.283
11 I thought it was hard	Challenge	.116	.005	.005	-.151	.001	**.679**	-.088	.583
26 I felt challenged	Challenge	-.032	.134	.049	.081	.059	**.661**	.085	.658
33 I had to put a lot of effort in to it	Challenge	.107	-.066	.137	.109	.034	**.563**	.094	.253
07 It gave me a bad mood	Negative Affect	**.782**	-.074	.009	.059	.012	-.078	.075	.470
09 I found it tiresome	Negative Affect	**.505**	-.240	.113	.063	-.043	.013	-.114	.692
16 I felt bored	Negative Affect	**.455**	-.213	.015	.089	-.043	-.208	-.212	.732
08 I thought about other things	Negative Affect	.278	.011	.102	.066	-.241	-.118	**-.327**	.467
06 I felt happy	Positive Affect	-.011	**.734**	.096	.097	.049	.049	-.184	.401
04 I thought it was fun	Positive Affect	-.041	**.703**	.111	- .006	-.012	.003	.273	.517
20 I enjoyed it	Positive Affect	-.073	**.653**	.095	.073	.020	.026	.184	.499
14 I felt good	Positive Affect	-.090	**.607**	.018	.213	.110	.049	-.097	.407
01 I felt content	Positive Affect	-.088	**.575**	.003	.217	.089	-.017	-.115	.596
After rotation Sums of Squares		3.17	3.70	3.03	2.90	2.61	2.07	1.44	
% of variance explained		9.7	11.2	9.2	8.8	7.9	6.3	4.4	

Table 5. Rotated pattern matrix of the EFA with 33 items loading on seven factors.

Exploratory factor analysis (EFA)

While the CFA is an overall test of the structure of a questionnaire, EFA allows for more interpretation and investigation into the reasons why the psychometric properties of the questionnaire may be insufficient. Hence, an EFA was conducted to identify weak items and theoretically meaningful factors. Bartlett's test indicated factorability ($\chi^2_{df=528} = 11586.3, p < .001$) as well as the average Kaiser-Meyer-Olkin factor adequacy measure (Overall MSA = 0.94, none below .8). Next we conducted a parallel analysis, which determines the optimal number of factors by comparing the factors extracted from the observed data with the number of factors extracted from a random data matrix [59]. Although parallel analysis suggested 6 factors, we decided to investigate a 7-factor structure – following the original 7-factor structure of the GEQ – to assess whether items need to be removed or reworded. Exploratory factor analysis using minres and oblimin (oblique) rotation extracted seven factors explaining 57% of the total variance. Oblique rotation was chosen because we expected some factors to be correlated. Factor loadings (MR1-7) and communalities (h2) of the 33 items are depicted in Table 5.

Results from the CFA and EFA suggest that the original factor structure of the GEQ is not adequate. In this case, items with communality values below .30 should first be investigated [23]. For the present study, two items of the Challenge component fell below this threshold, item 32 "I felt time pressure" and item 33 "I had to put a lot of effort into it". Next, we examined items loadings [23]. The .40-.30-.20 rule [23] states that an acceptable item should load at least .40 on its primary factor, not more than .30 on any other factor and the loading difference between the primary and secondary factor should be at least .20. Table 5 shows that items 12, 28, 23, 32 and 08 load less than .40 on their primary factor. Items 12, 05, 28, 23, 32, 08 also exhibit substantial cross-loadings.

Apart from these problematic psychometric properties, several issues emerge with regards to the interpretation of the factors. The factor MR2 was defined by high loadings of most Tension and Negative Affect items, as well as item 32 from the Challenge subscale, indicating that these components are not clearly separable within the GEQ. Based on our results, items 23 and 32 should be removed, leaving only Tension and Negative Affect items to form one factor. Positive Affect items clearly loaded onto one factor (MR1), without any issues regarding problematic items or cross-loadings of items from other constructs. Immersion as a distinct factor seemed to work reasonably well, however item 12, "It was aesthetically pleasing", should be removed. Competence also showed acceptable psychometric properties and no substantial influences of items from other components. Flow could possibly be improved by removing items 05 and 28, as they barely loaded onto the same factor (MR4) to an acceptable degree (i.e., .451 and .345 respectively). However, the remaining items "I lost connection with the outside world", "I forgot everything around me", and "I lost track of time" would arguably constitute a very narrow conceptualisation of flow, better described as "loss of sense for time". Item 32 "I felt time pressure" did not load with the other Challenge items onto one factor and should hence be removed. The remaining Challenge items may work as indicators of this component, but in light of the aforementioned concerns with regards to internal consistency, it seems risky to rely on only 3 items to measure Challenge. Finally, as Negative Affect and Tension formed one factor (MR2), the last factor MR7 does not represent a meaningful factor. It was defined by items 05 and 08. Clearly a 7-factor solution is not appropriate for the GEQ with this set of items.

DISCUSSION

Despite lacking a formal peer-reviewed validation, the GEQ has become one of the most prevalent instruments to measure different key dimensions of the player experience [50]. Indeed, our literature review counted 73 publications that employed the questionnaire, and many more publications included other modules and variants of the GEQ, or based their own questionnaires on it. This popularity arguably comes down to the GEQ items being readily available to researchers. However, in analysing various applications of the GEQ in current games research, we observed several inconsistencies and oddities with regards to reporting and citation practices surrounding the GEQ. We argue that these concerns can be to a large extent attributed to the semi-transparent process of disseminating the GEQ when it first emerged a decade ago. Further reflecting the uncritical use of the GEQ, we also observed substantial gaps with regards to reporting the instrument's psychometric properties. In fact, our own validation study confirms earlier scepticism [8] and empirical findings [6, 31] that the factor structure of the GEQ is not stable. In the following, we discuss the implications of these findings.

Cursory Literature Review

Specifically, the disarray around referencing the GEQ may imply a two-way problem. On the provider/author side, the confusion could have been mitigated if the originators of the GEQ had submitted their work for formal publication [50]. Given the comprehensiveness of the original FUGA deliverable [54], it is all the more surprising that the GEQ originators have not done so. On the consumer/researcher side, much more scrutiny should have been given to citing references, such as which variant of the GEQ was employed or whether they actually include the questionnaire items. Otherwise, this risks propagating erroneous information. Notably, Norman's [50] review of the GEQ referenced the 2007 FUGA project deliverable as published in 2009, where the GEQ consisted of 42 items. Norman stated that the number of items was 33, as revealed in "[a] copy of the questionnaire made available by the authors" ([50], p. 279). However, he did not specify which 'copy' nor when it was made available. Ironically, it appears that reviewers of that paper were also confused by what was being reviewed. Hence, this challenges Norman's [50] remark that "[...] by publicly fielding the [GEQ] prior to journal publication [the GEQ research group] call into question the practical utility of the peer review process as the gatekeeper of standardized questionnaires." Indeed, our systematic literature review attest the very significance of the peer review process on the GEQ, which could have mitigated the confusion and oddities we identified. Our findings also echo Marshall and Linehan's [45] analysis of how the work of Vandewater et al. [68] about the relationship between children's video game

use and obesity has been misinterpreted and proliferated in research on exergames.

Importance of Correct Citation

Our findings also reflect the disquieting tendency of "throw-away citation of prior work", a serious concern already addressed by Marshall et al. [46]. It can be argued that the undesirable practice of skimming related work might be caused by the pressure of citing signature papers in a specific area – an observation that many researchers could probably attest, that reviewers often comment on the omission of some "must-cite" papers, even though they may not be entirely relevant. Another plausible reason is the sheer volume of publications for a well-researched topic; some researchers unfortunately – likely under time pressure — choose to look at the abstract (and conclusion) of individual papers and then cite them. Nonetheless, there is absolutely no excuse that citing authors do not digest the paper cited, reflect on, and critique the ideas related to their own work. However, for the case of the GEQ, it is alarming to note that even the citation is not properly checked, let alone the content of the document cited.

Manuscript in Preparation

The notion of "Manuscript in preparation" is intriguing and worthy to be further investigated. How many years after a concept/tool has been conceived/developed and documented can we query its publication status as "Manuscript in preparation"?

As shown in Table 2, 11 out of the 26 papers cited the GEQ as manuscript in preparation were published after 2013, with four of them being published in 2017. Note that we do not argue that an "old" tool should not be used or manuscripts should not be cited. What we find puzzling is that both the citing authors and cited authors seem not bothered to clarify the publication status of the work when some apparent changes were made between 2007 and 2013.

Interestingly enough, the tag "Manuscript in preparation" does not appear in the title page or in fact anywhere in the publicly accessible GEQ documents released under the names of its originators. The tag could be automatically generated by a repository or search engine as a meta-data field when no formal publication venue was specified. Nonetheless, we surmise that a good number of researchers use Google Scholar, at least for an exploratory literature search on a specific topic. To give a rough impression, the search phrase "manuscript in preparation" returned about 724,000 hits (as of February 2018). Hence, there is a fair chance of encountering manuscripts in preparation when one looks up articles with this search engine. What strategies for using and referencing such a publication type should be taken to avoid the citation issues we witnessed in the GEQ? In this regard, we argue for shared responsibility of four major roles:

Researchers: It is imperative to follow an established citation practice such as the APA style, specifying where (i.e., URL) and when a manuscript was retrieved. In fact, such information is absent in many of the 73 papers we reviewed, aggravating the confusion. The status of a manuscript under preparation should be updated at most 5 years after its first release; while it may sound an arbitrary period, researchers are typically asked to establish and evaluate a three- to five-year publication plan, with concluding ongoing work being one of the tasks in the plan. Hence, in case of no update, it is advisable to contact the authors to clarify the status of the manuscript. When failing to yield a response, a very strict approach could be to not reference the manuscript and identify alternatives, if available. However, a more feasible solution is to mark the reference with a caveat. This information together with the location and date of retrieval increases transparency and can reduce the risk of mis-citation.

Authors: Authors are strongly encouraged to provide an update on their manuscript after a reasonable period of time (or longest by 5 years, as argued above) – what and why changes have (not) been made. Another scholarly practice that needs to be improved is the responsiveness to academic queries such as manuscript status. Furthermore, irrespective of the status of 'under preparation' or 'concluded/finalised', for a tool such as the GEQ that is aimed to be used by different researchers in different contexts, it is essential to provide clear and thorough guidelines. Especially when researchers tend to adapt the tool to address specific needs and constraints of their project. Indeed, our literature review showed that at least 34 of 73 papers modified the use of the GEQ involving its component, item, rating scale and/or language. The authors should advise how modifications can be implemented and what undesirable impact such modifications could have on the tool's reliability and validity.

Recently, O'Brien et al. [52] published a peer-reviewed study, which verified the factor structure of their User Engagement Questionnaire (UEQ) and shortened it from 31 to 12 items. In contrast, no report is published how the core GEQ has been changed from 42 to 33 items. While we do not comment on the UEQ per se, what we find commendable is their exemplary instructions on how to administer the long and short forms of the UEQ, and how to score as well as analyse them. Of particular relevance are their caveats on changing the wording of the items, altering the rating scale, utilising a subset of components, inserting/removing items, or translating the tool, because these could nullify the tool's established psychometric properties. With such a comprehensive guide, the chaos we observed in deploying the GEQ could have been curbed.

Reviewers: The References section may not always be thoroughly checked by reviewers, who arguably tend to look at certain items of interest instead of the complete list. We suggest that special attention be paid to references tagged with manuscript in preparation, identifying potential flaws and asking for clarification, when applicable.

Educators: The specific case of the GEQ made us ponder the implication for teaching next-generation HCI researchers, especially how to treat resources without a formal publication status. We amplify the call that researchers should critically analyse the quality of the work to be cited [45, 46] and cite it properly. Furthermore, it is crucial that statistical findings be reported systematically to ensure transparency and accountability of the work. These are all basic skills that research students must acquire as an integral part of research methods and ethics courses in their early training years. While this

recommendation sounds banal, the problematics we identified suggest that more needs to be done in actual teaching practices.

Apart from the aforementioned "personal" factors, some technical issues should be addressed. The same article can be shared in multiple platforms and repositories, which often are unfortunately not interoperable, resulting in inconsistency. A web service inviting authors to edit their list of publications has already been provided by agencies such as Google Scholar Citation[3]. Ideally, if a mechanism could be developed to tackle the issue of interoperability whereby a single update can be propagated to different sites, it would help clear up the issues of (mis-)citation.

Reliability and factor structure of the GEQ

Our own validation study combined with the reported reliability values in the literature review clearly show that at least the Challenge and Negative Affect components are highly problematic. Although internal consistency as measured by Cronbach's α is considered outdated [15], low values of Cronbach's α still point to problems with a scale [11]. We therefore recommend that authors who decide to use the GEQ should conduct a factor analysis to investigate the structure of the scale within their specific sample. Arguably, this is not a very practical solution for studies with low sample sizes, as exploratory factor analysis requires a large number of participants to be reliable [23]. Reviewers should request basic psychometric data from researchers, such as Cronbach's α, ideally including other indicators of reliability such as omega (see [15]) and results from exploratory factor analysis, if the sample size is above 200 and on a 5-to-1 participants to variable ratio [23]. If items share substantial cross-loadings, such as in the case of the Tension, Negative Affect and Challenge items, they should be removed, because item ratings might not be influenced by the underlying construct (e.g., Tension) but multiple other aspects, making the average score on Tension difficult to interpret and potentially misleading. Thus, we do not recommend using the GEQ in its current form. Some subscales appear reliable and structurally valid (e.g., Positive Affect and Immersion), however, other subscales that perform reasonably well with modifications (e.g., Flow when removing items 05 and 28) might not fully reflect the construct that was intended to be measured (i.e., low content validity).

There are several paths forward from this. Similar to the GEQ originators [55], a researcher may begin with examining common components of players' experiences with games and develop a questionnaire based on the extracted factors. However, this was beyond the scope of this paper. We therefore aimed to combine factors and identify problematic items to find a structure that fits the data. While revising the GEQ solely on this data would likely overlook some aspects of the player experience, such as a more narrow and limited conceptualisation of the Flow sub-component, it nevertheless provides a starting ground to improve and possibly extend the GEQ.

Limitations and Future Work

In the following we address some shortcomings inherent to our work and how these might be improved upon.

[3]https://scholar.google.co.uk/intl/en/scholar/citations.html

Analytical work / Literature review: There exist many information gaps in the papers reviewed, such as the number of items employed not reported in 47 out of 73 papers. We could painstakingly track these data by asking the respective authors to provide them, enabling us to fill some, but likely not all, of the gaps; it could be very challenging to yield such responses. In addition, there are omissions of basic statistical data essential for meta-analysis. It is daunting to take on such a clarification effort, which may not have a reasonable return on investment, because we need to rely heavily on the cooperativeness and generosity of the authors to look into their published work that they may regard as closed.

Empirical work / Validation: While our validation study exposed the instability of the original 7-factor structure of the GEQ and identified problematic items, we have not proceeded to conduct further empirical studies for improving on the GEQ – but refer to Johnson et al. [30] for a revised 5-factor structure. Potential drawbacks of deploying the online survey via Mechanical Turk may be relevant to our study such as the social desirability bias [1] and validity [9], which we attempted to mitigate with a careful filtering mechanism. Nevertheless, our findings are comparable – albeit not identical – to previous attempts testing the GEQ factor structure [6, 30, 31, 33].

CONCLUSION

Typically the main purpose of applying an evaluation tool is to identify strengths and weaknesses of the system under scrutiny. Clearly, if the tool is flawed, inferences drawn from the evaluation results can be erroneous, leading to a waste of effort or even drastic consequences. Hence, it is critical to ensure that such a tool is powerful in terms of its validity, reliability and sensitivity. Currently, as shown by our systematic literature review and validation study of the GEQ, the psychometric properties of this tool are yet to be fully established. While this issue of the GEQ has already been discussed by other researchers [6, 8, 30, 31, 50], what we find is an even more critical and worrying problem: *To what extent do citing authors know the work of the cited authors?* This query is related to the messiness we witnessed in referencing different versions of the GEQ. It suggests that some authors might take the questionnaire from somewhere and apply it without knowing its source, development history or properties. We highlight this as a wake-up call for reinforcing the proper citation practice as an integral part of educating the next generation of HCI and games researchers.

Probably the case of the GEQ is not unique; there may be other similar cases in other fields from which we can learn, especially how they typically cite and use manuscripts in preparation. We consider this as our future work. Another avenue for future work is the refinement of the GEQ or the development of a new tool serving the purpose of evaluating gameplay experience for a range of game genres. Certainly, this new tool must be rigorously validated, timely documented, properly tagged, and formally published.

ACKNOWLEDGMENTS

Special thanks to Lena Aeschbach and Joel Siebenmann.

REFERENCES

1. Judd Antin and Aaron Shaw. 2012. Social Desirability Bias and Self-reports of Motivation: A Study of Amazon Mechanical Turk in the US and India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 2925–2934. DOI: http://dx.doi.org/10.1145/2207676.2208699
2. Javier A. Bargas-Avila and Kasper Hornbæk. 2011. Old Wine in New Bottles or Novel Challenges: A Critical Analysis of Empirical Studies of User Experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2689–2698. DOI: http://dx.doi.org/10.1145/1978942.1979336
3. Jeanne H. Brockmyer, Christine M. Fox, Kathleen A. Curtiss, Evan McBroom, Kimberly M. Burkhart, and Jacquelyn N. Pidruzny. 2009. The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology* 45, 4 (2009), 624 – 634. DOI:http://dx.doi.org/10.1016/j.jesp.2009.02.016
4. Raffaello Brondi, Leila Alem, Giovanni Avveduto, Claudia Faita, Marcello Carrozzino, Franco Tecchia, and Massimo Bergamasco. 2015. Evaluating the Impact of Highly Immersive Technologies and Natural Interaction on Player Engagement and Flow Experience in Games. In *International Conference on Entertainment Computing*, Konstantinos Chorianopoulos, Monica Divitini, Jannicke Baalsrud Hauge, Letizia Jaccheri, and Rainer Malaka (Eds.). Springer International Publishing, Cham, 169–181. DOI:http://dx.doi.org/10.1007/978-3-319-24589-8_13
5. Raffaello Brondi, Giovanni Avveduto, Marcello Carrozzino, Franco Tecchia, Leila Alem, and Massimo Bergamasco. 2016. Immersive Technologies and Natural Interaction to Improve Serious Games Engagement. In *International Conference on Games and Learning Alliance*, Alessandro De Gloria and Remco Veltkamp (Eds.). Springer International Publishing, Cham, 121–130.
6. Florian Brühlmann and Gian-Marco Schmid. 2015. How to Measure the Game Experience?: Analysis of the Factor Structure of Two Questionnaires. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. ACM, New York, NY, USA, 1181–1186. DOI: http://dx.doi.org/10.1145/2702613.2732831
7. Barbara M Byrne. 2016. *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Routledge.
8. Paul Cairns, Anna L. Cox, Matthew Day, Hayley Martin, and Thomas Perryman. 2013. Who but not where: The effect of social play on immersion in digital games. *International Journal of Human-Computer Studies* 71, 11 (2013), 1069 – 1077. DOI: http://dx.doi.org/10.1016/j.ijhcs.2013.08.015
9. Jesse Chandler, Pam Mueller, and Gabriele Paolacci. 2014. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods* 46, 1 (01 Mar 2014), 112–130. DOI: http://dx.doi.org/10.3758/s13428-013-0365-7
10. Marcos Cordeiro d'Ornellas, Diego João Cargnin, and Ana Lúcia Cervi Prado. 2015. Evaluating the Impact of Player Experience in the Design of a Serious Game for Upper Extremity Stroke Rehabilitation.. In *MedInfo*. 363–367.
11. Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* 16, 3 (1951), 297–334.
12. Noirin Curran. 2013. Comments on the Article 'Characterising and Measuring User Experiences in Digital Games' by IJsselsteijn et al. (2007). *Interacting with Computers* 25, 4 (2013), 287–289. DOI: http://dx.doi.org/10.1093/iwc/iwt015
13. Yvonne AW De Kort, Wijnand A IJsselsteijn, and Karolien Poels. 2007. Digital games as social presence technology: Development of the Social Presence in Gaming Questionnaire (SPGQ). In *Proceedings of PRESENCE*, Vol. 195203.
14. Luis Gustavo Rotoly de Lima, André de Lima Salgado, and André Pimenta Freire. 2015. Evaluation of the User Experience and Intrinsic Motivation with Educational and Mainstream Digital Games. In *Proceedings of the Latin American Conference on Human Computer Interaction (CLIHC '15)*. ACM, New York, NY, USA, Article 11, 7 pages. DOI:http://dx.doi.org/10.1145/2824893.2824904
15. Thomas J Dunn, Thom Baguley, and Vivienne Brunsden. 2014. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology* 105, 3 (2014), 399–412.
16. Brian J. Gajadhar, Yvonne A. W. de Kort, and Wijnand A. IJsselsteijn. 2008. Shared Fun Is Doubled Fun: Player Enjoyment as a Function of Social Setting. In *Fun and Games*, Panos Markopoulos, Boris de Ruyter, Wijnand IJsselsteijn, and Duncan Rowland (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 106–117.
17. Hans-Peter Gasselseder. 2014. Those who played were listening to the music? Immersion and dynamic music in the ludonarrative. In *2014 4th International Workshop on Cognitive Information Processing (CIP)*. 1–8. DOI: http://dx.doi.org/10.1109/CIP.2014.6844512
18. Kathrin M. Gerling, Matthias Klauser, and Joerg Niesenhaus. 2011a. Measuring the Impact of Game Controllers on Player Experience in FPS Games. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (MindTrek '11)*. ACM, New York, NY, USA, 83–86. DOI: http://dx.doi.org/10.1145/2181037.2181052

19. Kathrin M. Gerling, Frank P. Schulte, and Maic Masuch. 2011b. Designing and Evaluating Digital Games for Frail Elderly Persons. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology (ACE '11)*. ACM, New York, NY, USA, Article 62, 8 pages. DOI: http://dx.doi.org/10.1145/2071423.2071501

20. Jennefer Hart, Ioanna Iacovides, Anne Adams, Manuel Oliveira, and Maria Margoudi. 2017. Understanding Engagement Within the Context of a Safety Critical Game. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '17)*. ACM, New York, NY, USA, 253–264. DOI: http://dx.doi.org/10.1145/3116595.3116633

21. Oliver Hohlfeld, Hannes Fiedler, Enric Pujol, and Dennis Guse. 2016. Insensitivity to Network Delay: Minecraft Gaming Experience of Casual Gamers. In *2016 28th International Teletraffic Congress (ITC 28)*, Vol. 03. 31–33. DOI:http://dx.doi.org/10.1109/ITC-28.2016.313

22. Daire Hooper, Joseph Coughlan, and Michael Mullen. 2008. Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods* 6, 1 (2008), 53–60.

23. Matt C. Howard. 2016. A Review of Exploratory Factor Analysis Decisions and Overview of Current Practices: What We Are Doing and How Can We Improve? *International Journal of Human-Computer Interaction* 32, 1 (2016), 51–62. DOI: http://dx.doi.org/10.1080/10447318.2015.1087664

24. Li-Tze Hu and Peter M. Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* 6, 1 (1999), 1–55. DOI:http://dx.doi.org/10.1080/10705519909540118

25. Wijnand IJsselsteijn, Yvonne De Kort, and Karolien Poels. 2013. The Game Experience Questionnaire. Eindhoven: Technische Universiteit Eindhoven. (2013). https://pure.tue.nl/ws/files/21666907/Game_Experience_Questionnaire_English.pdf

26. Wijnand IJsselsteijn, Yvonne De Kort, Karolien Poels, Audrius Jurgelionis, and Francesco Bellotti. 2007. Characterising and measuring user experiences in digital games. In *International conference on advances in computer entertainment technology*, Vol. 2. 27.

27. Wijnand IJsselsteijn, Karolien Poels, and Yvonne A W De Kort. 2008a. The Game Experience Questionnaire: Development of a self-report measure to assess player experiences of digital games. *TU Eindhoven, Eindhoven, The Netherlands* (2008). [CITATION] given in Google Scholar.

28. Wijnand IJsselsteijn, Wouter Van Den Hoogen, Christoph Klimmt, Yvonne De Kort, Craig Lindley, Klaus Mathiak, Karolien Poels, Niklas Ravaja, Marko Turpeinen, and Peter Vorderer. 2008b. Measuring the experience of digital game enjoyment. In *Proceedings of Measuring Behavior*. Noldus Information Technology Wageningen, Netherlands, 88–89.

29. Daniela Janßen, Christian Tummel, Anja Richert, and Ingrid Isenhardt. 2016. Towards Measuring User Experience, Activation and Task Performance in Immersive Virtual Learning Environments for Students. In *Immersive Learning Research Network*, Colin Allison, Leonel Morgado, Johanna Pirker, Dennis Beck, Jonathon Richter, and Christian Gütl (Eds.). Springer International Publishing, Cham, 45–58.

30. Daniel Johnson, M John Gardner, and Ryan Perry. 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 (2018), 38–46. DOI: http://dx.doi.org/10.1016/j.ijhcs.2018.05.003

31. Daniel Johnson, Lennart E. Nacke, and Peta Wyeth. 2015a. All About That Base: Differing Player Experiences in Video Game Genres and the Unique Case of MOBA Games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2265–2274. DOI: http://dx.doi.org/10.1145/2702123.2702447

32. Daniel Johnson, Peta Wyeth, Madison Clark, and Christopher Watling. 2015b. Cooperative Game Play with Avatars and Agents: Differences in Brain Activity and the Experience of Play. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 3721–3730. DOI: http://dx.doi.org/10.1145/2702123.2702468

33. Daniel Johnson, Peta Wyeth, Penny Sweetser, and John Gardner. 2012. Personality, Genre and Videogame Play Experience. In *Proceedings of the 4th International Conference on Fun and Games (FnG '12)*. ACM, New York, NY, USA, 117–120. DOI: http://dx.doi.org/10.1145/2367616.2367633

34. David O. Johnson, Raymond H. Cuijpers, Kathrin Pollmann, and Antoine A. J. van de Ven. 2016. Exploring the Entertainment Value of Playing Games with a Humanoid Robot. *International Journal of Social Robotics* 8, 2 (01 Apr 2016), 247–269. DOI: http://dx.doi.org/10.1007/s12369-015-0331-x

35. Ken Kelley. 2018. *MBESS: The MBESS R Package*. https://CRAN.R-project.org/package=MBESS R package version 4.4.3.

36. Uttam Kokil and José Luis González Sánchez. 2015. Exploring facets of playability: The differences between PC and tablet gaming. In *ACHI 2015 - 8th International Conference on Advances in Computer-Human Interactions*. 108–111.

37. Selcuk Korkmaz, Dincer Goksuluk, and Gokmen Zararsiz. 2014. MVN: An R Package for Assessing Multivariate Normality. *The R Journal* 6, 2 (2014), 151–162. https://journal.r-project.org/archive/2014-2/korkmaz-goksuluk-zararsiz.pdf

38. Nataliya Kosmyna, Franck Tarpin-Bernard, and Bertrand Rivet. 2015. Conceptual Priming for In-game BCI Training. *ACM Trans. Comput.-Hum. Interact.* 22, 5, Article 26 (Oct. 2015), 25 pages. DOI: http://dx.doi.org/10.1145/2808228

39. Effie Lai-Chong Law and Xu Sun. 2012. Evaluating user experience of adaptive digital educational games with Activity Theory. *International Journal of Human-Computer Studies* 70, 7 (2012), 478 – 497. DOI: http://dx.doi.org/10.1016/j.ijhcs.2012.01.007 Special Issue on User Experience (UX) in Virtual Learning Environments.

40. Gun A. Lee, Andreas Dünser, Seungwon Kim, and Mark Billinghurst. 2012. CityViewAR: A mobile outdoor AR application for city visualization. In *2012 IEEE International Symposium on Mixed and Augmented Reality - Arts, Media, and Humanities (ISMAR-AMH)*. 57–64. DOI: http://dx.doi.org/10.1109/ISMAR-AMH.2012.6483989

41. Yingzi Lin, Jeffrey Breugelmans, Maura Iversen, and David Schmidt. 2017. An Adaptive Interface Design (AID) for enhanced computer accessibility and rehabilitation. *International Journal of Human-Computer Studies* 98 (2017), 14 – 23. DOI: http://dx.doi.org/10.1016/j.ijhcs.2016.09.012

42. Tapani N Liukkonen, Tuomas Mäkilä, Hanna Ahtosalo, Toni Heinonen, Reetta Raitoharju, and Paula Pitkäkangas. 2015. Perceptions of the Elderly Users of Motion Tracking Exergames. In *Proceedings of the IADIS International Conference Game and Entertainment Technologies*. 52–64.

43. Abdullah Al Mahmud, Omar Mubin, Suleman Shahid, and Jean-Bernard Martens. 2010. Designing social games for children and older adults: Two related case studies. *Entertainment Computing* 1, 3 (2010), 147 – 156. DOI: http://dx.doi.org/10.1016/j.entcom.2010.09.001

44. Milena S. Markova. 2013. How Does the Tangible Object Affect Motor Skill Learning?. In *Proceedings of the 8th International Conference on Tangible, Embedded and Embodied Interaction (TEI '14)*. ACM, New York, NY, USA, 305–308. DOI: http://dx.doi.org/10.1145/2540930.2558150

45. Joe Marshall and Conor Linehan. 2017. Misrepresentation of Health Research in Exertion Games Literature. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 4899–4910. DOI: http://dx.doi.org/10.1145/3025453.3025691

46. Joe Marshall, Conor Linehan, Jocelyn Spence, and Stefan Rennick Egglestone. 2017. Throwaway Citation of Prior Work Creates Risk of Bad HCI Research. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. ACM, New York, NY, USA, 827–836. DOI: http://dx.doi.org/10.1145/3027063.3052751

47. Bernhard Maurer, Ilhan Aslan, Martin Wuchse, Katja Neureiter, and Manfred Tscheligi. 2015. Gaze-Based Onlooker Integration: Exploring the In-Between of Active Player and Passive Spectator in Co-Located Gaming. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '15)*. ACM, New York, NY, USA, 163–173. DOI: http://dx.doi.org/10.1145/2793107.2793126

48. Elisa D. Mekler, Julia Ayumi Bopp, Alexandre N. Tuch, and Klaus Opwis. 2014. A Systematic Review of Quantitative Studies on the Enjoyment of Digital Entertainment Games. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 927–936. DOI:http://dx.doi.org/10.1145/2556288.2557078

49. Jeremy Miles and Mark Shevlin. 2007. A time and a place for incremental fit indices. *Personality and Individual Differences* 42, 5 (2007), 869 – 874. DOI: http://dx.doi.org/10.1016/j.paid.2006.09.022 Special issue on Structural Equation Modeling.

50. Kent L. Norman. 2013. GEQ (Game Engagement/Experience Questionnaire): A Review of Two Papers. *Interacting with Computers* 25, 4 (2013), 278–283. DOI:http://dx.doi.org/10.1093/iwc/iwt009

51. Patricia Elena Nunez Castellar, Kimmo Oksanen, and Jan Van Looy. 2014. Assessing game experience: Heart rate variability, in-game behavior and self-report measures. In *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*. 292–296. DOI: http://dx.doi.org/10.1109/QoMEX.2014.6982334

52. Heather L. O'Brien, Paul Cairns, and Mark Hall. 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 (2018), 28 – 39. DOI: http://dx.doi.org/10.1016/j.ijhcs.2018.01.004

53. Rakesh Patibanda, Florian 'Floyd' Mueller, Matevz Leskovsek, and Jonathan Duckworth. 2017. Life Tree: Understanding the Design of Breathing Exercise Games. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '17)*. ACM, New York, NY, USA, 19–31. DOI: http://dx.doi.org/10.1145/3116595.3116621

54. Karolien Poels, Yvonne De Kort, and Wijnand IJsselsteijn. 2007a. D3.3 : Game Experience Questionnaire: development of a self-report measure to assess the psychological impact of digital games. Eindhoven: Technische Universiteit Eindhoven. (2007).

55. Karolien Poels, Yvonne de Kort, and Wijnand Ijsselsteijn. 2007b. "It is Always a Lot of Fun!": Exploring Dimensions of Digital Game Experience Using Focus Group Methodology. In *Proceedings of the 2007 Conference on Future Play (Future Play '07)*. ACM, New York, NY, USA, 83–89. DOI: http://dx.doi.org/10.1145/1328202.1328218

56. Karolien Poels, Wijnand Ijsselsteijn, and Yvonne de Kort. 2008. Development of the kids game experience questionnaire. In *Proceedings of Meaningful Play 2008*.

57. Aung Pyae, Mika Luimula, and Jouni Smed. 2017a. Investigating Players' Engagement, Immersion, and Experiences in Playing PokéMon Go. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition (C&C '17)*. ACM, New York, NY, USA, 247–251. DOI: http://dx.doi.org/10.1145/3059454.3078859

58. Aung Pyae, Luimula Mika, and Jouni Smed. 2017b. Understanding Players' Experiences in Location-based Augmented Reality Mobile Games: A Case of PokéMon Go. In *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '17 Extended Abstracts)*. ACM, New York, NY, USA, 535–541. DOI: http://dx.doi.org/10.1145/3130859.3131322

59. William Revelle. 2018. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. https://CRAN.R-project.org/package=psych R package version 1.8.4.

60. Wannes Ribbens, Steven Malliet, Richard Van Eck, and Damien Larkin. 2016. Perceived realism in shooting games: Towards scale validation. *Computers in Human Behavior* 64 (2016), 308 – 318. DOI: http://dx.doi.org/10.1016/j.chb.2016.06.055

61. Yves Rosseel. 2012. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* 48, 2 (2012), 1–36. http://www.jstatsoft.org/v48/i02/

62. Richard M. Ryan, C. Scott Rigby, and Andrew Przybylski. 2006. The Motivational Pull of Video Games: A Self-Determination Theory Approach. *Motivation and Emotion* 30, 4 (01 Dec 2006), 344–360. DOI: http://dx.doi.org/10.1007/s11031-006-9051-8

63. Dylan Schouten, Isabel Pfab, Anita Cremers, Betsy van Dijk, and Mark Neerincx. 2014. Gamification for Low-Literates: Findings on Motivation, User Experience, and Study Design. In *Computers Helping People with Special Needs*, Klaus Miesenberger, Deborah Fels, Dominique Archambault, Petr Peňáz, and Wolfgang Zagler (Eds.). Springer International Publishing, Cham, 494–501.

64. Suleman Shahid, Emiel Krahmer, Marc Swerts, and Omar Mubin. 2010. Child-robot Interaction During Collaborative Game Play: Effects of Age and Gender on Emotion and Experience. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction (OZCHI '10)*. ACM, New York, NY, USA, 332–335. DOI: http://dx.doi.org/10.1145/1952222.1952294

65. Janny C. Stapel, Yvonne A. W. de Kort, and Wijnand A. IJsselsteijn. 2008. Sharing Places: Testing Psychological Effects of Location Cueing Frequency and Explicit vs. Inferred Closeness. In *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '08)*. ACM, New York, NY, USA, 399–402. DOI:http://dx.doi.org/10.1145/1409240.1409298

66. James H. Steiger. 2007. Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences* 42, 5 (2007), 893 – 898.

67. Barbara G Tabachnick and Linda S Fidell. 2007. *Using multivariate statistics*. Allyn & Bacon/Pearson Education.

68. Elizabeth A Vandewater, Mi suk Shim, and Allison G Caplovitz. 2004. Linking obesity and activity level with children's television and video game use. *Journal of Adolescence* 27, 1 (2004), 71 – 85. DOI: http://dx.doi.org/https://doi.org/10.1016/j.adolescence.2003.10.003 Video Games and Public Health.

69. Blair Wheaton, Bengt Muthén, Duane F. Alwin, and Gene F. Summers. 1977. Assessing Reliability and Stability in Panel Models. *Sociological Methodology* 8 (1977), 84–136. DOI:http://dx.doi.org/10.2307/270754