

Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences



Vero Vanden Abeele^{a,*}, Katta Spiel^{a,b}, Lennart Nacke^c, Daniel Johnson^d, Kathrin Gerling^a

^a KU Leuven, Andreas Vesaliusstraat 13, 3000 Leuven, Belgium

^b Human-Computer Interaction Group, TU Wien, Austria

^c HCI Games Group, University of Waterloo, Canada

^d Queensland University of Technology, Brisbane, Australia

ARTICLE INFO

Keywords:

Player experience
PX
Game experience
Games user research
GUR
Scale development
Scale validation
Measurement instrument
Means-End theory

ABSTRACT

Games User Research (GUR) focuses on measuring, analysing and understanding player experiences to optimise game designs. Hence, GUR experts aim to understand how specific game design choices are experienced by players, and how these lead to specific emotional responses. An instrument, providing such actionable insight into player experience, specifically designed by and for GUR was thus far lacking. To address this gap, the *Player Experience Inventory* (PXI) was developed, drawing on Means-End theory and measuring player experience both at the level of *Functional Consequences*, (i.e., the immediate experiences as a direct result of game design choices, such as audiovisual appeal or ease-of-control) and at the level of *Psychosocial Consequences*, (i.e., the second-order emotional experiences, such as immersion or mastery). Initial construct and item development was conducted in two iterations with 64 GUR experts. Next, the scale was validated and evaluated over five studies and populations, totalling 529 participants. Results support the theorized structure of the scale and provide evidence for both discriminant and convergent validity. Results also show that the scale performs well over different sample sizes and studies, supporting configural invariance. Hence, the PXI provides a reliable and theoretically sound tool for researchers to measure player experience and investigate how game design choices are linked to emotional responses.

1. Introduction

The objective of much games research is ontological or epistemological in nature (Carter et al., 2014): defining what is a game, dissecting the core constituents of the player experience or modeling cultural interactions between games and society at large. However, equally much games research is characterized by more applied aspirations. Games User Research (GUR) is a relatively recent but blooming field, blending research from Human-Computer Interaction (HCI) and Game Development (Drachen et al., 2018; Nacke, 2017). GUR focuses on observing and understanding player experience, conceptualized as “the individual, personal experience held by the player during and immediately after the playing of the game (Wiemeyer et al., 2016)”, with the aim of designing games that meet players’ expectations and delivering actionable insights that can drive a game company’s activities.

To analyze and evaluate player experiences, GUR experts have many tools in their toolbox, e.g., play-testing protocols (Medlock et al., 2002; Pagulayan et al., 2003), playability heuristics (Desurvire and

Wiberg, 2009; Korhonen and Koivisto, 2006), game analytics (El-Nasr et al., 2016), biometrics (Mirza-Babaei et al., 2012; Nacke, 2013), etc. In the past years, we particularly witnessed a surge in the use of advanced data mining techniques to scrutinize player data (Baukhage et al., 2015; Drachen and Canossa, 2009). Yet, despite the advances in connected technologies and algorithms, players’ self-reports remain important to give meaning to these metrics. How should an increase in kill-death-ratios be interpreted? Does an increase in sweaty palms also reflect an increase in immersion (Nielsen, 2016)? There is still an identifiable need to triangulate objective, behavioral player data with subjective, introspective player self-reports. One way of collecting such self-reports is through questionnaires. Hence, several instruments exist that inquire into players’ subjective experiences (e.g., Brockmyer et al., 2009; Cheng et al., 2015; IJsselsteijn et al., 2007; Jennett et al., 2008; Ryan et al., 2006). However, some of these scales still await scientific validation (Johnson et al., 2018; Law et al., 2018). Moreover, these instruments are mostly focused on measuring psychosocial experiences. Therefore, to derive actionable insights on what game

* Corresponding author.

E-mail address: vero.vandenabeele@kuleuven.be (V.V. Abeele).

<https://doi.org/10.1016/j.ijhcs.2019.102370>

Received 31 January 2019; Received in revised form 29 June 2019; Accepted 18 October 2019

Available online 19 October 2019

1071-5819/© 2019 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

design elements may be contributing to these higher-level experiences, a GUR researcher needs to use complementary tools that help translate psychosocial experiences into operational guidelines.

In this article, we present the development and validation of an instrument to measure player experiences, with a specific focus on delivering actionable insight for GUR professionals. The Player Experience Inventory (PXI) allows GUR researchers to understand how lower-level game design choices (e.g., visual embellishments, choice of GUI elements, or the impact of making different game items available to players) are perceived by players, and how these contribute to higher-order psychological experiences (e.g., immersion, mastery, autonomy). To do so, the instrument measures player experience at both the level of *Functional Consequences*, i.e., the immediate, tangible consequences, experienced as a direct result of game design choices, and the level of *Psychosocial Consequences*, i.e., the emotional experiences, as a second-order response to game design choices. While other methods are available to GUR researchers to scrutinize the impact of game design choices, e.g., game heuristics (Desurvire and Wiberg, 2009; Korhonen and Koivisto, 2006), player experience cards (Lucero and Arrasvuori, 2010) or play-testing protocols (Medlock et al., 2002; Pagulayan et al., 2003), these methods require additional time and budgets. As GUR is characterized by its direct relevance for game industry, pressing release cycles more often than not constrain available time and budget. Moreover, when combining different methods, adequate expertise is needed to avoid piecemeal research results that lower overall scientific quality. The advantage of the PXI is that measurement is conducted with one scale, measuring at the same time at both levels.

In the past years, in GUR research, there has been increased attention towards psychometric performance and the lack of validated instruments. The PXI has been rigorously developed and validated over seven studies with 529 participants, and with strong involvement from game industry and academia, involving 64 GUR experts. Results of the studies support the theorized structure of the scale and provide evidence for both discriminant and convergent validity. Additionally, the scale has shown configural invariance across studies. Finally, criterion validity has been verified, as well as the underlying causal model. Hence, the results support the PXI as a reliable and theoretically sound tool for GUR experts.

2. Related work

When investigating currently available questionnaires, a GUR researcher may find it hard to make a choice, and come to the conclusion that *player experience* is an elusive concept to grasp. Many questionnaires prevail. This is a consequence of the many theories, models and concepts that have been put forward to define what constitutes a 'good' player experience. Different theories come with different questions to ask players. As of today, there is not one single accepted or integrating framework. However, there is agreement that player experience is a complex, multi-dimensional construct (IJsselsteijn et al., 2007), perhaps even a multi-paradigmatic construct. Different disciplines apply different lenses to study player experiences. Here we discuss the most relevant lenses and associated questionnaires.

2.1. Player experience as need fulfillment or psychological state attainment

A first lens conceptualizes humans as self-regulating beings. Hence, these theories and models consider playing games as a means to self-regulate, i.e. to regulate moods (Zillmann, 1988, 2015) or satisfy needs (Deci and Ryan, 1985; Ryan et al., 2006; Sherry et al., 2006). Players are motivated to play certain games to fulfill these needs, and a 'good' player experience results from the *satisfaction of these needs*. According to Self-Determination Theory (SDT), all humans have a universal need for growth through *Competence, Relatedness* and *Autonomy* (Deci and Ryan, 1985). Games are primarily motivating to the extent that players experience autonomy, competence and relatedness while playing. Hence,

the Player Experience of Need satisfaction questionnaire (PENS) (Ryan et al., 2006) as well as the Ubisoft Perceived Experience Questionnaire (UPEQ) (Azadvar and Canossa, 2018), set forward to measure to what extent a player finds these universal needs satisfied.

A satisfying player experience may also be conceptualized as the result of more broadly contextualized *gratifications sought and obtained* by media audiences. Stemming from this conceptualization, players actively choose to play games to satisfy certain gratifications. These gratifications, while not necessarily universal or leading to personal growth, are still shared by the audience, who seek their attainment. In a series of studies, Sherry and colleagues investigated what gratifications lie at the heart of gameplay (Greenberg et al., 2008; Lucas and Sherry, 2004; Sherry et al., 2006). As a result, the AVGUG (Analysis of Video Game Uses and Gratifications) instrument has been designed (Sherry et al., 2006) to measure the extent to which the gratification of *Competition, Challenge, Social Interaction, Diversion, Fantasy* and *Arousal* are met by a specific game for a specific audience.

A second lens focuses on humans as hedonists, striving for the attainment of one or more specific desired psychological states associated with gameplay, such as flow (Csikszentmihalyi, 1990; Csikszentmihalyi and Bennett, 1971), immersion (Brown and Cairns, 2004; Ermi and Mäyrä, 2005), cognitive absorption (Agarwal and Karahanna, 2000), presence (Bracken and Skalski, 2010; Takatalo et al., 2010; Tamborini and Bowman, 2010) or simply enjoyment (Mekler et al., 2014; Vorderer et al., 2004). Academic debate is still taking place on which of such states are more fundamental to a player experience and how they may be aligned or super-positioned in relation to each other, e.g., whereas some researchers conceptualize presence as a subdimension of flow (Novak et al., 1997), other researchers argue it is the other way round (Bryce and Rutter, 2001; Lessiter et al., 2001).

To measure the extent to which players experience the aforementioned psychological states during a game, again several questionnaires have been put forward. Cheng et al. (2015) developed and validated the Game Immersion Questionnaire (GIQ) with three subscales spanning seven constructs (Engagement with *Attraction, Time investment* and *Usability*; Engrossment with *Emotional attachment* and *Decreased perceptions*; and Immersion with *Presence* and *Empathy*), Jennett et al. (2008) construed the Immersive Experience Questionnaire (IEQ) with five constructs (*Cognitive Involvement, Emotional Involvement, Real World Dissociation, Challenge* and *Control*), Brockmyer et al. (2009) put forward the Game Engagement Questionnaire (GEQ) with four constructs (*Absorption, Flow, Presence* and *Immersion*), and IJsselsteijn et al. proposed the Game Experience Questionnaire (GEQIJ) (IJsselsteijn et al., 2008) with seven constructs (*Competence, Sensory & Imaginative Immersion, Flow, Tension/Annoyance, Challenge, Negative affect* and *Positive affect*).

This list of instruments to measure player experiences is not exhaustive, and each of the instruments differs to some extent in the psychological constructs considered as fundamental (Norman, 2013).

Certain theorizations and related instruments also span both perspectives of need fulfillment and psychological state attainment. Based on Social-cognitive theory (SCT), Grove et al. (2016) developed the Digital Games Motives Scale (DGMS). This scale acknowledges that outcome expectations can be game-internal (outcomes of gameplay that are intrinsically enjoyable such as *performance* or *narrative* related outcomes), game-external (outcomes of gameplay that serve as a mediator between the individual and their sociocultural context, such as *Social, Pastime, Habit, Escapism*) and normative (outcomes based on moral standards such as *moral self-reaction*) (De Grove et al., 2014). Moreover, some of the scales above, while originating from a need fulfillment perspective, still complement the instrument with desirable psychological states, or vice versa. For example, the PENS questionnaire, in addition to basing itself on three universal needs for personal growth, also added the desired state of *Presence* and *Intuitive controls*. The Player Experience Scale by Pavlas et al, includes on the one hand items on *Freedom* and *Autonomy*, as well as the construct of

Focus (i.e., immersion and presence).

Instruments also differ in the extent to which they are scientifically validated. Despite the large number of player experience questionnaires available, in most cases empirical validation is limited or absent. Over the past years, as the GUR field matures, this shortcoming has come under scrutiny (Law et al., 2018; Johnson et al., 2018). Law et al. (2018) found that the psychometric properties of the GEQ by Jjsselsteijn et al. (2008), one of the most used scales in GUR research, have yet to be established. In fact, their validation study shows that “the factor structure of the GEQ is not stable”. Similarly, Johnson and colleagues reported that “popular measures of video game player experience typically have not been empirically validated”. Johnson et al. (2018). They point out that this not only applies to the aforementioned GEQ, but equally to the PENS (Ryan et al., 2006), another scale popular among GUR experts. Although the PENS is backed by a strong theoretical model, authors write “[...] full empirical validation for either scale is yet to be published [...]”. Johnson et al. (2018) conducted factor-analytic investigation of not only the PENS but equally the GEQ and found, to a greater extent for the GEQ and a lesser extent for the PENS, the theorised structure cannot fully be supported. Denisova et al. (2016) provides support for the convergence of three scales (PENS Ryan et al., 2006, GIQ Jennett et al., 2008, GEQ by Brockmyer et al., 2009) on a higher order construct of engagement. However, this does not imply a formal validation of the underlying empirical models (i.e. underlying constructs) of these different scales. The authors also conclude that: “As things currently stand, all three [PENS, GIQ, GEQ-Brockmyer] seem to function as reasonable measures of player engagement in a game. However, we suggest that there is the opportunity to develop a more refined questionnaire based on these three [...].” Hence, for authors who wish to understand Player Experience beyond engagement, current widely used scales may not suffice or need further validation.

In sum, the aforementioned theories and questionnaires have different levels of validation and different perspectives on what constitutes a ‘good’ player experience. What these questionnaires have in common is their focus on motives or psychosocial experiences obtained by players, and not so much on players’ actions that give rise to these higher-level experiences. However, GU researchers often require *actionable* insights. They equally need to draw on models and methods to understand how game design choices give way to these higher-level player experiences.

2.2. Player experience as a consequence of player actions

To deliver actionable insight, a player experience can also be conceptualized as the result of good game design choices. The recognition that player experiences are the result of player actions during game play, which in turn are the result of a game designer’s choices is reflected in much game design literature. For example, this division can equally be found in Salen & Zimmerman’s framing of rules, play and culture (Salen and Zimmerman, 2003), Schell’s book of lenses (Schell 2014), Deterding’s argument for moving from game design elements to gamefulness (Deterding et al., 2011), the MDA-framework (MDA: A formal approach to game design and game research, 2004), in the many heuristics (Desurvire and Wiberg, 2009; Korhonen and Koivisto, 2006), design principles (Sweetser and Wyeth, 2005) or cards (Lucero and Arrasvuori, 2010) that help translate psychosocial experience concepts into operational guidelines.

In User Experience research, the interplay between a product’s design features and a user’s experiences is also reflected in Cockton’s treatment on value-centered design (Cockton, 2004), achieved by connecting means (product features) with ends (user values) (Cockton, 2008). This can also be found in Hassenzahl’s model (Hassenzahl, 2005), that distinguishes between the intended product character (as aimed for by the designer) on the one hand, and the apparent product character (as experienced by the user) on the other hand. In turn, this apparent character consists of pragmatic and hedonic attributes and

leads to different types of consequences: behavioural consequences (e.g., increased time spend with the product) or emotional consequences (e.g., pleasure, satisfaction) and ultimately a judgment about the products appeal (e.g., “It is good/bad”). To measure the apparent product character, Hassenzahl devised the AttrakDiff scale (Hassenzahl et al., 2003), measuring both pragmatic aspects (i.e., manipulation) and hedonic aspects (i.e., stimulation, identification and evocation).

2.2.1. Means-End theory

The acknowledgement that product attributes produce consequences at different levels of experience also has its epistemological roots in a well-researched domain in consumer psychology, i.e., Means End (ME) theory (Gutman, 1982; Reynolds and Gutman, 1988). As early as the 1970s, marketing researchers were putting forward that consumers do not think of products as a sum of attributes, but rather in terms of the benefits that they may bring (Gutman, 1978; Gutman and Reynolds, 1979). More specifically, Means-End theory posits that consumers choose a product or object not simply because it contains specific attributes (the ‘means’) but rather because these are instrumental to achieving certain benefits or desired ‘consequences’. These consequences are desired because they align with personal values (the ‘ends’). In other words, users’ product preferences, usage behavior and experiences are dependent on how they perceive certain product attributes as most likely to produce certain desired consequences which enable them to meet their values. ME theory, in turn rooted in Expectancy-Value theory (Fishbein and Ajzen, 1975; Grunert and Bech-Larsen, 2005) has built a rich research tradition for over 50 years (Olson and Reynolds, 1983; Reynolds and Olson, 2001) in social psychology and consumer research.

2.2.2. The means-end chain

In its most basic form a ME chain takes the shape of chain: from product *Attributes* to desired *Consequences* for the users, ultimately aligning with their *Values* (see Fig. 1). Whereas product attributes exist separately from the user, and values persist, irregardless of whatever product at hand, consequences unfold in the interaction between the product and the user. Attributes on their own have no consequences, and thus, no relevance. Equally, personal values or motives that cannot be tied to product use are not meaningful either for researchers who aim to derive insight in a user’s preferences. Yet, consequences capture the experience with the product (Olson and Reynolds, 2001). Therefore, Means-End theory explicitly advocates a focus on consequences to understand a user’s experience of and preference for a product.

Although consequences can be modeled at varying levels of abstraction and different fine grained divisions in Means-End chains exist (Kardes et al., 2010), eventually two levels of consequences were found sufficient for most marketing analyses purposes. In consumer research, the four-level chain, ranging from Attributes to *Functional* Consequences to *Psychosocial* Consequences and Values, has seemed most useful and become a de facto standard (Reynolds and Olson, 2001).

Functional consequences are situated at the usage level. These are the immediate and tangible consequences that are experienced directly by consumers, during the use of the product. *Psychosocial* consequences

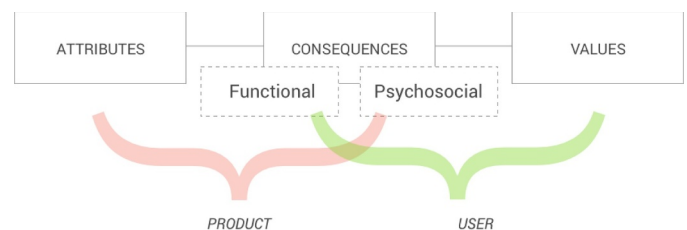


Fig. 1. A means-end chain consists of attributes causing certain desired functional and psychosocial consequences for consumers, aligning with their personal values.

exceed the immediate usage level and reach into the social or, psychological level. These consequences are the more emotional experiences and may be shaped even after the usage of the product.

Means-End theory originated in Marketing research to understand the consumer buying decision process. However, it has been successfully applied in user experience research as well to derive insights and formulate ‘implications for design’ (e.g., Kwak et al., 2014; Tuunanen and Govindji, 2016; Wu et al., 2014). Yet, adaptations were made to serve the particular disciplinary needs in the field of HCI and user experience design (Vanden Abeele et al., 2012a; Wu et al., 2014; Zaman, 2008).

The means-end chains (MECs) that are revealed in an experimental UX context are typically less elaborate than MECs that consumer researchers find with respect to established products. Also, it should be taken into account that for certain values, people rely on hypotheses instead of actual user experiences within a real life context. Therefore, values listed by users might not be as reliable or simply nonexistent (Vanden Abeele and Zaman, 2009). Despite these differences, ME approaches have been found useful in UX research.

2.2.3. Means-End Theory applied to games

Means-End theory approaches have also successfully been applied to investigate player experiences (see e.g., Celis et al., 2013; Sundström et al., 2014; Vanden Abeele et al., 2009; Vanden Abeele et al., 2012b; Zaman, 2008). Many of these studies investigate how different game variants lead to different functional and psychosocial consequences.

As aforementioned, the notion that *low-level* player actions are linked to *higher-level* experiences is not new. Salen and Zimmerman detail how meaningful play in a game “emerges from the relationship between player action and system outcome,” but equally how it only occurs when “The relationships between actions and outcomes in a game are both discernable and integrated into the larger context of the game, play and culture” (Salen and Zimmerman, 2003). In “The Art of Game Design”, Schell discusses how the “The experience rises out of a game” (Schell, 2014, Ch.3) and presents a set of lenses to help designers think about what in-game actions can bring about e.g., *Curiosity*, *Surprise* or *Fun*.

These linkages between product features and functional consequences and psychosocial consequences also resonates with the Mechanics-Dynamics-Aesthetics (MDA) framework (MDA: A formal approach to game design and game research, 2004). This framework emphasizes the causal link between design choices made and the experience held by the player. In particular, the *Mechanics* (components of a game) are set into motion during the run-time of the game, causing *Dynamics*. These direct interactions between a player’s inputs and the game’s outputs lead to *Aesthetics*, i.e., emotional responses evoked in the player. From the above account, it becomes apparent that the concept of Mechanics in the MDA framework aligns with the concept of Attributes in ME theory. Moreover, the concept of Dynamics aligns with Functional Consequences, i.e., the immediate consequences while playing a game. The concept of Aesthetics, i.e., the emotional response, aligns with Psychosocial Consequences in ME theory. Hence, the Means-End approach lends theoretical support to the MDA framework.

Implicitly, the importance of measuring player experience at the functional level as well, is also acknowledged in some of the above questionnaires. For example, the PENS, while rooted in Self-Determination Theory, has added the construct of *intuitive controls*. While having controls that are intuitive is not necessarily a universal need fostering personal growth, it was found important enough by the PENS researchers to include in their instrument. In a similar vein, Cheng et al., (2015) incorporated the construct of usability in their Game Immersion Questionnaire, to assess player experience at the functional level.

However, from a conceptual and theoretical level, none of the current questionnaires has an explicit focus on including constructs at the *Functional Consequences* level or understanding linkages across the elements at different levels.

Considering the importance of better understanding how player actions contribute to a player experience, and given this void in the current instruments to measure player experience at different levels of abstraction, in this paper, the development and validation of such an instrument is presented. The ambition of the Player Experience Inventory (PXI) is to present a rigorously validated scale to support GU researchers to measure player experience at both the level of *Functional Consequences* and at the level of *Psychosocial Consequences*. In this manner, the PXI also supports exploring linkages across the constructs at the different levels, investigating to what extent certain functional consequences are causal to certain psychosocial consequences.

3. Method

Developing a measurement instrument is a longitudinal process (DeVellis, 2012). In total, seven studies were conducted to complete the process of scale conception, scale construction and scale validation, as outlined in Table 1. In *study 1*, 31 experts active in the Games User Research field, were asked to review a first selection of constructs and items, and discuss whether these were important to include. Based on the insights generated by these experts, the theoretical model was devised. In *study 2*, 33 GUR experts provided feedback on this revised model, via a Q-sorting procedure (Anderson and Gerbing, 1991). Given the positive results of the Q-sort, a survey was set up in *study 3*: data was collected from 228 students who were asked to evaluate a salient play experience. An exploratory factor analysis confirmed the existence of the constructs, but pruning the scale was necessary. After this, a first confirmatory factor analysis was conducted, revealing moderate model fit. Further pruning was performed to improve model fit and to improve the parsimony of the scale. At the end, the model consisted of ten constructs, evaluated by three items each, and providing good model fit. Next, in *study 4*, the model fit as well as the structural and metric invariance was validated, via data from an additional survey study conducted with a new sample of 138 students. In *study 5*, configural invariance was again validated, this time with data stemming from play testing or experimental evaluations, rather than evaluating a salient play experience (the data was collected directly after game play). In *study 6*, a final evaluation of the model fit, convergent and discriminant validity was carried out on the combined data collected in study 3, 4 and 5. As a final step, in *study 7* criterion validity was assessed via the data collected during an experimental evaluation of player experience, with 40 players. Additionally, the causal model was tested by the data from study 6, as well as the data from studies 3, 4, and 5. In all, scale development and validation included 529 participants and 64 experts.

4. Scale conception

4.1. Study 1 - Interviews and card sorting with 31 GUR experts

At the initial stage, three game researchers and co-authors of this paper reviewed 124 scales containing over 800 constructs¹, used in game research, with the aim to devise a scale to measure player experience across a broad range of game genres, including serious games and gamified applications. Hence, we focused on the core elements that contribute to player experience and that are generalizable across most playful products. Therefore, we decided to exclude constructs that

¹ Fourteen databases were scanned: ProQuest, Ebscohost, ACM, IEEE Xplore, Springerlink, CogPrints, Emerald, InfoSci, Web of Science, Scopus, ScienceDirect, Informit, Project Muse and a Synthesis of the Digital Library of Engineering and Computer Science. The search strategy used was: (evaluat* OR model OR scale OR questionnaire OR survey OR measur* OR immers* OR flow OR motiv* OR presence OR enjoy* OR engag* OR fun) AND (“computer games” OR “video games” OR “videogames”). The full list of scales and constructs can be found in the Data in Brief accompanying this article

Table 1
Overview of the different studies, number of participants, age and gender.

	N	Dropped cases	Age		Gender	
			Mean Age	SD	Male	Female
SCALE CONCEPTION						
Study 1: Q-Sorting (closed) and GUR expert interviews	31	0			22	9
Study 2: Q-Sorting (open) with GUR experts	33	0			26	7
SCALE CONSTRUCTION						
Study 3: Exploratory and confirmatory factor analysis	228	9	20.0	2.66	173	53
SCALE VALIDATION						
Study 4: Multigroup Invariance CFA- Assessing configural and metric invariance	138	8	21.5	4.64	100	37
Study 5: Validation of the model with experimental data	163	12				
Study 6: Model fit, convergent& discriminant validation all data	529	29				
Study 7: Criterion validation	40	0	27	7.6	23	17

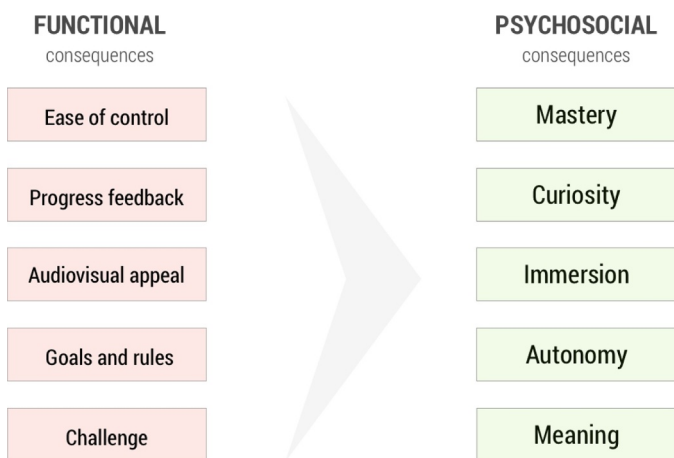


Fig. 2. Revised model of Player Experience Inventory; constructs are separated at the different levels of Functional and Psychosocial Consequences.

polled specifically for social interaction, given that this is often missing in single-player games. Likewise, constructs that focused strongly on a ‘narrative’ were excluded as this would not apply to certain game genres. After three iterations, consensus was reached and the following constructs were included in the first study: *ease-of-control*, *aesthetic appeal*, *absorption*, *interest*, *competence*, *autonomy*, *effort*, *meaning* and *enjoyment*. Per construct, five to seven items were generated by the authors to poll for this construct. In this first study on scale conception, the theoretical model was not yet finalized, and constructs still included items at both the functional as well as the psychosocial level. This resulted in a first selection of nine constructs and 53 items (see the Data in Brief accompanying this article).

Next, following DeVellis scale development guidelines (DeVellis, 2012), GUR experts were contacted to participate. GUR experts were defined as people who were active in the field of game design and development, game evaluation and game research. Graduate students, as well as senior researchers and industry professionals were included in the sample. These game researchers were found via the personal contact lists of the authors of this paper, and via snowball sampling. In the end, 31 GUR experts responded favorably (22 from academia and 9 from game industry), and reviewed and discussed the constructs and items. They first completed an item-sort exercise, according to the Q-sorting procedure (Anderson and Gerbing, 1991) (i.e., closed card) sorting. GU researchers were given the list of constructs and items via the online tool OptimalSort (User Experience). Next, they were asked to assign items to the construct that, in their judgment, matched best. In addition, 24 out of these 31 GU researchers adhered to a think-aloud protocol during the sorting procedure, and participated in an in-depth interview after the sorting, reflecting on the constructs and items and how they conceptualized player experience.

The results of the Q-sort (i.e., the proportion of substantive agreement among experts and the substantive-validity coefficient) showed that the majority of items were correctly assigned and reached the desired cut-offs ($\geq .7$) (Anderson and Gerbing, 1991). However, the in-depth interviews also revealed the dislike of GU researchers towards certain labels (e.g., *absorption* or *aesthetics*). More profoundly, GU researchers found some constructs less relevant, e.g., *effort*. Particularly, game researchers in industry emphasized that items were still heavily focused on psychosocial consequences. To provide actionable information for them, constructs that polled specifically at the functional level were found missing. Experts also pointed out the need to refine the theoretical model.

4.1.1. Revising the theoretical model

Based on the results of this expert study, we re-inspected the constructs and elaborated the theoretical model drawing on Means-End (ME) theory; we articulated the two levels of the instrument and the causal link between *Functional* consequences and *Psychosocial* consequences. Constructs and items were revised: the construct of *effort* was removed, three new constructs (*Progress feedback*, *Goals and Rules*, *Challenge*) were added. In addition, we renamed constructs and reworded/removed problematic items and added new items (see Fig. 2). The final model is comprised of five constructs polling for *functional* consequences and five constructs polling for *psychosocial* consequences. The *functional* level contains the following constructs: *Ease of Control* to be understood as ‘The extent to which a player finds the actions to control the game clear and intuitive’ (5 items), *Challenge* to be understood as ‘The extent to which the specific challenges in the game match the players skill level’ (5 items), *Progress Feedback* to be understood as ‘The extent to which it is clear to the player how well he or she is doing in the game’ (5 items), *Goals and Rules* to be understood as ‘The extent to which the overall objective and rules are clear to the player’ (5 items) and *Audiovisual Appeal* to be understood as ‘The extent to which a player appreciates the audiovisual styling of the game’ (5 items). The *Psychosocial Consequences* level contains the following constructs: *Meaning* to be understood as ‘A sense of connecting with the game, resonating with what is important’ (5 items), *Immersion* to be understood as ‘A sense of immersion and cognitive absorption, experienced by the player’ (6 items), *Mastery* to be understood as ‘A sense of competence and mastery derived from playing the game’ (6 items), *Curiosity*: ‘A sense of interest and curiosity roused by the game’ (5 items) and *Autonomy* to be understood as ‘A sense of freedom and autonomy to play the game as desired’ (5 items).

4.2. Study 2 - Q-sorting with 33 GUR experts

Given the substantial changes and the fully revised theoretical model, constructs and items were again reviewed by 33 GUR experts who were not involved in study 1. Again, this was done via a clustering, i.e., a Q-sort procedure via OptimalSort (User Experience). However,

Table 2
Study 3. Exploratory factor analysis with 35 items, N = 228.

	AA	CUR	MEA	IMM	GR	GR	AUT	CH	PF	EC
MEA_3	.035	.056	.899	-.047	-.056	-.007	-.091	.002	.019	.014
MEA_4	-.009	.013	.763	-.005	.075	.115	-.055	.007	.051	-.139
MEA_5	-.045	-.071	.817	.151	-.018	-.011	.071	-.040	-.034	.126
MAS_2	-.104	.043	.053	-.088	.153	.774	.036	.010	.028	-.098
MAS_3	-.059	.029	-.045	-.089	-.035	.842	-.043	.091	-.078	.112
MAS_4	.034	-.077	.129	.030	-.021	.609	.150	-.019	-.004	.012
MAS_5	.026	-.067	.290	.044	-.079	.427	.081	.028	.003	.026
IMM_1	.053	.250	-.133	.566	-.027	.190	-.068	-.018	-.069	-.040
IMM_4	-.094	-.066	.004	.853	-.002	-.267	.159	-.067	.027	.097
IMM_5	-.007	-.027	.074	.744	.021	-.032	.000	.097	.032	.061
IMM_6	.042	-.037	.129	.601	.008	.136	-.056	-.020	.005	-.165
AUT_2	.079	.117	-.001	.128	.077	.013	.641	.051	-.069	-.152
AUT_3	-.057	.050	-.013	-.020	.015	.053	.789	.060	.008	-.110
AUT_4	.032	.106	-.072	.050	-.092	.025	.692	.075	.020	-.013
CUR_1	-.005	.835	-.086	.063	-.014	.033	-.058	-.045	.006	.011
CUR_2	-.026	.842	.007	-.018	.003	-.039	.115	-.043	.042	.024
CUR_3	.078	.532	.221	-.002	-.023	.031	.150	-.039	-.061	.114
CUR_4	-.027	.834	.021	-.089	-.005	-.013	.122	.043	.048	.041
EC_3	-.161	.091	-.111	.093	.099	.207	-.172	-.029	.127	.504
EC_4	.033	.038	.056	-.033	.060	-.100	-.121	.081	-.108	.780
EC_5	.200	.042	-.012	.030	-.032	.143	-.002	.019	-.019	.500
CH_2	-.066	.045	.112	-.075	-.002	.043	-.019	.596	.028	.065
CH_3	.062	-.054	-.080	-.040	-.094	.019	.113	.773	.100	.046
CH_5	.001	-.078	-.057	.091	.112	.086	.137	.607	-.084	-.001
PF_3	-.017	.167	.054	-.007	.117	-.154	-.037	.056	.720	-.106
PF_4	.012	-.167	-.103	.022	-.030	.119	.211	-.128	.745	.134
PF_5	.034	.045	.073	.011	-.046	-.018	-.148	.103	.787	-.076
AA_2	.901	.013	.010	-.035	.049	-.162	.022	.045	-.050	-.087
AA_3	.928	.004	.010	.006	.015	-.135	-.062	.084	-.009	.048
AA_4	.854	-.045	-.066	-.048	-.056	.236	.061	-.143	.026	.020
AA_5	.686	-.005	.056	.016	.077	-.017	.011	-.022	.113	.088
GR_1	.111	-.024	-.022	.061	.532	.135	-.234	.055	.025	-.118
GR_3	.122	.114	-.063	.073	.612	.049	-.123	-.017	-.018	-.017
GR_4	-.061	-.073	.010	.016	.580	-.026	-.058	.093	.025	.198
GR_5	-.036	-.059	.037	-.110	.862	-.038	.285	-.098	-.001	.078

Extraction method: Principal Axis Factoring (PAF). Rotation method: Promax.

different from study 1, this time experts were presented the 52 items of the ten constructs (see Fig. 2) without labels, akin to an open card sorting. Experts were asked to cluster items in as many groups as they saw fit, and to come up with names for the labels themselves. This resulted in a more challenging sorting exercise, but also allowed for more input from the GU researchers.

Based on the open Q-sorting, first a similarity matrix was computed, showing the percentage of participants who agree with each item pairing, i.e., the percentage of experts that similarly grouped two items (see the Data in Brief accompanying this article) and clustering those items that are grouped together more often. The results of this clustering by experts lent support to the theoretical model, and confirmed the existence of the theorized constructs. Next, average pair agreements were computed (i.e., how often one item was grouped with another item) per cluster. For the ten constructs, these pair agreements ranged between 95.5% and 66.3%, whereas the average pair agreement between items of different constructs was 6.1%. Giving these encouraging results of study 2, we set forth to construct the measurement model for the Player Experience Inventory.

5. Scale construction

5.1. Study 3 - Creating the measurement model

To create the measurement model, a survey was handed out containing the 52 items, to 237 students who participated in a summer school. The items were randomized, and to be rated on a 7-point Likert scale. In addition to the 52 items, the survey also asked for the name of the game (which the player was evaluating), and included extra items polling for overall enjoyment and appreciation of the game to assess

criterion validity. Finally, the survey also polled for gender and age. Students were given the instrument as a paper and pencil survey. Participants were asked to fill out the survey for a specific game in mind for which the player experience was still salient: a game they played recently or a game they had played often. Participants enlisted games came from a variety of video game genres. The top ten genres were puzzle games (17.5%), followed by action adventure games (15.7%), first person shooters (13.4%), sport simulation games (10.6%), multi-player online battle arenas (7.8%), massively multiplayer online role-playing games (6.5%), racing games (5.5%), real-time strategy (5.5%), action role playing games (4.1%), and social simulation games (3.7%). Filling out this survey took approximately 15 minutes. The demographics of participants can be found in Table 1.

Data cleaning Data cleaning was carried out according to the process put forward by Gaskin and Carpenter (2018). When more than 10% of data was not provided by the participant, or when suspicious answer patterns were found (graphical patterns in the answers or limited variance in answers suggesting disengagement), these were omitted from the data set. In total nine surveys were dropped; 228 samples were retained. Data of eight missing values were imputed by using the median of the scores on the other construct items.

Preliminary reliability testing Reliability was checked by verifying Cronbach's α , item-whole correlations and squared multiple correlations of the different items. Items performing suboptimally on item- or construct-level were removed. In particular, an item was removed when one or more of the following factors was present: extreme means (< -2 or $> +2$), limited item variance < 1.0 , a low item-whole correlation $< .4$ or a low square multiple correlation $< .3$, or when removing the item would improve Cronbach's Alpha. Four items were dropped, 48 items remained.

Table 3

Cut-off criteria for model fit indices, according to [Cabrera-Nguyen \(2010\)](#); [Hooper et al. \(2008\)](#); [Hu and Bentler \(1999\)](#); [Tabachnick and Fidell \(2007\)](#).

	Acceptable	Excellent
X2/df	< 5	< 2
CFI	> .90	> .95
TLI (NNFI)	> .90	> .95
RMSEA	< .08	< .06
SRMR	< .09	< .08

Exploratory Factor analysis Next, we conducted exploratory factor analysis (EFA), using IBM SPSS version 25, using Principal Axis Factoring with Promax rotation. Based on our theoretical model, we set the extraction fixed to ten factors. With a Kaiser-Meyer-Olkin (KMO) index of 0.874 and a significant Bartlett's test of sphericity ($\chi^2 = 6404$, $p < .001$), sampling adequacy was considered good. The ten factors explained 66% of the variance. However, as some items showed undesirable cross-loadings, additional pruning was needed. Thirteen more items were removed from the analysis. A final EFA with 35 items was retained, sampling adequacy was good (KMO-index = 0.855, $\chi^2 = 4468$, $p < .001$, communalities of all items were above $> .4$), explaining 73% of variance in the dataset (see [Table 2](#)). All items loaded highly ($\geq .5$) and uniquely on their factors ($\leq .3$ on other factors) suggesting good convergent and discriminatory validity (see [Table 2](#)).

Confirmatory factor analysis To further investigate the factor structure and fit of the model, a confirmatory factor analysis in AMOS was conducted. The fit of the measurement model is evaluated by means of several fit indices (see [Table 3](#)). Researchers suggest combinational rules ([Cabrera-Nguyen, 2010](#); [Hooper et al., 2008](#); [Hu and Bentler, 1999](#); [Tabachnick and Fidell, 2007](#)) based on the following fit indices and norms (Relative or normed Chi-Square (χ^2/df), the Root Mean Square Error of Approximation (RMSEA), Standardised Root Mean Square Residual, Tucker-Lewis Index (TLI, also known as the Non-Normed Fit Index (NNFI), and the CFI (Comparative Fit Index)) as shown in [Table 3](#), but also acknowledge that for smaller sample sizes, or for more complex models, ($N \leq 250$) this stringent criteria may be harder to obtain ([Cabrera-Nguyen, 2010](#); [Hu and Bentler, 1999](#)).

Running the model with 35 items resulted only in a moderate fit (CFA = .90, RMSEA = .059, $\chi^2 = 1.79$), (see [Table 4](#), Study 3: Model with 35 items). To improve the model fit and to increase parsimony of the scale, five additional items were removed. Candidates for removal were those items that showed either lower convergent validity or higher cross loadings, according to the procedure put forward by [Gaskin and Carpenter \(2018\)](#). Additionally, we aimed for a parsimonious scale, but still retaining minimally 3 items per factor. This model with 30 items (3 items per factor) significantly improved model fit (CFI = .935, RMSEA = .050, $\chi^2 = 1.57$) (see [Table 4](#), Study 3: Model with 30 items). Hence, this model with 30 items was used for further validation.

Table 4

Model fit indices for the different studies.

	N	χ^2/df	TLI	CFI	SR-MR	RMSEA	CI90 -	CI90 +
Study 3: Model 35 it.	228	1.790	.884	.900	.058	.059	.053	.065
Study 3: Model 30 it.	228	1.572	.922	.935	.050	.050	.042	.058
Study 4: Model 30 it.	138	1.434	.903	.928	.065	.056	.045	.067
Study 5: Model 30 it.	163	1.653	.919	.937	.063	.063	.054	.072
Study 3 and 4	366	1.628	.948	.957	.044	.041	.035	.047
Study 3, 4 and 5	529	2.089	.947	.956	.041	.045	.041	.050

6. Scale validation

6.1. Study 4: Multigroup invariance CFA - Assessing configural and metric invariance

To test for configural and metric invariance, an additional sample of data was collected. This extra sample allows testing the factor stability of the final model on a new sample. Moreover, this extra group of data also allows for a *multiple-group invariance confirmatory factor analysis* (MGCFAs) to be carried out ([Byrne, 2016](#)). MGCFAs is used to compare latent variable means, variances, and co-variances across groups while holding measurement parameters invariant. This procedure can lend support to the constructs having the same theoretical structure and meaning across studies, and, hence, this technique is suitable to confirm the structural validity of the measurement model.

Data collection and cleaning In a similar manner as study 3, data was collected via surveys, from student populations in Belgium and Austria. Participants were asked to fill out the survey with a specific game in mind for which the player experience was still salient. In total 146 responses were collected (see [Table 1](#)). The top ten of participants' enlisted genres were puzzle games (23.7%), followed by action adventure games (13.2%), first person shooters (11.8%), multiplayer online battle arenas (7.9%), sport simulation games (7.9%), social simulation games (7.9%), real-time strategy (6.6%), action role playing games (6.6%), racing games (5.3%) and massively multiplayer online role-playing games (3.9%). Data cleaning was carried out in a similar manner as with study 3 ([Gaskin](#)). When more than 10% of data was not provided, or when suspicious answer patterns were found, these were omitted from the data set. In total eight surveys were dropped; 138 surveys were retained. Data for nine missing values were imputed by using the median of the scores on the other construct items.

Configural and metric invariance tests First, a confirmatory factor analysis in AMOS was conducted, the results suggest the model had an acceptable fit (CFI = .928, RMSEA = .056, $\chi^2 = 1.434$). In addition, the data of study 3 and 4 was combined, and this provided an excellent model fit (CFI = .957, RMSEA = .041, $\chi^2 = 1.628$).

Second, configural invariance was tested according to the 'multiple-group invariance confirmatory factor analysis' (MGCFAs) technique. First the model fit of the general model was assessed across the two data sets (i.e. the data from study 3 and the data from study 4), in an unconstrained manner. Good model fit was achieved, (CFI = .932, RMSEA = .037, $\chi^2 = 1.503$) suggesting good configural invariance (see [Table 5](#)). As a second step, the measurement weights of factor loadings (regression weights) were constrained. Good model fit was again achieved. Moreover the fit measures were tested between unconstrained and constrained models and were not found to be significantly different ($\chi^2 = 27.25$, $df = 30$, $p = .610$) (see [Table 6](#)). Hence, this suggests metric invariance of the model as well.

In sum, based on the good fit indices, and the small incremental change in fit indices between unconstrained and constrained models, it can be concluded that the measurement model has configural and metric invariance.

Table 5
Fit indices for the Multigroup Invariance Confirmatory Factor Analysis.

	χ^2/df	NFI	RFI	IFI	TLI	CFI	RM-SEA	AIC	BCC
Unconstrained	1.503	.825	.789	.934	.918	.932	.037	1502	1597
Constrained	1.479	.821	.792	.934	.922	.933	.036	1470	1550

6.2. Study 5: validating the model with data from experimental studies

As a last test of configural invariance, we gathered data from experimental game evaluations and play tests. Different from study 3 and 4, we collected data this time not drawing on delayed recall, where students were asked to score a salient game experience. Rather, players were asked to complete the Player Experience Inventory immediately after playing the game. Given the objective of the Player Experience Inventory to support GUR researchers, the data of study 5 has a higher ecological validity; this last data set aligns with the actual usage situations of the instrument.

Data collection and cleaning As experimental evaluations typically rely on smaller group sizes, study 5 is actually a composite of four different studies collected via three different GUR researchers, active in Canada (N= 29, 1 case dropped, evaluation of commercial prototype), Australia (N=38, 6 cases dropped; N=56, 5 cases dropped, evaluation of COTS MOBA game) and the United Kingdom (N=40, no cases dropped, evaluation of student prototype of a First-Person Shooter game). When experimental evaluations were based on a repeated measures design, only one measurement per participant was included. Because of the reliance on third parties and because of the privacy of individuals participating in playtesting protocols, it was not possible to assess gender or mean age.

The different researchers shared data sets as a CSV file with clearly labeled headers to distinguish the different constructs and items. Data cleaning was carried out in a similar manner as in previous studies. When more than 10% of data was not provided, or when suspicious answer patterns were found, participants were omitted from the data set. In total twelve surveys were dropped; 163 surveys were retained. No missing values were imputed.

Assessing model fit, configural and metrical invariance Confirmatory factor analysis in AMOS was first conducted, on the basis of the data of study 5 alone, the results suggest the model has an acceptable fit (CFI = .937, RMSEA = .063, $\chi^2/df= 1.653$), see Table 3. Hence, this result demonstrates the fit of the model when working with data, collected on the basis of immediate recall, rather than delayed recall. Next, we tested configural invariance again according to the MGCFA technique. First, the model fit of the general model was assessed across the two data sets, i.e. the data from delayed recall (study 3 and 4) and the data from immediate recall (study 5), in an unconstrained manner. Good model fit was achieved, (CFI = .946, RMSEA = .035, $\chi^2/df= 1.662$) suggesting good configural invariance, see Table 7. As a second step, the measurement weights of factor loadings (regression weights) were constrained. Again, good model fit was obtained. However, we did not achieve metrical invariance, factor loadings changed significantly between delayed and immediate recall ($\chi^2 = 77.5$, $df = 30$, $p < .001$), see Table 8.

6.3. Study 6: Assessing model fit, convergent and discriminant validity of the combined data set

As a last step to validate the model fit, and to verify once more convergent and discriminant validity, we combined the data of study 3,

Table 6
Incremental fit indices.

Model Comparison	df	χ^2	p	NFI (delta1)	IFI (delta2)	RFI (rho1)	TLI (rho2)
Measurment. weights	30	27.25	.610	.004	.005	-.003	-.004

Table 7
Fit indices for the Multigroup Invariance Confirmatory Factor Analysis on the basis of immediate versus delayed recall.

	χ^2/df	NFI	RFI	IFI	TLI	CFI	RM-SEA	AIC	BCC
Unconstrained	1.662	.877	.852	.947	.935	.946	.035	1737	1826
Constrained	1.699	.869	.848	.943	.931	.941	.036	1754	1834

4 and 5 (N = 529). No additional data cleaning was carried out.

Assessing model fit First, confirmatory factor analysis in AMOS was conducted. This provided an excellent model fit (CFI = .956, RMSEA = .045, $\chi^2/df= 2.089$).

Assessing convergent and discriminant validity Next, convergent validity and discriminant validity of constructs were by looking at the composite reliability (CR), the average variance extracted (AVE), the maximum and shared variance (see Table 9). Good convergent validity is evidenced by the AVE (all constructs are $\geq .5$ with the exception of Ease-of-Control (.462) which is just below this score). Convergent validity is also evidenced by the composite reliability (CR) which is $\geq .7$ for all factors. Finally, discriminant validity is also good, as the square root of the AVE (the values on the diagonal in Table 10) are greater than any of the inter-construct correlations.

6.4. Study 7 - Assessing criterion validation

As a final step in scale development, the criterion validity of the instrument was assessed. In particular, two specific actions were carried out. First, criterion validity was assessed for the different constructs themselves. Secondly, the mediation model underlying the instrument was investigated.

Data collection and cleaning Data was obtained via a 2 X 2 repeated measures study design, where both game (either a casual game or a First-Person Shooter game) was manipulated as well as the presence of visual embellishments, with the aim of understanding their effect on player experience. Two custom built games were created: one game replicated the mechanics from the well-known arcade game Frogger. This game was chosen its casual arcade style of gameplay. The second game provides a more in-depth and sophisticated 3D game experience, featuring game mechanics of a first person shooter. Both games presented a variant with and without visual embellishments.. After each of the four play sessions, players were asked to fill out a set of different measurement instruments, among which the PXI as well as the PENS (Ryan et al., 2006) and the AttrakDiff2 (AD) (Hassenzahl et al., 2003) scale. PENS was chosen given that it includes constructs closely related to the most used scale in PX at the current moment. However, it lacks items that can poll at the functional level. Therefore, the measurement was complemented with items from AttrakDiff, which is a well known in UX research, conceptually related to ME theory and containing items that poll at the different levels (pragmatic and hedonic). At the end of each game session, players were also asked to rate how much they enjoyed the game itself on a 7-point Likert scale via the item "Please rate how much you enjoyed the game you played?". No data cleaning was needed.

Criterion validity of the PXI constructs First, a mapping was made between the constructs of the PXI and those constructs from the PENS and AttrakDiff2 that are conceptually related. It was hypothesized that constructs that are conceptually related measure related aspects and, hence, should covary. Some mappings were straightforward (e.g., PXI Audiovisual appeal and AD Beauty, PXI Ease of Control and PENS

Table 8
Incremental fit indices.

Model Comparison	df	χ^2	p	NFI (delta1)	IFI (delta2)	RFI (rho1)	TLI (rho2)
Measur. weights	30	77.5	.000	.004	.004	.004	.004

Table 9
Convergent and discriminant validity of PXI constructs and items.

Constructs & Items	CR	AVE	MSV
MEANING:	0.885	0.720	0.420
<i>Playing the game was meaningful to me.</i>	0.821	0.606	0.391
<i>The game felt relevant to me.</i>			
<i>Playing this game was valuable to me.</i>			
MASTERY:	0.791	0.563	0.420
<i>I felt capable while playing the game.</i>			
<i>I felt I was good at playing this game.</i>			
<i>I felt a sense of mastery playing this game.</i>			
IMMERSION:	0.855	0.664	0.403
<i>I was no longer aware of my surroundings while I was playing.</i>			
<i>I was immersed in the game.</i>			
<i>I was fully focused on the game.</i>			
AUTONOMY:	0.908	0.767	0.403
<i>I felt a sense of freedom about how I wanted to play this game.</i>			
<i>I felt free to play the game in my own way.</i>			
<i>I felt like I had choices regarding how I wanted to play this game.</i>			
CURIOSITY:	0.727	0.473	0.391
<i>I felt eager to discover how the game continued.</i>			
<i>I wanted to explore how the game evolved.</i>			
<i>I wanted to find out how the game progressed.</i>			
EASE OF CONTROL:	0.790	0.556	0.267
<i>I thought the game was easy to control.</i>			
<i>The actions to control the game were clear to me.</i>			
<i>It was easy to know how to perform actions in the game.</i>			
CHALLENGE:	0.821	0.605	0.412
<i>The game was challenging but not too challenging.</i>			
<i>The game was not too easy and not too hard to play.</i>			
<i>The challenges in the game were at the right level of difficulty for me.</i>			
PROGRESS FEEDBACK:	0.922	0.797	0.265
<i>The game gave clear feedback on my progress towards the goals.</i>			
<i>I could easily assess how I was performing in the game.</i>			
<i>The game informed me of my progress in the game.</i>			
AUDIOVISUAL APPEAL:	0.808	0.584	0.412
<i>I enjoyed the way the game was styled.</i>			
<i>I liked the look and feel of the game.</i>			
<i>I appreciated the aesthetics of the game.</i>			
GOALS AND RULES:			
<i>The goals of the game were clear to me.</i>			
<i>I grasped the overall goal of the game.</i>			
<i>I understood the objectives of the game.</i>			

Table 10
Discriminant validity, Square root of AVE greater than inter-construct correlations.

	MEA	MAS	IMM	AUT	CUR	EC	CH	PF	AA	GR
Meaning	.849									
Mastery	.589	.778								
Immersion	.648	.476	.750							
Autonomy	.628	.519	.474	.815						
Curiosity	.569	.463	.553	.635	.876					
Ease of Control	.244	.625	.335	.281	.355	.688				
Challenge	.370	.517	.474	.456	.425	.462	.746			
Progress feedback	.374	.444	.368	.363	.278	.508	.477	.778		
Audiovisual appeal	.515	.370	.512	.414	.508	.410	.389	.469	.893	
Goals and rules	.246	.504	.316	.135	.237	.600	.365	.642	.425	.764

Table 11
Mapping constructs from the Player Experience Inventory to conceptually related constructs of the PENS and AttrakDiff scale.

Construct PXI	Construct PENS or AD	Pearson's R	Pearson'S R
Meaning	AD Attractiveness	0.638**	0.436**
Mastery	PENS Competence	0.884**	0.785**
Immersion	PENS Presence	0.596**	0.543**
Autonomy	PENS Autonomy	0.720**	0.603**
Curiosity	PENS Presence	0.697**	0.653**
Ease of Control	PENS Intuitive controls	0.856**	0.836**
Audiovisual appeal	AD Beauty	0.732**	0.815**
Challenge	PENS Competence	0.722**	0.325*
Goals and Rules	AD Pragmatic Quality	0.711**	0.733**
Progress feedback	AD Pragmatic Quality	0.643**	0.522**

Intuitive controls, PXI Autonomy and PENS Autonomy, PXI Mastery and PENS Competence). Other mappings were realized upon inspection of the items, e.g., the PENS construct of Presence is a broadly measured construct with nine items, covering both aspects of immersion as well as curiosity, therefore it was paired both with PXI Immersion and PXI Curiosity. The mapping can be found in Table 11. In a similar vein, PXI meaning was paired with AD Goodness upon inspection of the AD items that poll for motivation and appeal. Finally, AD Pragmatic was paired with both PXI Goals and Rules and Progress feedback, as this construct polls for 'usability' related aspects. Upon the mapping, a bivariate correlation analysis was conducted (see Table 11). It was found that all paired constructs are highly correlated. This lends support to the criterion validity of the instrument.

Mediation analysis Finally, we aimed to inspect the theoretical model underlying the instrument. More specifically, we expect that *Functional consequences* positively predict *Enjoyment* (hypothesis 1). We also expect that *Functional consequences* positively predict *Psychosocial Consequences* (hypothesis 2). Finally, we predict that *Psychosocial Consequences* positively predict *Enjoyment* (hypothesis 3). However, we also expect a mediation effect, in particular we expect that the effect of *Functional consequences* on *Enjoyment* is mediated via *Psychosocial Consequences* (hypothesis 4).

To perform this mediation analysis, again we used the combined dataset of study 3, study 4 and study 5. No extra datacleaning was carried out. Game enjoyment was measured by computing the average of three items: "I liked playing the game", "The game was entertaining" and "I had a good time playing the game". *Functional consequences* score was computed as the average of the means for the constructs of *Ease of Control*, *Challenge*, *Progress feedback*, *Audiovisual appeal* and *Goals and*

Rules. *Psychosocial consequences* score was computed as the average of the means for the constructs of *Meaning, Mastery, Immersion, Autonomy* and *Curiosity*. The mediation analysis was conducted in AMOS, and direct and indirect effect were tested using a bootstrap estimation approach with 2000 samples according to the procedure in Williams and MacKinnon (2008).

Results indicated that *Functional Consequences* were found to be a significant predictor of *Game Enjoyment*, $b = 0.791$, $SE = 0.044$, $p < 0.001$, (standardized regression coefficient 0.614), supporting hypothesis 1. Moreover, *Functional Consequences* were also found to be a significant predictor of *Psychosocial Consequences*, $b = 0.750$, $SE = 0.039$, $p < 0.001$ (standardized regression coefficient 0.637), supporting hypothesis 2. These results also support that a mediation analysis can be carried out. After controlling for the effect of the mediator *Psychosocial Consequences* on *Game Enjoyment*, the effect of *Functional Consequences* on *Game Enjoyment* was approximately halved, $b = 0.388$, $SE = 0.050$, $p < 0.001$ (standardized regression coefficient 0.301), but remained significant. The effect of *Psychosocial Consequences* on *Game enjoyment* was found at $b = 0.536$, $SE = 0.043$, $p < 0.001$ (standardized regression coefficient 0.490). The significance testing of the indirect effects was done using a bootstrap estimations. These results indicated the indirect effect were significant, $p < 0.001$ (see Fig. 3). Hence, these results suggest that a partial mediation is present, supporting hypothesis 4.

7. Discussion

The aim of this study was to develop and rigorously validate a scale that provides insight into how specific game design choices are experienced by players (*Functional Consequences*), and how these lead to specific emotional responses (*Psychosocial Consequences*). We contribute with an instrument that builds on Means-End Theory, and enables researchers and game developers to investigate player experience with a focus on functional and psychosocial aspects of play. The model underlying our work was refined by feedback from 64 Games User Research experts. Development and validation of the scale was carried out in five subsequent studies including 529 players. Discriminant and convergent validity of constructs was tested as well as configural and metrical invariance. Results show that the scale performs well over different sample sizes and studies (both delayed recall via paper-based surveys and immediate recall during play-testing approaches).

The PXI may be particularly useful for those researchers active in industry or game development. These games user researchers may wish to better understand how a diverse set of specific game design choices (i.e. the use of certain control schemes, the use of visual embellishments, the design of certain obstacle-levels) contribute (i.e. are associated with, mediate or moderate) psychosocial experience (mastery, immersion, curiosity...). To the best of the authors knowledge, there is no other scale in player experience research that separates between

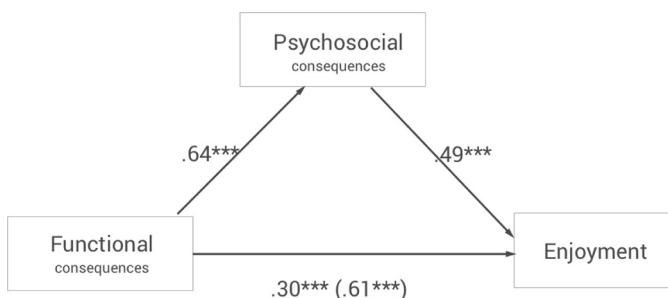


Fig. 3. Standardized regression coefficients for the relationship between Functional Consequences and Game Enjoyment as mediated by Psychosocial consequences. The standardized regression coefficient for the relationship between Functional Consequences and Game Enjoyment when controlling for Psychosocial consequences is in parenthesis. *** < 0.001 .

psychosocial and functional consequences at the construct level. For measuring this last category in particular, researchers are often left on their own, creating items and constructing themselves or relying on a combination of several other scales or methods. Use a combination of different scales may result in lengthier questionnaires, conceptually overlapping constructs and items, and possibly fatigue by participants. Hence, this may, in the end, lower overall scientific quality of the study, and be particularly problematic for the rapid playtesting cycles that typify the games industry. With the PXI, GUR experts now have one instrument to measure 10 constructs with 30 items, in a comprehensive yet parsimonious manner.

Leveraging Means-End Theory for Games Research The theoretical model underlying our work was informed by Means-End Theory (Gutman, 1982; Reynolds and Gutman, 1988), drawing from its previous application in user experience (UX) and player experience (PX) research. Here, we demonstrate that it can also serve as theoretical foundation for scale development; the development process and validation results of our scale suggest that the two different abstraction levels of consequences that have been found sufficient for most marketing analyses purposes are also useful for the analysis of player experiences. However, we need to acknowledge that the application of Means-End Theory for measuring player experience faces the same challenges as other fields in UX and PX, e.g. in situations where there is no active choice or no prolonged gaming experience, ME chains may be less articulated. Despite these limitations, our work does provide support for the MDA framework, and for models of play that take into account the game design elements to contribute to play experiences, opening up new perspectives for the development of empirically-grounded theories of play.

Measurement of Player Experience at the Functional and Psychosocial Level The measurement of functional and psychosocial aspects of play experiences also enables the games research community to derive new insights into what constitutes positive player experience: our work reveals that functional consequences affect overall enjoyment of the game, but this is partially mediated by psychosocial consequences. In this article, mediation has only been investigated at a global level. However, this scale offers the possibility of a finer-grained analysis, i.e., further path modeling at the construct level can be conducted. It allows GUR experts to investigate effects of game design choices on player experience, by means of building mediation models of how certain game choices affect in-game behaviors and emotional responses. For example, a game development studio might decide to redesign the actions to control a certain game character. GUR researchers could then investigate the scoring of *Ease-of-Control* as compared to the previous version of the game, and further investigate how this may effect feelings of *Mastery* or *Immersion*. The advantage of the Player Experience Inventory is that measurement of both levels can be done with one scale, at one time.

To further the research on player experience, this scale is free to use by other researchers. Moreover, data has been collected from 529 players already, evaluating certain games as part of certain game genres. In the spirit of open science, the dataset on which this scale has been built is available to other researchers and can be found at in the Data in Brief, accompanying this article. Hence, researchers can use this set to conduct further analysis on how certain games or genres score on different constructs, and how path-analysis can explain certain preferences.

8. Limitations

The work presented in this article needs to be interpreted in the light of a number of limitations. First, scale development and validation is an ongoing process. While different studies have been carried out with 529 participants already, across different continents and settings, we need to acknowledge that it was a predominantly young adult male audience. Hence, more studies are needed to assess how the PXI

performs across different game audiences. In this sense, we also need to acknowledge that we did not find metrical invariance across delayed and immediate recall data sets. We argue that this may be due to the different composition of game genre rather than the difference in salience of the game experience. However, this remains speculation and warrants further research. Second, to assess criterion validity we relied on scales (PENS Ryan et al., 2006, AttrakDiff2 Hassenzahl et al., 2003) that likely have influenced game experts when contributing to our selection of scales and constructs. A more robust criterion validation would include analysis of correlations with players' actual behaviors and utterances. We plan to explore this in future research. Finally, no claim can be made about the extent to which the constructs included at the level of *Functional* and *Psychosocial* consequences are exhaustive. It remains hard to define the exact nature of a game (Wittgenstein, 2009) and different genres and audiences further make it hard to draw exact boundaries (Juul, 2011) or delineate all relevant constructs to fully measure the player experience for a specific genre, audience and context. For this scale, we aspired to include a set of constructs that is generalizable across game genres and audiences. However, it is likely that for certain different game genres and audiences certain Functional and/or Psychosocial Consequences are not yet captured, e.g., constructs on relatedness and/or on narrative might be necessary to be included to capture different variations of *player experiences*. Hence, future research could focus on testing the scale with specific games, game audiences or methods, or on extending the scale with extra modules.

9. Conclusion

In this paper, we presented the conception, construction and validation of the Player Experience Inventory. Unique to the PXI is that it allows for GUR researchers to measure player experience at both the level of *Functional* and *Psychosocial* consequences. In this manner, the PXI aspires to provide actionable insight, enabling a better understanding of how game design choices impact the player actions during the runtime of the game, and how they shape emotional responses. The scale was devised on the basis of Means-End theory, but equally with the help of 64 GUR experts. The construction and validation of the measurement model was carried out over five studies, with 529 participants. Therefore, the scale is a reliable and valid tool in the toolbox of a GUR researcher.

Acknowledgements

We like to thank all GUR experts and participants contributing to this data set. A specific thanks goes out to Kieran Hickley and Gnter Wallner for contributing to the dataset, and to Penelope Sweetser and Peta Wyeth for sharing with us the list of scales to measure Player Experiences.

References

- Agarwal, R., Karahanna, E., 2000. Time flies when you're having fun: cognitive absorption and beliefs about information technology usage. *Mis. Q.* 665–694.
- Anderson, J.C., Gerbing, D.W., 1991. Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology* 76 (5), 732–740.
- Azadvar, A., Canossa, A., 2018. UPEQ: Ubisoft Perceived Experience Questionnaire: A Self-determination Evaluation Tool for Video Games. Proceedings of the 13th International Conference on the Foundations of Digital Games. ACM, New York, NY, USA, pp. 5:1–5:7. <https://doi.org/10.1145/3235765.3235780>. Event-place: Malmö, Sweden
- Bauchhage, C., Drachen, A., Sifa, R., 2015. Clustering game behavior data. *IEEE Trans. Comput. Intell. AI Games* 7 (3), 266–278. <https://doi.org/10.1109/TCAIG.2014.2376982>.
- Bracken, C.C., Skalski, P., 2010. Immersed in media: Telepresence in everyday life. Routledge. Google-Books-ID: w6GRagAAQBAJ
- Brockmyer, J.H., Fox, C.M., Curtiss, K.A., McBroom, E., Burkhart, K.M., Pidruzny, J.N., 2009. The development of the game engagement questionnaire: a measure of engagement in video game-playing. *J. Exp. Soc. Psychol.* 45 (4), 624–634. <https://doi.org/10.1016/j.jesp.2009.02.016>.

- Brown, E., Cairns, P., 2004. A Grounded Investigation of Game Immersion. CHI '04 Extended Abstracts on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 1297–1300. <https://doi.org/10.1145/985921.986048>.
- Bryce, J., Rutter, J., 2001. In the Game In the Flow: Presence in Public Computer Gaming. Poster presented at. IT University of Copenhagen.
- Byrne, B.M., 2016. Structural equation modeling with AMOS : basic concepts, applications, and programming, third edition. Routledge <https://doi.org/10.4324/9781315757421>.
- Cabrera-Nguyen, P., 2010. Author guidelines for reporting scale development and validation results in the journal of the society for social work and research. *J. Soc. Social Work Res.* 1 (2), 99–103.
- Carpenter, S., 2018. Ten steps in scale development and reporting: a guide for researchers. *Commun. Methods Meas.* 12 (1), 25–44. <https://doi.org/10.1080/19312458.2017.1396583>.
- Carter, M., Downs, J., Nansen, B., Harrop, M., Gibbs, M., 2014. Paradigms of Games Research in HCI: A Review of 10 Years of Research at CHI. Proceedings of the First ACM SIGCHI Annual Symposium on Computer-human Interaction in Play. ACM, New York, NY, USA, pp. 27–36. <https://doi.org/10.1145/2658537.2658708>.
- Celis, J., Husson, J., Abeele, V.V., Loyez, L., Van den Audenaeren, L., Ghesquié, P., Goeleven, A., Wouters, J., Geurts, L., 2013. Translating preschoolers' game experiences into design guidelines via a laddering study. Proceedings of the 12th International Conference on Interaction Design and Children. ACM, New York, NY, USA, pp. 147–156. <https://doi.org/10.1145/2485760.2485772>.
- Cheng, M.-T., She, H.-C., Annetta, L., 2015. Game immersion experience: its hierarchical structure and impact on game-based science learning. *J. Comput. Assisted Learn.* 31 (3), 232–253. <https://doi.org/10.1111/jcal.12066>.
- Cockton, G., 2004. Value-centred HCI. Proceedings of the third Nordic conference on Human-computer interaction. ACM, Tampere, Finland, pp. 149–160. <https://doi.org/10.1145/1028014.1028038>.
- Cockton, G., 2008. Designing worth—connecting preferred means to desired ends. *interactions* 15 (4), 54–57. <https://doi.org/10.1145/1374489.1374502>.
- Csikszentmihalyi, M., 1990. *Flow: The Psychology of Optimal Experience*. Harper and Row, New York.
- Csikszentmihalyi, M., Bennett, S., 1971. An exploratory model of play. *Am. Anthropol.* 73 (1), 45–58.
- De Grove, F., Cauberghe, V., Van Looy, J., 2014. In pursuit of play: toward a social cognitive understanding of determinants of digital play. *Commun. Theory* 24 (2), 205–223. <https://doi.org/10.1111/comt.12030>.
- Deci, E.L., Ryan, R.M., 1985. *Intrinsic Motivation and Self-determination in Human Behavior*. Springer.
- Denisova, A., Nordin, A.I., Cairns, P., 2016. The Convergence of Player Experience Questionnaires. Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play. ACM, New York, NY, USA, pp. 33–37. <https://doi.org/10.1145/2967934.2968095>.
- Desurvire, H., Wiberg, C., 2009. Game usability heuristics (PLAY) for evaluating and designing better games: the next iteration. *International Conference on Online Communities and Social Computing*. Springer, pp. 557–566.
- Deterding, S., Dixon, D., Khaled, R., Nacke, L., 2011. From game design elements to gamefulness: defining gamification. Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments. pp. 9–15.
- DeVellis, R.F., 2012. *Scale Development: Theory and Applications*. SAGE. Google-Books-ID: vmwBHYuchfAC
- Drachen, A., Canossa, A., 2009. Towards gameplay analysis via gameplay metrics. Proceedings of the 13th international MindTrek conference: Everyday life in the ubiquitous era. ACM, pp. 202–209.
- Drachen, A., Mirza-Babaei, P., Nacke, L.E., 2018. *Games User Research*. Oxford University Press. Hardcover
- El-Nasr, M.S., Drachen, A., Canossa, A., 2016. *Game Analytics*. Springer.
- Ermí, L., Mäyrä, F., 2005. Player-centred game design: experiences in using scenario study to inform mobile game design. *Game Stud.: Int. J. Comput. Game Res.* 5 (1).
- Fishbein, M., Ajzen, I., 1975. *Belief, attitude, intention and behavior: An introduction to theory and research*. Addison-Wesley Pub.
- Gaskin, J., Data screening.
- Greenberg, B.S., Sherry, J.L., Lachlan, K., Lucas, K., Holmstrom, A., 2008. Orientations to video games among gender and age groups. *Simulation Gaming*. <https://doi.org/10.1177/1046878108319930>. 1046878108319930
- Grove, F.D., Cauberghe, V., Looy, J.V., 2016. Development and validation of an instrument for measuring individual motives for playing digital games. *Media Psychol.* 19 (1), 101–125. <https://doi.org/10.1080/15213269.2014.902318>.
- Grunert, K.G., Bech-Larsen, T., 2005. Explaining choice option attractiveness by beliefs elicited by the laddering method. *J. Econ. Psychol.* 26 (2), 223–241. <https://doi.org/10.1016/j.joep.2004.04.002>.
- Gutman, J., 1978. Uncovering the distinctions people make versus the use of multi-attribute models: do a number of little truths make wisdom. Proceedings of the 20th Annual Conference of the Advertising Research Foundation. Advertising Research Foundation New York, pp. 71–76.
- Gutman, J., 1982. A means-end chain model based on consumer categorization processes. *J. Mark.* 46 (2), 60–72.
- Gutman, J., Reynolds, T.J., 1979. An investigation of the levels of cognitive abstraction utilized by consumers in product differentiation. *Attitude Res. Sun* 128–150.
- Hassenzahl, M., 2005. The Thing and I: Understanding the Relationship between User and Product. In: Blythe, M., Overbeeke, K., Monk, A., Wright, P. (Eds.), *Funology: from usability to enjoyment*. Kluwer Academic Publishers, Norwell, MA, USA, pp. 31–41.
- Hassenzahl, M., Burmester, M., Koller, F., 2003. AttrakDiff: Ein Fragebogen Zur Messung Wahrgenommener Hedonischer Und Pragmatischer Qualität. *Mensch & Computer* 2003. Springer, pp. 187–196.

- Hooper, D., Coughlan, J., Mullen, M., 2008. Structural equation modelling: guidelines for determining model fit. *Articles* 2.
- Hu, L.-t., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Multidiscip. J.* 6 (1), 1–55. <https://doi.org/10.1080/10705519909540118>.
2004. MDA: A formal approach to game design and game research. Vol. 4.
- Ijsselstein, W., de Kort, Y.A., Poels, K., Jurgelionis, A., Bellotti, V., 2007. Characterising and Measuring User Experiences in Digital Games. ACM Press, Salzburg, Austria.
- Ijsselstein, W. A., de Kort, Y. A., Poels, K., 2008. The game experience questionnaire: development of a self-report measure to assess the psychological impact of digital games. In press.
- Jennett, C., Cox, A.L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., Walton, A., 2008. Measuring and defining the experience of immersion in games. *Int. J. Hum.-Comput. Stud.* 66 (9), 641–661. <https://doi.org/10.1016/j.ijhcs.2008.04.004>.
- Johnson, D., Gardner, M.J., Perry, R., 2018. Validation of two game experience scales: the player experience of need satisfaction (PENS) and game experience questionnaire (GEQ). *Int. J. Hum. Comput. Stud.* 118, 38–46. <https://doi.org/10.1016/j.ijhcs.2018.05.003>.
- Juul, J., 2011. *Half-real: Video Games Between Real Rules and Fictional Worlds*. MIT press.
- Kardes, F., Cronley, M., Cline, T., 2010. *Consumer Behavior*, 1. South-Western College Pub.
- Korhonen, H., Koivisto, E.M.I., 2006. Playability Heuristics for Mobile Games. Proceedings of the 8th Conference on Human-computer Interaction with Mobile Devices and Services. ACM, New York, NY, USA, pp. 9–16. <https://doi.org/10.1145/1152215.1152218>.
- Kwak, M., Hornbæk, K., Markopoulos, P., Bruns Alonso, M., 2014. The Design Space of Shape-changing Interfaces: A Repertory Grid Study. Proceedings of the 2014 Conference on Designing Interactive Systems. ACM, New York, NY, USA, pp. 181–190. <https://doi.org/10.1145/2598510.2598573>.
- Law, E.L.-C., Brühlmann, F., Mekler, E.D., 2018. Systematic review and validation of the game experience questionnaire (geq) - implications for citation and reporting practice. Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play. ACM, New York, NY, USA, pp. 257–270. <https://doi.org/10.1145/3242671.3242683>.
- Lessiter, J., Freeman, J., Keogh, J., Davidoff, J., 2001. A cross-media presence questionnaire: the ITC sense of presence inventory. 10 (3), pp282–297.
- Lucas, K., Sherry, J.L., 2004. Sex differences in video game play: a communication-based explanation. *Commun. Res.* 31 (5), 499–523. <https://doi.org/10.1177/0093650204267930>.
- Lucero, A., Arrasvuori, J., 2010. PLEX Cards: A Source of Inspiration when Designing for Playfulness. Proceedings of the 3rd International Conference on Fun and Games. ACM, New York, NY, USA, pp. 28–37. <https://doi.org/10.1145/1823818.1823821>.
- Medlock, M.C., Wixon, D., Terrano, M., Romero, R., Fulton, B., 2002. Using the RITE method to improve products: a definition and a case study. *Usability Professionals Association* 51.
- Mekler, E.D., Bopp, J.A., Tuch, A.N., Opwis, K., 2014. A Systematic Review of Quantitative Studies on the Enjoyment of Digital Entertainment Games. Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 927–936. <https://doi.org/10.1145/2556288.2557078>.
- Mirza-Babaei, P., Nacke, L., Fitzpatrick, G., White, G., McAllister, G., Collins, N., 2012. Biometric Storyboards: Visualising Game User Research Data. CHI '12 Extended Abstracts on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 2315–2320. <https://doi.org/10.1145/2212776.2223795>.
- Nacke, L.E., 2013. An Introduction to Physiological Player Metrics for Evaluating Games. *Game Analytics*. Springer, London, pp. 585–619. https://doi.org/10.1007/978-1-4471-4769-5_26.
- Nacke, L.E., 2017. Games user research and gamification in human-computer interaction. *XRDS* 24 (1), 48–51. <https://doi.org/10.1145/3123748>.
- Nielsen, J., 2016. *Games User Research: Whats Different?* Nielsen Normal Group.
- Norman, K.L., 2013. GEQ (Game engagement/experience questionnaire): a review of two papers. *Interact. Comput.* 25 (4), 278–283. <https://doi.org/10.1093/iwc/iwt009>.
- Novak, T.P., Hoffman, D.L., Yung, Y.-F., 1997. *Modeling the Structure of the Flow Experience Among Web Users*, Boston, MA, USA.
- Olson, J., Reynolds, T.J., 2001. The Means-End Approach to Understanding Consumer Decision Making. *Understanding Consumer Decision Making: The Means-end Approach To Marketing and Advertising Strategy*. Psychology Press, Mahwah, NJ, pp. 3–20.
- Olson, J.C., Reynolds, T.J., 1983. Understanding Consumers' Cognitive Structures: Implications for Advertising Strategy. *Advertising and Consumer Psychology*, L. Percy and a. woodside. Lexington Books, Lexington, MA, pp. 77–90.
- User Experience | UX Design Tools | Optimal Workshop. <https://www.optimalworkshop.com/?gclid=>.
- Pagulayan, R., Keeker, K., Wixon, D., Romero, R.L., Fuller, T., 2003. User-centered Design in Games. *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications* (pp. 883–906): L. Erlbaum Associates Inc, pp. 217–235.
- Reynolds, T.J., Gutman, J., 1988. Laddering theory, method, analysis, and interpretation. *J. Advert. Res.* 28 (1), 11–31.
- Reynolds, T.J., Olson, J.C., 2001. *Understanding Consumer Decision Making: The Means-end Approach To Marketing and Advertising Strategy*. Psychology Press.
- Ryan, R.M., Rigby, C.S., Przybylski, A., 2006. The motivational pull of video games: a self-determination theory approach. *Motiv. Emot.* 30 (4), 344–360. <https://doi.org/10.1007/s11031-006-9051-8>.
- Salen, K., Zimmerman, E., 2003. *Rules of Play: Game Design Fundamentals*. MIT Press.
- Schell, J., 2014. *The Art of Game Design: A book of lenses*. AK Peters/CRC Press.
- Sherry, J.L., Lucas, K., Greenberg, B.S., Lachlan, K., 2006. Video Game uses and Gratifications as Predictors of use and Game Preference. *Playing video games: Motives, responses, and consequences* pp. 213–224.
- Sundström, P., Baumgartner, A., Beck, E., Döttlinger, C., Murer, M., Randelshofer, I., Wilfinger, D., Meschtscherjakov, A., Tscheligi, M., 2014. Gaming to sit safe: the restricted body as an integral part of gameplay. Proceedings of the 2014 conference on Designing interactive systems. ACM, pp. 715–724.
- Sweetser, P., Wyeth, P., 2005. GameFlow: a model for evaluating player enjoyment in games. *Comput. Entertain.* 3 (3). <https://doi.org/10.1145/1077246.1077253>. 3–3
- Tabachnick, B.G., Fidell, L.S., 2007. *Using multivariate statistics*, 5th ed. Allyn & Bacon/Pearson Education, Boston, MA.
- Takatalo, J., Häkkinen, J., Lehtonen, M., Kaistinen, J., Göte, N., 2010. Presence, Involvement and Flow in Digital Games. *Evaluating User Experience in Games*, 1st. Springer, London, pp. 277.
- Tamborini, R., Bowman, N.D., 2010. Presence in Video Games. *Immersed in media: Telepresence in everyday life* pp. 87–109.
- Tuunanen, T., Govindji, H., 2016. Understanding flow experience from users requirements. *Behav. Inf. Technol.* 35 (2), 134–150. <https://doi.org/10.1080/0144929X.2015.1015167>.
- Vanden Abeele, V., Gajadhar, B., De Schutter, B., 2009. Gaming Naturally is more Fun Together: The Influence of Controller Type on Player Experience. *Advances in Computing Entertainment/ Digital Interactive Media Entertainment and Arts*, Athens, Greece, October 29–31 2009.
- Vanden Abeele, V., Hauters, E., Zaman, B., 2012. Increasing the Reliability and validity of Quantitative Laddering Data with LadderUX. Extended abstracts on Human factors in computing systems. Austin, Texas, USA.
- Vanden Abeele, V., Zaman, B., 2009. Laddering the user experience!. *User Experience Methods, Interact 2009*, Uppsala, Sweden. Uppsala, Sweden.
- Vanden Abeele, V., Zaman, B., De Grooff, D., 2012. User experience laddering with preschoolers: unveiling attributes and benefits of cuddly toy interfaces. *Personal Ubiquitous Comput.* 16 (4), 451–465. <https://doi.org/10.1007/s00779-011-0408-y>.
- Vorderer, P., Klimmt, C., Ritterfeld, U., 2004. Enjoyment: at the heart of media entertainment. *Commun. Theory* 14 (4), 388–408. <https://doi.org/10.1111/j.1468-2885.2004.tb00321.x>.
- Wiemeyer, J., Nacke, L., Moser, C., Mueller, F., 2016. Player Experience. *Serious Games*. Springer, Cham, pp. 243–271. https://doi.org/10.1007/978-3-319-40612-1_9.
- Williams, J., MacKinnon, D.P., 2008. Resampling and distribution of the product methods for testing indirect effects in complex models. *Struct. Equ. Model.* 15 (1), 23–51.
- Wittgenstein, L., 2009. *Philosophical Investigations*. John Wiley & Sons.
- Wu, Y.-L., Chang, C.-W., Chang, C.-J., 2014. Exploring the Value of Cloud Services and QoE Factors on IPCentrex via the Means-End Chain Framework. Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia. ACM, New York, NY, USA, pp. 363–366. <https://doi.org/10.1145/2684103.2684173>.
- Zaman, B., 2008. Introducing contextual laddering to evaluate the likeability of games with children. *Cogn. Technol. Work* 10 (2), 107–117.
- Zillmann, D., 1988. Mood management through communication choices. *Am. Behav. Scientist* 31 (3), 327–340.
- Zillmann, D., 2015. Mood Management: Using Entertainment to the Full Advantage. *Communication, Social Cognition, and Affect (PLE: Emotion)*. <https://doi.org/10.4324/9781315743974-16>.