Validity Threats in Quantitative Data Collection with Games: A Narrative Survey

David Gundry and Sebastian Deterding

University of York

Author Note

This work was supported by the EPSRC Centre for Doctoral Training in Intelligent Games & Games Intelligence (IGGI) [EP/L015846/1] and the Digital Creativity Labs (digitalcreativity.ac.uk), jointly funded by EPSRC/ AHRC/Innovate UK under grant no. EP/M023265/1.

Correspondence concerning this article should be addressed to David Gundry, Department of Computer Science, University of York, UK, YO10 5GH. Contact: deg500@york.ac.uk

Abstract

Background. Games are increasingly used to collect **scientific data**. Some suggest that game features like high cognitive load may limit the inferences we can draw from such data, yet no systematic overview exists of potential validity threats of **game-based methods**.

Aim. We present a **narrative survey** of documented and potential threats to validity in using games for **quantitative data collection**.

Method. We combined an unsystematic bottom-up **literature review** with a systematic top-down application of standard validity threat typologies to games to arrive at a **systematisation** of game-characteristic **validity threats**.

Results. We identify three game characteristics that potentially impact validity: Games are **complex systems**, impeding the predictable control and isolation of treatments. They are rich in unwanted **variance** and diversity. And their **social framing** can differ from and interact with the framing of research studies or non-game situations they are supposed to represent. The diversity of gamers and their differences to general populations bring further complications.

Discussion and Conclusions. The wealth of potential validity threats in game-based research is met by a dearth of systematic methodological studies, leading us to outline several future **research directions.**

Keywords: data collection, game-based methods, research gamification, games with a purpose, applied games, validity, validity threats

Validity Threats in Quantitative Data Collection with Games: A Narrative Survey

We have seen a surge in the use of games to collect data for research questions outside games research itself, variously called *gamifying research (*Deterding et al., 2015) or *game-based methods* (Slegers et al., 2016). For instance, economists have long had to wrestle with the fact that they couldn't run true macro-economic experiments – neither would governments let them, nor could they really set up and compare two identical real-life economies. Hence, economists like Castronova, Williams, Ratan, and Keegan (2009) or Živić, Andjelković, Özden, Dekić, and Castronova (2017) have been exploring the use of virtual economies of massively multiplayer online games as economic petri dishes, based on the observation that the macroeconomic aggregates of virtual economies match real-world ones. Scholars in linguistics or ergonomics are adapting existing and creating new games as lab and online experiments (Hawkins, Rae, Nesbitt, & Brown, 2012; Oladimeji, Thimbleby, Curzon, Iacovides, & Cox, 2012). Researchers in e.g. psychology and epidemiology are re-purposing *game intelligence* – large-scale data from existing entertainment games – to answer basic research questions (Devlin et al., 2014; Williams, Contractor, Poole, Srivastava, & Cai, 2011). Finding close correlations between people's in-game performance and out-of-game traits like fluid intelligence (Kokkinakis, Cowling, Drachen, & Wade, 2017), they suggest that games can be used as alternative psychometric instruments. Qualitative researchers in human-computer interaction (HCI) and other fields are increasingly using board and card games to structure user and design research processes (Hannula & Harviainen, 2016; Slegers et al., 2016). So-called *citizen science games* are enlisting thousands of volunteer players to crowd-source scientific data collection and processing tasks like recording pollution levels, classifying images of galaxies, or identifying protein foldings (Cooper, 2015).

Data Collection with Games: Forms and Reasons

More systematically, one may distinguish *research games*, the use of full-fledged games for research purposes, from *research gamification*, the use of game design elements within research designs like surveys (Keusch & Zhang, 2017), and *game intelligence*, the opportunistic research use of the data exhaust of existing entertainment games (Devlin et al., 2014). Research games can be incorporated as part of an overall research design (e.g. Denisova & Cairns (2015)) or function as the total research setting (e.g. Zendle, Cairns, & Kudenko (2015)). They might provide an experimental treatment (e.g. Johnstone (1996)), a measurement instrument (e.g. Foroughi, Serraino, Parasuraman, & Boehm-Davis (2016)), or both. Data collection may be internal to the game, for example by using game telemetry, or external to the game, for example through video-taping game events or pre/post-game interviews or questionnaires. Importantly, research games differ in how much they are controlled by the researcher: researchers can simply use existing entertainment games (McMahan, Ragan, Leal, Beaton, & Bowman, 2011), which may provide ready access to large, ecological valid data sets, particularly in the case of online games. Designing bespoke games in contrast affords fine-grained control over the game (e.g. Zendle et al. (2015)). Finally, creating modified ("modded") versions of existing games presents a middle ground between design control and ecological validity (Elson & Quandt, 2016).

Looking across these varied instances, one can make out four main reasons games are used for scientific data collection. The first is *motivation*: Good games are enjoyable and intrinsically motivating. Hence, turning an experiment or survey into a game may motivate people to participate voluntarily (reducing recruitment costs), and motivate them to stick through to the end, reducing churn and invalid data points. Second, networked online games provide *potential large-scale population and data access* – the global population of digital game players is estimated to be 2.6 billion (McDonald, 2017). In this, they are not fundamentally different from other online research paradigms. However, third, popular entertainment online games like WORLD OF WARCRAFT (Blizzard Entertainment, 2004) or gamified crowdsourcing platforms like ZOONIVERSE (Citizen Science Alliance, 2009) provide *actual opportunistic access* to large pre-assembled, willing audiences. Fourth and finally, digital games allow *perfect control and tracking*: By virtue of being already fully-digital, *every* aspect of a game environment can be deliberately held constant or manipulated, and *every* player action is already captured at the level of individual keystrokes and mouse movements. And with the rise of pervasive games, ubiquitous sensing, and natural and immersive interfaces like motion control and augmented reality, digital games are increasingly able to capture literally every movement, anywhere, any time.

Validity of Data Collection with Games

Whenever data is collected, the question of *validity* arises: To what extent does the data support the inferences we draw from it? Together with reliability, validity is a key construct and quality criterion of scientific research, with a long history in quantitative research (Jenkins, 1946), but also in wide use in contemporary qualitative research (Morse, Barrett, Mayan, Olson, & Spiers, 2008; Nahid Golafshani, 2003). Following Messick's popular conceptualisation, "[v]alidity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (Messick, 1995, p. 471). Beginning with Campbell (1957), researchers have developed multiple typologies of validity and validity threats, usually structured around aspects of claimed causal relations between the treatment and outcome of an experimental study. In a classic typology, Shadish, Cook, and Campbell (2002, pp. 38, 42–93) distinguish four kinds of validity:

- *Statistical Conclusion Validity*: Do the presumed cause and effect covary and if so, how strongly? Common threats to this validity types are statistical issues like low statistical power, inappropriate statistical methods (e.g. because the data violates underlying assumptions like normal distribution), or uncorrected fishing for results.
- *Internal Validity*: Is the observed covariation between treatment and outcome actually due to their causal link or other factors in the research design? This connects to validity threats like confounds or selection bias. For instance, an observed negative correlation between age and gameplay performance may not be due to age, but due to a third factor (= confound) that covaries with age, such as gaming socialisation. Similarly, if gaming socialisation affects gameplay performance and researchers want to study the effect of alcohol consumption on gameplay performance, they need to ensure during sampling that treatment and control group don't have very different gameplay socialisations (= selection bias).
- *Construct Validity*: Do the particular samples, treatments, and outcome measures used in the experiment accurately operationalise the constructs that are studied? Many threats to validity in this category have to do with what researchers call reactivity (Shadish et al., 2002, p. 77). Participants actively make sense of and respond to the study they participate in: they may answer what they think the experimenter wants to hear, cheat to get a high score on an assessment, or are reminded by the study of stereotypes regarding their own aptitude in the activity studied, which may induce self-doubts and anxieties dampening performance.
- *External Validity*: Do observed correlations hold across other people, settings, treatments, and measures? In other words, do the study results generalize? Part of this is the question of so-called ecological validity: to what extent does the experimental situation reflect actual situations as people would experience them in their day-to-day life? For example, that people choose one game over another in a lab experiment where they only have two games to choose from may tell us little about how people actually choose games in their everyday life, where they are faced with hundreds of thousands of games in an app store.

When it comes to games research, various validity threats have been mentioned: the use of different games as experimental conditions that vary in more than the targeted aspect (Ferguson, 2015); failing to recognise differences in gaming expertise linked to different game genres (Latham, Patston, & Tippett, 2013); or the high cognitive load of games, making it difficult for players to play a game in a natural way *and* report on their gameplay experience, as standard think-aloud methods in HCI ask for (Hoonhout, 2008). However, these and similar observations have remained scattered and piecemeal, and chiefly concern entertainment and educational games research, not the use of games for other research questions (see e.g. Louvel (2018) for a survey of ecological validity issues in lab-based games user research and S.

P. Smith, Blackmore, and Nesbitt (2015) for a meta-analysis of data collection forms in serious games). In fact, while authors like Williams (2010) and Deterding (2016) have called for systematic research programmes on the correlation between people's in-game and out-of-game behaviour, there has been little integrative, let alone systematic work in this area – one recent exception being a narrative review of issues in gamified surveys by Keusch and Zhang (2017).

Structure of this Article

To move the field forward, this article surveys known validity threats in the use of games for data collection, as well as highlighting potential, as-of-yet unstudied threats where there is a compelling argument for them. In interest of space, we constrain the discussion to well-established validity issues of quantitative research that are particularly pertinent to games. We organise our discussion along three key features of games that appear to underlie most validity issues we identified, namely *systemic complexity*, *variance*, and *framing*. First, the complex systemic nature of games and gameplay make it hard to isolate and manipulate individual game elements without inadvertently affecting other properties as well, leading to confounds threatening *internal validity*. Second, games and gameplay are high in diversity and uncontrolled variance. This means that given the same effect size, larger samples are needed to make valid inferences about true effects, and increases the likelihood of detecting false positive effects (type 1 errors) as well as failing to detect true effects (type 2 error), impacting *statistical conclusion validity*. Furthermore, high variance threatens *construct validity*, as it becomes hard to hold everything but the operationalisation of the construct in question constant. Gameplay is a very particular kind of social situation of frame that differs both from experimental setups and other types of social situations we wish to make inferences about. This threatens the generalisability of findings, and with it, *external validity*. A fourth and final section discusses *player-related* validity threats. Despite the mainstreaming of video gaming, player demographics of particular games still can deviate from the general population, likewise impeding *external validity* of data collected via games. Interpersonal differences in play styles and genre expertise may confound results, while turning an activity into a game may lead players to deviate in strategic ways from how they would spontaneously behave in a non-game version of the same activity (i.e. cheat or "game the system"), with negative ramifications for *internal, external, and construct validity*. Figure 1 provides a schematic overview of game characteristics and validity implications.

Characteristic	**Validity Implication**	**Validity Threat**			
		Statistical	Internal	Construct	External
Systemic Complexity					
Games are rich, complex stimuli	Cognitive load and induced arousal can interact with measurement or confound		■		
Games and gameplay are complex systems	Manipulation may have unexpected emergent effects		■		
Games are novelty-based and learned	Learning effects over repeat measures		■		
Variance					
Games are divergent	Different games as treatment/control can differ on more than desired dimension		■	■	
	Data from one game may not replicate in others				■
Game setups are divergent	Measurement errors and confounds when different participants use different setups		■	■	
Game content is varied	Uncontrolled, non-random variance in stimuli and conditions	■	■	■	
Gameplay is emergent and varied		■	■	■	
Commercial games are not fixed		■	■	■	
Framing					
Play frame may differ from target situation	Gameplay behaviour may not generalise				■
	Gameplay may motivate dishonest responding		■		■
Research studies may differ from play	Forced gameplay may turn game preferences into a confound		■		
	Knowing alternative game conditions may produce resentful demoralisation or compensatory rivalry			■	
Players					
Gamers are not the general population	Games may attract a biased sample				■
	Games may differentially produce stereotype threat, evaluation apprehension, and social desirability in (non)gamers		■	■	
Gamers are diverse	Chosen game genre and platform may attract a biased sample				■
	Difference in player types or genre expertise may confound results		■		

Figure 1: Overview of game characteristics. Shaded cells highlight relevant validity implications

We arrived at this systematisation through a combined bottom-up and top-down approach: Bottom-up, we conducted an opportunistic literature search across major relevant databases (Web of Science, Scopus, ACM Digital Library, Google Scholar), searching for "validity" and "games", resulting in 10 relevant papers (see References). Top-down, we used the typology by Shadish et al. (2002) to systematically ask for each validity threat they classify where and how it may manifest around digital games. Clustering reported and hypothesised validity threats, we arrived at a smaller subset of threats that we then tried to variously organise, arriving at three high-level characteristics of games that appeared responsible for them. In response to reviewer suggestions, we factored out player-related threats into a separate fourth category. We expressly view our proposed systematisation as a first draft.

For each of the four groups, we will first introduce underlying characteristics and then report observed and potential validity threats. In the discussion and conclusion, we will draw some overarching observations on the opportunities and challenges of using games and game design in data collection and outline areas for future research.

Systemic Complexity

Many even simple games form *complex systems* (Salen & Zimmerman, 2004), meaning that they are made of a network of many constituent parts which interrelate, mutually depend, and interact in ways that are hard to analyse, model, manipulate, or predict in isolation (Auyang, 1999). This manifests particularly in *gameplay* and *player experience*, the process in which players interact with a game and the experiences they have of this process (Hunicke, LeBlanc, & Zubek, 2004). Both strongly *emerge* from the interaction of game system and players in ways that are *nonlinear* and *time-dependent*. Changing even a small parameter of a single game mechanic, such as drawing not one but two cards in a card game, can make one in-game action more powerful than another. This in turn can change what winning strategies are, how long the game takes to play, how satisfying its challenges are, or what kinds of information need to be communicated between the players. Notably, the emergent properties of particular design changes or elements do not necessarily hold across games. For instance, adding a time constraint to chess matches turn chess into *speed chess*, a game with such recognised differences in gameplay and player experience that it warrants its own name. In contrast, adding a match time limit to the real-time game QUAKE (id Software, 1996) changes gameplay and player experience only minimally. Game development is therefore highly iterative, continually building and playtesting design changes to assess and tune their actual emergent effects (Hunicke et al., 2004).

What this means for researchers is that isolating and manipulating constituent parts – let alone psychologically "active ingredients" (Michie & Johnston, 2013, p. 469) – of a game is inherently challenging. But without isolating features we risk confounds in our experimental manipulations – third variables that provide a competing explanation for our results. Confounds threaten internal validity, meaning that we may not be able to justify that the treatment has observed an effect.

Within this section, we identify and address three characteristics of games that are involved in their complex systemic constitution:

1. Games are rich, complex stimuli
2. Games and gameplay are complex systems
3. Games are novelty-based and learned

Games are Rich, Complex Stimuli

Stimulus complexity *per se* is not unique to games: traditional experimental materials may be more or less complex, ranging from the relative sparseness and uniformity of e.g. standardised paper questionnaires to the richness of social psychological experiments like the famous bystander study in which confederates role-played a seizure during a discussion to assess whether the presence of other bystanders affected participants' responses (Darley & Latané, 1968). Compared with other typical media psychological materials like text, imagery, music, or film, games sit on the high end of stimulus complexity and richness, combining the above into a total art work. Apart from the resulting stimulus variance across and within games (see below, *Variance*), this introduces significant potential confounds of its own, namely arousal and cognitive load. Gameplay often induces arousal (Anderson & Bushman, 2001), which is a well-known potential confound particularly in self-report studies, leading to e.g. selective attention (Pham, 1996). Gameplay is also often immersive and engaging, which increases cognitive load (Schrader & Bastiaens, 2012). Extraneous cognitive load in turn is known in games-based learning to impede learning (Kiili, 2005, p. 21). The cognitive load of educational games can interfere with their pedagogical effectiveness, and confound studies that compare game and non-game conditions without controlling for cognitive load (Wouters, Nimwegen, Oostendorp, & van Der Spek, 2013). Similarly, the high combined cognitive load of gameplay and certain data collection methods like think-aloud (Hoonhout, 2008) may overload participants, such that measurement interacts with and potentially confounds gameplay and player experience.

Games and Gameplay are Complex Systems

The systemic interrelation and interaction of elements of games makes it is hard to change one aspect of a game in isolation (Kim & Shute, 2015). In psychological parlance (Littman & Rosen, 1950), changes on the *molecular* level of individual game elements or player actions tend to play out as wholesale changes on the *molar* level of the whole game or gameplay – often in emergent, nonlinear ways. Conversely, the same molar *gestalt* – a particular player experience, gaming strategy, or gameplay dynamic – may be realised through many divergent molecular game features and player actions. Communication research has developed some theoretical paradigms that acknowledge such systemic dynamics (Früh & Schönbach, 2005), and some theoretical models do describe games, gameplay and player experience on different levels of organisation (e.g. Klimmt (2006)). Still, much design research and guidance in applied gaming for data collection and other purposes focuses on individual, molecular game features or elements (Deterding, 2015), and the complex systemic constitution of games runs counter to the *de facto* linear, reductionist assumptions embodied in standard experimental research designs.

For game-based data collection, we see two particular ramifications. First, any manipulation (e.g. imported from a standard non-game study design) may generate unforeseen interactions and emergent dynamics in addition to the causal effects it is hypothesised to produce. Therefore, researchers should prototype and pretest manipulations to check for these confounds. This means that researchers need a clear idea of what confounds to look out for. For example, Carnagey and Anderson (2005) used two modes of the same racing game to manipulate whether violence is rewarded in the game. In one condition, players were rewarded for killing pedestrians and race opponents. In the other, they were prevented from doing so. While most differences between the conditions were controlled by using the same game, the difference in used game mode still arguably had an unintended knock-on effect on competitiveness (Adachi & Willoughby, 2011). Concretely, there were different numbers of things to compete on in each condition. The condition where players were rewarded for destroying race opponents had two sources of competition: winning the race and surviving the free-for-all. This second source of competition was not present in the control condition where destroying race opponents was prevented. To us, this suggests that researchers should incorporate game design expertise in pretests, as these kinds of unexpected dynamics are likely more readily apparent to experienced game designers.

Relatedly, the development process itself has a potential significant impact on the conditions designed. In games, it is typical (and practical) to develop a first part of the game (“first playable”, “vertical slice”) in high detail to establish the game’s core mechanics and gameplay. All subsequently developed parts or levels are effectively variations and extensions on this core gameplay. Hence, designers are constrained in the development of subsequent content (or conditions) in a way that they are not in the first, and there will often be a general difference in quality, be that in terms of fun or balance, between initial and subsequent conditions developed in this way. As McMahan et al. (2011, p. 4) assert for experimental designs in general: “in our experience, researchers may spend more time or effort implementing a condition that they subconsciously (or consciously) favor, biasing the study toward that condition.”

Games are Novelty-based and Learned

Novelty is the property of an experience being new (Silvia, 2006). Simply put, the first experience of a stimulus or measurement instrument is different from subsequent encounters. In situations where novelty features strongly, the validity threat of *learning effects* arises: participants perform differently when they had past experience with a stimulus or measurement instrument (Shadish et al., 2002).

This relates to two time-related characteristics of games. First, interest and curiosity are two major intrinsic motives and sources of enjoyment in gameplay. Games feature properties like dramatic conflicts, novel content, puzzles, or randomness to afford uncertainty that draws attention and is satisfying to resolve (Costikyan, 2013). If a part of a game entails largely the same outcomes and experience on replay, less uncertainty, curiosity, and interest are likely to arise. This is of particular relevance to games focusing on narrative: after a first play-through, the novelty and dramatic tension of their plot is largely exhausted (Roth, Vermeulen, Vorderer, & Klimmt, 2012). As a result, if players play the same section of a game twice – once as a treatment and once as a control – the diminishing uncertainty, curiosity, and interest may become major learning effects.

A second time-related game characteristic is that players over time learn to play them well. In fact, a large part of their enjoyment arises from the competence experience of learning to master the game (Deterding, 2015; Koster, 2005). Most games are therefore designed with a careful scaffolding of required and taught skills and knowledge, increasing difficulty in lock-step with growing player skill (Chen, 2007). If a player returns to replay earlier game sections they've already mastered, they are likely to find it easy to overcome its challenges, and will thus likely experience mastery or learning – again, a strong possible learning effect. Another potential learning-related confound is the difference between learning a new skill and performing a mastered skill, which can express itself in e.g. error rates, time taken, exploratory versus goal-oriented behaviour, and the like. Players' game knowledge may even threaten construct validity. For example, if multiple play-throughs allow players to memorise puzzle solutions rather than solving puzzles anew, game performance may be indicative of short term memory rather than problem-solving abilities. The amount of time participants have to learn a game before engaging with the game section that constitutes the experimental manipulation/control may also confound results. Too little time and players lack of skill may prevent them from effectively completing the experiment. Too much time may lead to ceiling effects or converging upon optimal strategies. Finally, which section of the overall sequence of a game players play may significantly impact what level of difficulty and required skills they encounter. These strong potential learning effects and other confounds are of particular concern in within-subject, repeated measure designs where the same subject is presented with control and experimental condition in sequence. Apart from choosing different study designs, one common mitigation is to vary the sequence of control and manipulation conditions as part of randomisation. A second mitigation strategy would be to use techniques like adaptive procedural content generation and pre-testing to ensure that game content in each condition is equally novel and difficult.

Variance

Next to complex systems, another popular way of framing games is as a possibility space (Squire, 2008). Games open a space or tree of possible states and partly relinquish control over in-game events to extraneous influences like player choice or randomness. Control, however, is the *sine qua non* of experimental design. It ensures that all participants experience the same manipulated or control condition as reported – essential for construct validity. It also minimises unwanted variance between conditions. Where such unwanted variance is randomly distributed, it statistically obscures true effects, requiring the use of larger samples. Where it is non-random, it becomes a confounding variable. Research games are thus faced with somewhat conflicting requirements to be both a controlled research tool and a game offering possibility spaces. This dilemma manifests itself around at least five different forms of unwanted variance:

1. Games are divergent
2. Game setups are divergent
3. Game content is varied
4. Gameplay is varied and emergent
5. Commercial games are not fixed

Games are Divergent

Games are a highly diverse medium, with different interfaces and controls (e.g. desktop monitor plus mouse and keyboard versus mobile touch screen versus motion control plus VR headset), different social contexts and configurations (e.g. public competitive Esports play versus private cooperative or competitive multiplayer gaming versus solitary play), and different genres (e.g. open world exploration, casual puzzler, idle game, RPG, first person shooter), each affording different demands and experiences. No single game can therefore be taken as representative of all games. The selection of a particular game, including its interface, controls, social context and configuration potentially threatens internal validity if different games (or game configurations) are used for different experimental conditions, and impinges on external validity in terms of how well or widely any findings generalise. Indeed, the use of different games to operationalise different experimental conditions has been cited as a critical internal validity issue in the violence in video games literature (Adachi & Willoughby, 2011; Ferguson, 2015). Because the games used differ on more dimensions than just violent content, these dimensions present potentially confounding variables (Elson et al., 2013).

In response, some scholars have suggested ways to match games on certain criteria (Adachi & Willoughby, 2011), such that key features considered relevant to the investigation vary only in desired respects. For example, research on violence in video games has adopted this approach to match violent versus non-violent treatment and control games on competitiveness (Adachi & Willoughby, 2011), difficulty of controls (Przybylski, Rigby, & Ryan, 2010), and frustration (Przybylski, Rigby, Deci, & Ryan, 2014). Two difficulties arise with this matching strategy. Firstly, games may not successfully be matched on the given factor. Secondly, games may remain divergent on unmatched factors that reveal themselves to pose confounds. For instance, Anderson and Dill (2000) selected the two games WOLFENSTEIN 3D (id Software, 1992) and MYST (Cyan, 1993) as violent treatment and nonviolent control because they matched for "blood pressure, heart rate, frustration, difficulty, action pace, enjoyment and excitement" (Adachi & Willoughby, 2011, p.58). Yet in this, Anderson and colleagues missed that the games also differed in competitiveness, which proved to be a significant confounding variable (Adachi & Willoughby, 2011).

Another approach is to adapt a single game to provide treatment and control conditions, a so-called modified game paradigm (Hilgard, Engelhardt, & Rouder, 2017). This adaptation can often be achieved through modding an existing game (e.g. Elson and Quandt, (2016), Engelhardt, Hilgard, and Bartholow (2015), Mohseni, Liebold, and Pietschmann (2015)), or developing bespoke games (e.g. Zendle et al. (2015)). This allows researchers to control the experimental conditions far better, and avoids the potential confounds of different games. The challenge, as explained earlier, remains that a small manipulation within a game can have unforeseen and undesired emergent systemic effects on gameplay and player experience.

Game Setups are Divergent

Related to the diversity of games, the way digital games are technically delivered and instrumented can vary greatly. This leads to potential measurement errors or confounds, as the means for collecting, transmitting, and recording data are subject to error, interference, or unwanted variance. This issue is particularly acute in remote/online designs, where researchers have less control over gaming hardware, controls, and networks. Variance in participants' computers, controls, or networking bandwidth can cause issues with tasks and measures such as those involving reaction times (Hilbig, 2016; Reimers & Stewart, 2007).

Game Content is Varied

Even within a single chosen game, the interactivity of games means that the actual content (levels, puzzles, rewards, challenges) players experience will vary between players and game sessions, often significantly and to a not fully predictable extent. This threatens construct validity, as it may be hard to ensure that or discern whether players experienced the desired stimulus, and to ensure that or discern whether they experienced other, undesired variance in stimuli. If game performance is also used as a measurement instrument, such as in educational assessment, chance variation may overwhelm the meaningful information it contains. On the structural level of a game's design, there are at least three sources of emergent variance.

Games Provide Player Choice. Games relinquish significant control to the player (Klimmt, Vorderer, & Ritterfeld, 2007), supporting choice in what character they embody, what goals they pursue, what strategies they use and what actions they take. Not only is such meaningful choice directly fuelling engaging and enjoyable autonomy experiences (Deterding, 2016): it can regularly lead different players to perform different actions and experience different outcomes. Within an experiment, in contrast, tasks performed are usually strictly controlled, and undesired variance in outcome minimised. While almost all games offer some degree of player choice, the amount of choice offered differs markedly between game genres and games. Where some *rail-road* players along the same trajectory, other open *sand-boxes* with a wide range of possible goals and actions.

Games Often Include Random Events. Particularly so-called games of chance relinquish control over key game events to randomness. In other games, randomness is incorporated in the design through procedurally generated content to afford replayability: a random seed is used to e.g. generate a different game map every time. Where such events are *truly* random, they merely require a large enough sample to detect true and reject false covariance. However, they often are *pseudo*-random and thus potential confounds, e.g. using pseudo-random seeds or biasing randomness in desired directions. The dropping of in-game relevant rewards ("loot") in role-playing games for instance is known to have carefully crafted chances to optimise player engagement and not negatively affect the in-game economy.

The starting situation can vary. The starting situation of a game can be fixed or variable. For instance, many games allow players to customise their characters before start, configure controls, or set the game difficulty. These configuration options are one of the easier variables to control. For instance, a save game or starting setup can be prepared and loaded to ensure consistency across participants. The important thing is to account for this potential variance, especially in remote designs. However, some game details change separately and automatically, for instance the game-wide unlocking of new items, levels, or achievements, which affects the possibility space of all subsequent players.

Gameplay is Emergent and Varied

By relinquishing control over the game state to player agency, games open the systematic possibility that players act differently with every game session. Game theoretically, one can model this possibility space of actions as a decision tree (Elias, Garfield, & Gutschera, 2012). If players were fully informed and rational actors in rational choice terms, solely motivated to win the game, one could calculate and predict the strategically optimal move, and such game theoretic calculations are indeed a common tool among game designers (J. H. Smith, 2006). However, especially in games with two or more interdependent actors (human or artificial), possible choices and game states quickly compound to a point where calculating the optimal move becomes humanly impossible: three pairs of turns into chess, there are 121 million possible game states, for instance. Nevertheless, game sessions display higher-level dynamics and player communities and expert players evolve higher-level strategies and heuristics to reduce this complexity, which again interact and change in hard to fully predict ways (Elias et al., 2012). In active player communities around games like LEAGUE OF LEGENDS (Riot Games, 2009), for instance, shared views about optimal high-level strategies like character choice ("the meta") are in constant flux. Complicating the picture further, player actions are regularly shaped by more concerns than mere winning (Gundry & Deterding, 2018). Overall, this means that especially in interdependent multiplayer games and other games with so-called emergent gameplay (sic), gameplay actions and experiences are hard to control and predict on a low level and showcase emergent but again not fully predictable nor controllable patterns on a higher level of organisation.

Commercial Games are Not Fixed

Current trends in game development mean that even an individual game's variance in content is not necessarily stable, but may change between sessions and over time.

Game Updates and A/B Tests. The widespread availability of high-speed internet connections has enabled a new development and business model usually called *games as a service*, where games are increasingly provided as a continuing online service. As a result, there is often no single, canonical version of a game that holds constant across studies. Rather, continuous game changes have become commonplace, including patches and bug fixes, downloadable content (DLC), seasonal in-game events, and more. Even worse, developers now make frequent use of A/B or even multivariate testing, serving different players different versions of the same game in parallel to assess which works better. Researchers working with existing entertainment games, especially online ones, therefore run the risk of unintentionally or even unknowingly serving participants different game versions at different points, a so-called history effect threatening internal validity.

This risk can be easily avoided in games purpose-made for research. When using existing entertainment games, researchers should decide on a canonical version of the game for the purposes of the study wherever possible, document the version of the game used, and ideally make a copy of it available for future researchers interested in replication. In some cases, a canonical version may be safeguarded by downloading a local copy, disconnecting the game from the internet, or disabling updates. Where a stable version cannot be ensured and a game is updated during an experiment, the researcher should consider and report any potential impact this may have had.

Adaptation and Content Generation. Many games tailor the experience to the individual player. Single-player games commonly use techniques like dynamic difficulty adjustment (Hunicke, 2005) to give players a satisfying experience by adjusting difficulty based on their past gameplay performance, or even procedurally generate whole levels to keep content novel for players (Shaker, Togelius, & Nelson, 2016). While these systems aim to players with an overall evenly enjoyable experience, they reduce control and predictability of actual moment-to-moment game content for researchers. Whether or not to choose a game with dynamic difficulty adjustment or procedural content generation thus becomes an important research design consideration: if an even higher-level player experience is desired, such games may be the best option (but need pretesting). If low-level control is needed, they are to be avoided.

Multiplayer Games. Other players in multiplayer games are sources of variation. For example, one participant may face an easy opponent, while another faces an experienced opponent. While competitive multiplayer games use ranking and matchmaking systems to provide players with an overall even, fair experience (Sarkar, Williams, Deterding, & Cooper, 2017), individual match experiences still differ in many respects. Similarly, online player communities may differ between servers and change over time in their size, activity, demographics, norms, and practices (Bartle, 2004). Thus two players of the same multiplayer game may have different experiences based on when they played and who they played with. This variance of multiplayer games can be somewhat controlled by using confederates or bots. However, this may in turn produce history effects (when confederates tire out over multiple plays) or threaten ecological validity, as scripted human or bot play may differ from spontaneous play. Where multiple study participants play together, it may be appropriate to use a Group Randomised Trial design to adjust for intraclass correlation (Murray, 1998).

Framing

People's everyday life is organised into different kinds or types of social situations that each come with particular roles, norms, and expectations: going to the movies, shopping at a store, giving a lecture, etc. During socialisation, people learn what kinds of situations exist in their society, and how to understand and act appropriately within them. Sociologist Erving Goffman (1986) first extensively studied these kinds of situations, calling them *frames*. Whenever an activity is transplanted from its naturalistic situation into a different frame, like an experiment or game, this different framing may have a significant effect on people's experience and behaviour. In psychology, the class of validity threats called *demand characteristics* refers to exactly this "totality of cues and mutual role expectations that inhere in a social context, (e.g., a psychological experiment or therapy situation), which serve to influence the behavior and/or self-reported experiences of the research participant or patient" (Orne & Whitehouse, 2000, p. 182). For instance, by taking on the role of a good participant in the frame of an experiment, study participants may act in ways they hope help the researcher accomplish their study goal, rather than how they would spontaneously act in a different situation. Conversely, games research has highlighted that many kinds of behaviours that would be inappropriate in everyday interaction become acceptable or even desired in a gaming frame, such as aggressively competitive and strategic behaviour, bluffing, or teasing and taunting (Deterding, 2014). These frame differences may threaten *construct validity* when demand characteristics become unwittingly part of the treatment. They can threaten *external validity* in that behaviours and experiences occurring during play may not hold outside of the play frame.

Play May Differ from the Target Situation

One oft-mentioned feature of digital games is that they allow to *simulate* parts of the real world with great verisimilitude, especially when employing contemporary immersive technologies like virtual reality. Yet no matter how realistic the simulation, *playing at* an activity is always socio-materially different from performing the activity without a play frame: social and material consequences are usually muted (the virtual lion doesn't really bite, the as if-breakup is not a real breakup); and norms and expectations for behaviour framed as play differ from those for behaviour framed as earnest (Deterding, 2014). Thus, simply *calling* an activity a game can already change the experience and observed behaviour (Lieberoth, 2015). This raises the question of mapping: for what kinds of behaviours and contexts does in-game behaviour correlate with behaviour in the real world? (Deterding, 2016) For instance, while aggregate economic behaviour in online game auction houses mirrors economic behaviour in real-world auction houses, communication norms in online game chat markedly differs from those of everyday face-to-face conversations (Deterding, 2016). While scholars like Williams (2010) raised this as a basic meta-methodological question of games-based research, eight years on, there is still little if any work on this issue. Instead, we here want to highlight only two obvious differences as starting points for future work.

Games Mute Socio-Material Consequences. As mentioned, in-game events usually have lowered practical and symbolic consequences compared to their non-play-framed counterparts. While gambling and the rise of real-money trading and microtransactions around in-game items provide plenty counterexamples, in many games, there is no bodily or economic risk involved in game outcomes. Players may therefore be more risk-taking in games than they would be in the real world. There is rich related debate in economics on how much participants need to be paid and what real-world payout consequences there need to be for participant decisions during a study for the study to count as ecologically valid

(Camerer, Hogarth, Budescu, & Eckel, 1999). Several studies on economic games find differences in player choice when monetary incentives are added (Schlenker & Bonoma, 1978).

Games Invite Strategic Action. Games are one of the few social contexts in which ruthlessly rational, strategic, self-interested action is allowed and even desired: a player who doesn't try hard to calculate and take optimal moves in order to win would be considered a spoilsport (Deterding, 2014). This norm of gameworthiness is counterbalanced with norms of playworthiness – having fun together – which may result in suboptimal behaviour like self-handicapping. However, different game contexts and genres come with different norms how ruthlessly one is allowed and expected to play (Deterding, 2014), and these norms may differ from the situational norms of the activity of context that one wishes to collect data on. For instance, if a game is designed to elicit people's preferences about different flavours of ice cream, and there is a strategic in-game advantage to answer "chocolate" even if one *actually* prefers strawberry flavour, the game's design will confound the responses (see Gundry & Deterding (2018) for a detailed discussion and design guidelines to mitigate these effects). The focus on winning the game may also override participants' desire to be a good study subject and lead them to cheat. In games, cheating is the use of mechanisms unintended or forbidden in the game to achieve in-game advantage (Consalvo, 2007). In experimental terms, cheating occurs when participants use means that were not intended to complete an experimental task. For instance, in an online game eliciting players' hand-eye coordination speed, very motivated players may change their screen contrast or use different controllers like an auto-fire mouse to score higher than they would under standard conditions.

Research Studies May Differ from Play

Like play, research studies also constitute a social frame with norms and expectations of their own – what psychology calls demand characteristics (Orne & Whitehouse, 2000). These pose game-characteristic validity threats where they interact or clash with the norms and expectations of gameplay now being re-framed as a research study.

Experiments May Force Gameplay Against Player Preference. First, in leisurely gameplay, players expect autonomy over what game they play when and how long (Deterding, 2016). However, in research studies, participants are generally assigned to predetermined gameplay conditions. This may lead to frustration as participants may be made to play games they would not usually choose to play and may not like (Ferguson et al., 2017). Certain games may be more widely acceptable than others. For example, players of first-person shooters may be happy to play a casual puzzle game, whereas the reverse may not be true. More generally, genre, controls, or required energy and time to learn may all present differential barriers to engagement with different games (Brown & Cairns, 2004). Thus, they may all interact with player dispositions (genre preference, controller familiarity, etc.) that may covary with other player features (age, gender) to produce patterned differences in play outcomes, engagement, and the like that may confound the treatment-outcome correlation under study.

Game Conditions May Differ. If study participants learn about the differences in treatment and control groups, they may adjust their behaviour accordingly, a validity threat that is usually discussed under the labels of compensatory rivalry and resentful demoralization (Shadish et al., 2002). Compensatory rivalry is the phenomenon wherein a control group puts in extra effort in order to compete against a group receiving an intervention. This may be exacerbated in game-based research given the competition-embracing social norms of gameplay (Deterding, 2014).

Resentful demoralisation is the opposite effect, where one group is demoralised by being put in the inferior condition. Games are generally expected to be fun, but often two experimental conditions cannot be equally fun. Participants who view their game condition as inferior may be subject to resentful demoralisation, e.g. if their condition is excessively difficult or particularly easy. Similarly, if participants are recruited on the basis of playing a game, they may be demoralised to find the game is different to what they anticipated, or they are in a non-game control condition.

Player Factors

In the preceding sections, we discussed principled issues that arise from the constitution of games, no matter the *particular* participants that engage with them. In this section, we summarise player-specific threats to validity, that is, issues that arise due to the constitution of video game players as participants, and issues that arise from interpersonal differences between players:

1. Gamers are not the general population
2. Gamers are diverse

Gamers are not the General Population

To draw inference from a sample to a wider population, the sample must be representative of that population. Else, the external validity of the study is threatened. This is a particular concern for games-based research in that gaming is a voluntary pursuit: individuals self-select to play games, which may lead to sampling bias. Problems arise when characteristics of being a *gamer* moderate study outcomes.

For some, gaming is part of their identity, while others regularly play games without self-identifying as gamers. Yet no matter if they self-identify as gamers or not, people who play games often share certain characteristics such as *gaming capital* (Consalvo, 2007), their accumulated knowledge, experience, and attitudes towards games and gaming. This includes gaming literacy, meaning the degree to which an individual feels confident in understanding games and how they are played. Such concepts, skills and knowledge can be highly transferable. For example, participants who have played a game with a particular control scheme (keyboard and mouse to navigate a first-person shooter) are likely to have an advantage at similar games with similar control schemes over someone who is not familiar with them. Similarly, playing shooter games has been found to improve spatial cognitive abilities (Granic, Lobel, & Engels, 2014).

It seems likely that gamers will be more interested in taking part in a study involving games compared to non-gamers. This is evidenced in online surveys, where proficient players may be over-represented (Khazaal et al., 2014). As a result, games-based research may attract a participant sample with particular skills and knowledge that deviate from the general population. Relatedly, gamer identity and gaming capital may interact with using a game for data collection. Participants who do not identify as a gamer (e.g. older adults (McLaughlin, Gandy, Allaire, & Whitlock, 2012)) may suffer stereotype threat: by anticipating that they will perform badly, their actual performance is decreased (J. L. Smith, 2004). Similarly, expectations about gameplay may heighten evaluation apprehension. While self-identifying gamers may find it socially desirable to perform well in a game and exert extra effort, non-gamers still often view games as a waste of time. Thus, non-gamers may want to downplay their investment in and performance at games, unless the gameplay has a socially acceptable justification (Deterding, 2017). Put differently, social desirability may produce significant performance differences between participants identifying as gamers or non-gamers.

Finally, the use of games as a research instrument may have differential effects on attrition, the drop-out of study participants over time. Some degree of attrition is common with long-running studies, and game-based interventions are in fact sometimes use to promote long-term behaviour change, such as ZOMBIES, RUN! (Six to Start, 2012) or SUPERBETTER (McGonigal, 2012) (Johnson et al., 2016). However, self-identifying gamers or participants with high gaming capital might be more likely to persist with a game-based experiment or treatment, or in contrast, abandon a study earlier because they have higher expectations of game design or perceive the intervention as less novel.

Gamers are Diverse

The gamer stereotype of a white heterosexual male teen (Shaw, 2012) doesn’t reflect the growing diversity of people who play games (Williams, Yee, & Caplan, 2008). That said, playing a game is generally seen as a voluntary activity that individuals self-select into based on their preferences (Deterding, 2016). Existing gaming literacy, socio-economic status, and the like may also affect what kinds of gaming devices and games people access. Hence, certain player characteristics may therefore be over- or under-represented among the users of certain games, genres, or platforms. While recruiting from a console multiplayer first-person shooter may result in a more white, male, gamer-identifying *core* *gamer* sample, recruiting from a mobile casual puzzle game may result in an older, more female sample. These demographic characteristics all present potential confounds.

A related problem frequently raised in the literature is that players typically show different degrees of expertise in different game genres (Latham et al., 2013). Games may differ substantially in the skills they involve, one game may train twitch skills, while another may train executive processes. Differential effects of training with different video games were identified by Subrahmanyam and Greenfield (1994). Put differently, gaming literacy is not a unitary construct. Researchers should control for genre-related expertise to ensure differential representation across condition doesn't confound results.

A third common difference among players are so-called *player types* (Hamari & Tuunanen, 2014). Different people display different stable preferences in play activity and style: they may prefer exploration, socialising, winning, or something else altogether. These interpersonal difference again may confound results if not controlled for.

Discussion and Conclusion

Games are increasingly popular ways to harness research data from large (online) populations – be it as the natural exhaust of entertainment games, be it games intentionally chosen or designed as research instruments. While there has been some work on the validity of games-centred media effects and learning research, little has been done on the validity threats of using games to collect data for non-game-related research questions. We therefore offered a systematisation of potential validity threats of game-based research. While game-based research is just as fallible to general threats to validity (Shadish et al., 2002) such as publication bias (Ferguson, 2007) or issues surfaced in the current debate on reproducible research (Munafò et al., 2017), we particularly focused validity issues characteristic for games and their players.

The latter are maybe the most straightforward to address. Games and especially particularly game genres still attract particular populations with particular preferences and abilities. To some extent, this issue is self-correcting as game-playing becomes ever more prevalent and normalised across populations. Remaining bias can be controlled for with relatively standard research design measures, or simply documented as a limitation.

A less straightforward validity issue is that games are complex systems from which gameplay and player experience emerge non-linearly. This makes it fundamentally difficult to manipulate just one game parameter (as an experimental treatment) without potentially also changing many others, producing potential confounds that threaten internal validity. And because much of the enjoyment and engagement of games revolve around curiosity stoked by novelty and competence fuelled by experiences of learning, games can show strong maturation and attrition effects: playing the same content twice just isn't as fun or challenging as the first time around. The standard methodological response is larger sample sizes with between-subject designs or within-subject designs with randomised ordering of conditions (Shadish et al., 2002). Another solution may be to pretest manipulations prior to the actual study, involving game design expertise to ensure no unexpected emergent confounds manifest.

Furthermore, we found that games, game content, and gameplay are highly varied, and researchers often have relatively little control over what players do and experience in a game and in what order. This makes statistical testing more challenging (or at least often requires larger sample sizes), and threatens internal, construct, and external validity. Different games vary on many dimensions, including crucially game genres. This cautions against operationalising different conditions of constructs as different games, or generalizing findings from one game to another, let alone other game genres. In educational research, authors like Squire (2011) have therefore called to replicate studies on the effects of particular design or instructional strategies across *multiple divergent games* before making any more general claims as to their effectiveness. Because games are far less standardised than e.g. psychometric tasks and instruments, researchers should also document the games used in any publication in as much detail as possible, including screenshots or video figures demonstrating gameplay, the version and section of the game played when using existing entertainment games, and ideally, an executable mirror of the actual game. Else, reviewers and readers will have difficulty assessing the validity of the reported findings, as will researchers interested in replicating the study.

Beyond such first stabs at mitigating strategies, we think the issues of systemic complexity and variance point to a more fundamental research need in games that is as much theoretical as methodological. Games and gameplay can be described on multiple levels of organisation (Klimmt et al., 2007). Developer experience suggests that some or even most of the socially and psychologically functioning mechanisms are located on higher, molar levels of organisation (Hunicke et al., 2004). For instance, providing *personalisation* during character creation should in aggregate increase autonomy need satisfaction, even if each individual player may have interacted with the personalisation interface and therefore have experienced a different moment-to-moment chain of actions and screen events (Turkay & Adinolf, 2015). However, there is little if any consensus nor methodological good practice on what level of organisation to study and manipulate games, or how to theorise and identify causally active mechanisms as constructs. Arguably the most progress in this respect has been made in gamification research, but even here, researchers are mainly pointing out a massive agenda of future desiderata (Deterding, 2015; Landers, Auer, Collmus, & Armstrong, 2018). By comparison, health sciences have now engaged in a decade plus of work modelling and identifying *behaviour change techniques* as the psychological active ingredients in health interventions and are still far from any comprehensive consensus (Michie & Johnston, 2013). Even if a consensus construct (let alone taxonomy) existed, we would still need methods to quickly and reliably identify which kinds of active ingredients a given game held and what subcomponents of said games are involved in each. And yet all this future work would only address the variance and emergent complexity of *games*, not *gameplay*. What differences in gameplay actually make a difference? When and why can we disregard low-level differences in player behaviour because they all instantiate the same molar types of action (Baum, 2002)? What kinds of higher-level dynamics does gameplay reliably gravitate towards (Vahlo, Kaakinen, Holm, & Koponen, 2017)? We are arguably even further from being able to answer these questions.

A final game characteristic threatening validity we identified was social framing: research studies and gameplay are both very particular types of social situations with very particular orderings, norms, roles, and expectations, which may already be triggered simply by verbally and visually labelling a situation as a game or an experiment. While some evidence suggests a close correlation of in-game and real-life behaviour, some suggests marked differences (Deterding, 2016). Scholars like Dmitri Williams (2010) have therefore called for a systematic research programme on the mapping of real and virtual worlds. Almost ten years later, we are still dearly in need of such a concerted effort. By highlighting two characteristic features of play situations – lowered consequence and a license for strategic action – we hope to have given some starting points for it.

And in light of the preceding pages, one may add a second research programme, concerned not just with the meta-methodological preconditions of game-based research (when and where can games even function as research instruments?), but with its *design*. In the wider field of applied games, we now have substantive literatures on how to *integrate* persuasive, educational, or motivational purposes into the design of serious games and gamified systems (Bogost, 2007; Deterding, 2015; Squire, 2011). In research games, such knowledge is dearly missing. By surveying the validity threats that arise in collecting data with games, we hope to have made a first step in closing this gap.

References

Adachi, P. J., & Willoughby, T. (2011). The effect of video game competition and violence on aggressive behavior: Which characteristic has the greatest influence? *Psychology of Violence*, *1*(4), 259–274. *https://doi.org/10.1037/a0024908*

Adachi, P. J., & Willoughby, T. (2011). The effect of violent video games on aggression: Is it more than just the violence? *Aggression and Violent Behavior*, *16*(1), 55–62. *https://doi.org/10.1016/j.avb.2010.12.002*

Anderson, C. A, & Dill, K. E. (2000). Video Games and Aggressive Thoughts, Feelings, and Behavior in the Laboratory and in Life. *Journal of Personality and Social Psychology*, *78*(4), 772–790.

Anderson, C. A., & Bushman, B. J. (2001). Effects of Violent Video Games on Aggressive Behavior, Aggressive Cognition, Aggressive Affect, Physiological Arousal, and Prosocial Behavior: A meta-analytic review of the scientific literature. *Psychological Science*, *12*(5), 353–359. *https://doi.org/10.1111/1467-9280.00366*

Auyang, S. Y. (1999). *Foundations of Complex-System Theories in Economics, Evolutionary Biology, and Statistical Physics*. Cambridge, England: Cambridge University Press. *https://doi.org/10.1017/CBO9780511626135*

Bartle, R. (2004). *Designing Virtual Worlds*. Berkeley, CA: New Riders.

Baum, W. M. (2002). From molecular to molar: a paradigm shift in behavior analysis. *Journal of the Experimental Analysis of Behavior*, *78*(1), 95–116. *https://doi.org/10.1901/jeab.2002.78-95*

Blizzard Entertainment. (2004). *World of Warcraft*. [Video Game]. Blizzard Entertainment

Bogost, I. (2007). *Persuasive Games: The expressive power of videogames*. Cambridge, MA: MIT Press.

Brown, E., & Cairns, P. (2004). A Grounded Investigation of Game Immersion. In *CHI '04 extended abstracts on human factors in computing* (pp. 1297–1300). *https://doi.org/10.1145/985921.986048*

Camerer, C. F., Hogarth, R. M., Budescu, D. V., & Eckel, C. (1999). The Effects of Financial Incentives in Experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, *19*(1-3), 7–42. *https://doi.org/10.1023/A:1007850605129*

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312. *https://doi.org/10.1037/h0040950*

Carnagey, N. L., & Anderson, C. A. (2005). The effects of reward and punishment in violent video games on aggressive affect, cognition, and behavior. *Psychological Science*, *16*(11), 882–889. *https://doi.org/10.1111/j.1467-9280.2005.01632.x*

Castronova, E., Williams, D., Ratan, R., & Keegan, B. (2009). As real as real? Macroeconomic behavior in a large-scale virtual world. *New Media & Society*, *11*(5), 685–707. *https://doi.org/10.1177/1461444809105346*

Chen, J. (2007). Flow in games (and everything else). *Communications of the ACM*, *50*(4), 31–34. *https://doi.org/10.1145/1232743.1232769*

Citizen Science Alliance. (2009). Zooniverse. Retrieved from http://www.zooniverse.org

Consalvo, M. (2007). *Cheating: Gaining advantage in videogames* (p. 228). Cambridge, MA: MIT Press.

Cooper, S. (2015). Massively Multiplayer Research: Gamification and (Citizen) Science. In S. P. Walz & S. Deterding (Eds.), *The gameful world: Approaches, issues, applications* (pp. 487–500). Cambridge, MA: MIT Press.

Costikyan, G. (2013). *Uncertainty in games*. Cambridge, MA: MIT Press.

Cyan. (1993). *Myst*. [Video Game] Brøderbund.

Darley, J. M., & Latané, B. (1968). Bystander Intervention in Emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*(4), 377–383. *https://doi.org/10.1037/h0025589*

Denisova, A., & Cairns, P. (2015). The Placebo Effect in Digital Games. In *Proceedings of the 2015 annual symposium on computer-human interaction in play - CHI PLAY '15* (pp. 23–33). *https://doi.org/10.1145/2793107.2793109*

Deterding, S. (2014). *Modes of play : A frame analytic account of video game play* (PhD thesis). University of Hamburg.

Deterding, S. (2015). The Lens of Intrinsic Skill Atoms: A method for gameful design. *Human-Computer Interaction*, *30*(3-4), 294–335. *https://doi.org/10.1080/07370024.2014.993471*

Deterding, S. (2016). Gameplay: Map or Frame? In *CHI 2016 workshop games as an HCI method.* Retrieved from: http://eprints.whiterose.ac.uk/100129/

Deterding, S. (2017). Alibis for Adult Play: A Goffmanian account of escaping embarrassment in adult play. *Games and Culture 13(3), 260–279. https://doi.org/10.1177/1555412017721086*

Deterding, S., Canossa, A., Harteveld, C., Cooper, S., Nacke, L., & Whitson, J. (2015). Gamifying research: Strategies, opportunities, challenges, ethics. In *Conference on human factors in computing systems* (pp.2421–2424). New York, NY: ACM. *https://doi.org/10.1145/2702613.2702646*

Devlin, S., Cowling, P. I., Kudenko, D., Goumagias, N., Nucciareli, A., Cabras, I., Li, F. (2014). Game Intelligence. In *Computational intelligence and games* (pp. 1–8). Dortmund: IEEE. *https://doi.org/10.1109/CIG.2014.6932917*

Elias, G. S., Garfield, R., & Gutschera, K. R. (2012). *Characteristics of Games*. Cambridge, MA: MIT Press.

Elson, M., & Quandt, T. (2016). Digital games in laboratory experiments: Controlling a complex stimulus through modding. *Psychology of Popular Media Culture*, *5*(1), 52–65. *https://doi.org/10.1037/ppm0000033*

Elson, M., Breuer, J., Looy, J. V., Kneer, J., Quandt, T., Van Looy, J., Quandt, T. (2013). Comparing Apples and Oranges? Evidence for pace of action as a confound in research on digital games and aggression. *Psychology of Popular Media Culture*, *4*(2). *https://doi.org/10.1037/ppm0000010*

Engelhardt, C. R., Hilgard, J., & Bartholow, B. D. (2015). Acute exposure to difficult (but not violent) video games dysregulates cognitive control. *Computers in Human Behavior*, *45*, 85–92. *https://doi.org/10.1016/j.chb.2014.11.089*

Ferguson, C. J. (2007). The good, the bad and the ugly: A meta-analytic review of positive and negative effects of violent video games. *Psychiatric Quarterly*, *78*(4), 309–316. *https://doi.org/10.1007/s11126-007-9056-9*

Ferguson, C. J. (2015). Do Angry Birds Make for Angry Children? A meta-analysis of video game influences on children’s and adolescents’ aggression, mental health, prosocial behavior, and academic performance. *Perspectives on Psychological Science*, *10*(5), 646–666. *https://doi.org/10.1177/1745691615592234*

Ferguson, C. J., Colon-Motas, K., Esser, C., Lanie, C., Purvis, S., & Williams, M. (2017). The (Not So) Evil Within? Agency in video game choice and the impact of violent content. *Simulation and Gaming*, *48*(3), 329–337. *https://doi.org/10.1177/1046878116683521*

Foroughi, C. K., Serraino, C., Parasuraman, R., & Boehm-Davis, D. A. (2016). Can we create a measure of fluid intelligence using Puzzle Creator within Portal 2? *Intelligence*, *56*, 58–64. *https://doi.org/10.1016/j.intell.2016.02.011*

Früh, W., & Schönbach, K. (2005). Der dynamisch-transaktionale Ansatz III: Eine Zwischenbilanz. *Publizistik*, *50*(1), 4–20. *https://doi.org/10.1007/s11616-005-0115-7*

Goffman, E. (1986). *Frame Analysis: An essay on the organization of experience*. Boston, MA: Northeastern University Press.

Granic, I., Lobel, A., & Engels, R. C. M. E. (2014). The benefits of playing video games. *The American Psychologist*, *69*(1), 66–78.

Gundry, D. E., & Deterding, S. (2018). Intrinsic Elicitation: A model and design approach for games collecting human subject data. In *Foundations of digital games 2018*. New York, NY: ACM Press. *https://doi.org/10.1145/3235765.3235803*

Hamari, J., & Tuunanen, J. (2014). Player Types: A meta-synthesis. *ToDIGRA*, *1*(2), 29–53. *https://doi.org/10.26503/todigra.v1i2.13*

Hannula, O., & Harviainen, J. T. (2016). Efficiently Inefficient: Service design games as innovation tools. *Fifth Service Design and Innovation Conference*, 1–12.

Hawkins, G. E., Rae, B., Nesbitt, K. V., & Brown, S. D. (2012). Gamelike features might not improve data. *Behavior Research Methods*, *45*(2), 301–318. *https://doi.org/10.3758/s13428-012-0264-3*

Hilbig, B. E. (2016). Reaction time effects in lab- versus Web-based research: Experimental evidence. *Behavior Research Methods*, *48*(4), 1718–1724. *https://doi.org/10.3758/s13428-015-0678-9*

Hilgard, J., Engelhardt, C. R., & Rouder, J. N. (2017). Overstated evidence for short-term effects of violent games on affect and behavior: A reanalysis of Anderson et al. (2010). *Psychological Bulletin*, *143*(7), 757–774. *https://doi.org/10.1037/bul0000074*

Hoonhout, J. (2008). Let the Game Tester Do the Talking: Think aloud and interviewing to learn about the game experience. In K. Isbister & N. Schaffer (Eds.), *Game usability: Advice from the experts for advancing the player experience* (pp. 65–78). Burlington: MA: Morgan Kaufmann.

Hunicke, R. (2005). The case for dynamic difficulty adjustment in games. In *Proceedings of the 2005 ACM SIGCHI international conference on advances in computer entertainment technology* (pp. 429–433). ACM. *https://doi.org/10.1145/1178477.1178573*

Hunicke, R., LeBlanc, M., & Zubek, R. (2004). MDA: A formal approach to game design and game research. In *Proceedings of the AAAI workshop on challenges in game aI*.

id Software. (1992). *Wolfenstein3D*. [Video Game] Apogee Software.

id Software. (1996). *Quake*. [Video Game] GT Interactive.

Jenkins, J. G. (1946). Validity for What? *Journal of Consulting Psychology*, *10*, 93–98. *https://doi.org/10.1037/h0059212*

Johnson, D., Deterding, S., Kuhn, K.-A., Staneva, A., Stoyanov, S., & Hides, L. (2016). Gamification for health and wellbeing: A systematic review of the literature. *Internet Interventions*, *6*, 89–106. *https://doi.org/10.1016/j.invent.2016.10.002*

Johnstone, T. (1996). Emotional speech elicited using computer games. In *Fourth international conference on spoken language* (pp. 1985–1988). IEEE.

Keusch, F., & Zhang, C. (2017). A Review of Issues in Gamified Surveys. *Social Science Computer Review*, *35*(2), 147–166. *https://doi.org/10.1177/0894439315608451*

Khazaal, Y., Singer, M. van, Chatton, A., Achab, S., Zullino, D., Rothen, S., Thorens, G. (2014). Does self-selection affect samples' representativeness in online surveys? An investigation in online video game research. *Journal of Medical Internet Research*, *16*(7). *https://doi.org/10.2196/jmir.2759*

Kiili, K. (2005). Digital game-based learning: Towards an experiential gaming model. *Internet and Higher Education*, *8*(1), 13–24. *https://doi.org/10.1016/j.iheduc.2004.12.001*

Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers and Education*, *87*, 340–356. *https://doi.org/10.1016/j.compedu.2015.07.009*

Klimmt, C. (2006). *Computerspielen als Handeln: Dimensionen und Determinanten des Erlebens interaktiver Unterhaltungsangebote.* Köln: Herbert von Halem.

Klimmt, C., Vorderer, P., & Ritterfeld, U. (2007). Interactivity and Generalizability: New media, new challenges. *Communication Methods and Measures*, *1*(3), 169–179. *https://doi.org/10.1080/19312450701434961*

Kokkinakis, A. V., Cowling, P. I., Drachen, A., & Wade, A. R. (2017). Exploring the relationship between video game expertise and fluid intelligence. *Plos One*, *12*(11). *https://doi.org/10.1371/journal.pone.0186621*

Koster, R. (2005). *Theory of Fun for Game Design*. Sebastopol, CA: O'Reilly Media.

Landers, R. N., Auer, E. M., Collmus, A. B., & Armstrong, M. B. (2018). Gamification Science, Its History and Future: Definitions and a research agenda. *Simulation and Gaming*. *https://doi.org/10.1177/1046878118774385*

Latham, A. J., Patston, L. L. M., & Tippett, L. J. (2013). Just how expert are "expert" video-game players? Assessing the experience and expertise of video-game players across "action" video-game genres. *Frontiers in Psychology*, *4*. *https://doi.org/10.3389/fpsyg.2013.00941*

Lieberoth, A. (2015). Shallow gamification: Testing psychological effects of framing an activity as a game. *Games and Culture*, *10*(3), 229–248. *https://doi.org/10.1177/1555412014559978*

Littman, R. A., & Rosen, E. (1950). Molar and molecular. *Psychological Review*, *57*(1), 58–65. *https://doi.org/10.1037/h0056560*

Louvel, G. (2018). 'Play as if you were at home': dealing with biases and test validity. In A. Drachen, P. Mirza-Babaei, & L. E. Nacke (Eds.), *Games user research* (pp. 393–402). Oxford, England: Oxford University Press.

McDonald, E. (2017). The Global Games Market Will Reach $108.9 Billion in 2017 With Mobile Taking 42%. Retrieved from: https://newzoo.com/insights/articles/the-global-games-market-will-reach-108-9-billion-in-2017-with-mobile-taking-42/

McGonigal, J. (2012). SuperBetter. Retrieved from https://www.superbetter.com/

McLaughlin, A., Gandy, M., Allaire, J., & Whitlock, L. (2012). Putting fun into video games for older adults. *Ergonomics in Design*, *20*(2), 13–22. *https://doi.org/10.1177/1064804611435654*

McMahan, R. P., Ragan, E. D., Leal, A., Beaton, R. J., & Bowman, D. A. (2011). Considerations for the use of commercial video games in controlled experiments. *Entertainment Computing*, *2*(1), 3–9. *https://doi.org/10.1016/j.entcom.2011.03.002*

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.

Michie, S., & Johnston, M. (2013). Behavior change techniques. In *Encyclopedia of behavioral medicine* (pp. 182–187). New York, NY: Springer. *https://doi.org/10.1007/978-1-4419-1005-9_1661*

Mohseni, M. R., Liebold, B., & Pietschmann, D. (2015). Extensive modding for experimental game research. In *Game research methods* (pp. 323–340). Pittsburgh, PA: ETC Press.

Morse, J. M., Barrett, M., Mayan, M., Olson, K., & Spiers, J. (2008). Verification Strategies for Establishing Reliability and Validity in Qualitative Research. *International Journal of Qualitative Methods*, *1*(2), 13–22. *https://doi.org/10.1177/160940690200100202*

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie Du Sert, N., Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9. *https://doi.org/10.1038/s41562-016-0021*

Murray, D. M. (1998). *Design and Analysis of Group-Randomized Trials*. Oxford, England: Oxford University Press.

Nahid Golafshani. (2003). Understanding Reliability and Validity in Qualitative Research. *The Qualitative Report*, *8*(4), 597–607.

Oladimeji, P., Thimbleby, H., Curzon, P., Iacovides, I., & Cox, A. (2012). Exploring unlikely errors using video games : An example in number entry research. In *Workshop on safety-critical systems and video games: Contradictions and commonalities, held at fun and games* (pp. 3–7).

Orne, M. T., & Whitehouse, W. G. (2000). Demand characteristics. In A. Kazdin (Ed.), *Encyclopedia of psychology* (pp. 469–470). Washington, DC: American Psychological Association.

Pham, M. T. (1996). Cue representation and selection effects of arousal on persuasion. *Journal of Consumer Research*, *22*(4), 373–387. *https://doi.org/10.1086/209456*

Przybylski, A. K., Rigby, C. S., & Ryan, R. M. (2010). A Motivational Model of Video Game Engagement. *Review of General Psychology*, *14*(2), 154–166.

Przybylski, A. K., Rigby, C. S., Deci, E. L., & Ryan, R. M. (2014). Competence-impeding electronic games and players' aggressive feelings, thoughts, and behaviors. *Journal of Personality and Social Psychology*, *106*(3), 441–457. *https://doi.org/10.1037/a0034820*

Reimers, S., & Stewart, N. (2007). Adobe Flash as a medium for online experimentation: A test of reaction time measurement capabilities. *Behavior Research Methods*, *39*(3), 365–370. *https://doi.org/10.3758/BF03193004*

Riot Games (2009*). League of Legends*. [Video Game]. Riot Games

Roth, C., Vermeulen, I., Vorderer, P., & Klimmt, C. (2012). Exploring Replay Value: Shifts and continuities in user experiences between first and second exposure to an interactive story. *Cyberpsychology, Behavior, and Social Networking*, *15*(7), 378–381. *https://doi.org/10.1089/cyber.2011.0437*

Salen, K., & Zimmerman, E. (2004). *Rules of Play: Game design fundamentals*. Cambridge, MA: MIT Press.

Sarkar, A., Williams, M., Deterding, S., & Cooper, S. (2017). Engagement Effects of Player Rating System-Based Matchmaking for Level Ordering in Human Computation Games. In *FDG'17*. New York: ACM Press. *https://doi.org/10.1145/3102071.3102093*

Schlenker, B. R., & Bonoma, T. V. (1978). Fun and Games : The validity of games for the study of conflict. *The Journal of Conflict Resolution*, *22*(1), 7–38. *https://doi.org/10.1177/002200277802200102*

Schrader, C., & Bastiaens, T. J. (2012). The influence of virtual presence: Effects on experienced cognitive load and learning outcomes in educational computer games. *Computers in Human Behavior*, *28*(2), 648–658. *https://doi.org/10.1016/j.chb.2011.11.011*

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (pp. 1–643). Boston, MA: Houghton Mifflin.

Shaker, N., Togelius, J., & Nelson, M. J. (2016). *Procedural Content Generation in Games*. Cham, Switzerland: Springer. *https://doi.org/10.1007/978-3-319-42716-4*

Shaw, A. (2012). Do you identify as a gamer? Gender, race, sexuality, and gamer identity. *New Media and Society*, *14*(1), 28–44. *https://doi.org/10.1177/1461444811410394*

Silvia, P. J. (2006). *Exploring the Psychology of Interest*. Oxford, England: Oxford University Press.

Six to Start (2012). *Zombies, Run!* [Video Game]. Six to Start

Slegers, K., Maurer, B., Bleumers, L., Krischkowsky, A., Duysburgh, P., & Blythe, M. (2016). Game-based HCI Methods: Workshop on playfully engaging users in design. *CHI Extended Abstracts on Human Factors in Computing Systems*, 3484–3491. *https://doi.org/10.1145/2851581.2856476*

Smith, J. H. (2006). *Plans and Purposes: How videogame goals shape player behaviour* (Doctoral dissertation). IT University of Copenhagen. Retrieved from *http://jonassmith.dk/weblog/wp-content/dissertation1-0.pdf*

Smith, J. L. (2004). Understanding the process of stereotype threat: A review of mediational variables and new performance goal directions. *Educational Psychology Review*, *16*(3), 177–206. *https://doi.org/10.1023/B:EDPR.0000034020.20317.89*

Smith, S. P., Blackmore, K., & Nesbitt, K. (2015). A Meta-Analysis of Data Collection in Serious Games Research. In C. S. Loh (Ed.), *Serious games analytics* (pp. 31–55). Cham, Switzerland: Springer. *https://doi.org/10.1007/978-3-319-05834-4_2*

Squire, K. D. (2011). *Video Games and Learning: Teaching and participatory culture in the digital age*. New York, NY: Teachers College Press.

Squire, K. D. (2008). Video Games and Education: Designing learning systems for an interactive age. *Educational Technology Magazine: The Magazine for Managers of Change in Education, 48*(2), 17–26.

Subrahmanyam, K., & Greenfield, P. M. (1994). Effect of video game practice on spatial skills in girls and boys. *Journal of Applied Developmental Psychology*, *15*(1), 13–32. *https://doi.org/10.1016/0193-3973(94)90004-3*

Turkay, S., & Adinolf, S. (2015). The effects of customization on motivation in an extended study with a massively multiplayer online roleplaying game. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *9*(3). *https://doi.org/10.5817/CP2015-3-2*

Vahlo, J., Kaakinen, J. K., Holm, S. K., & Koponen, A. (2017). Digital Game Dynamics Preferences and Player Types. *Journal of Computer-Mediated Communication*, *22*(2), 88–103. *https://doi.org/10.1111/jcc4.12181*

Williams, D. (2010). The mapping principle, and a research framework for virtual worlds. *Communication Theory*, *20*(4), 451–470. *https://doi.org/10.1111/j.1468-2885.2010.01371.x*

Williams, D., Contractor, N., Poole, M. S., Srivastava, J., & Cai, D. (2011). The Virtual Worlds Exploratorium: Using large-scale data and computational techniques for communication research. *Communication Methods and Measures*, *5*(2), 163–180. *https://doi.org/10.1080/19312458.2011.568373*

Williams, D., Yee, N., & Caplan, S. E. (2008). Who plays, how much, and why? Debunking the stereotypical gamer profile. *Journal of Computer-Mediated Communication*, *13*(4), 993–1018. *https://doi.org/10.1111/j.1083-6101.2008.00428.x*

Wouters, P., Nimwegen, C. van, Oostendorp, H. van, & van Der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology*, *105*(2), 249–265. *https://doi.org/10.1037/a0031311*

Zendle, D., Cairns, P., & Kudenko, D. (2015). Higher Graphical Fidelity Decreases Players' Access to Aggressive Concepts in Violent Video Games. In *Proceedings of the 2015 annual symposium on computer-human interaction in play* (pp. 241–251). London: ACM. *https://doi.org/10.1145/2793107.2793113*

Živić, N., Andjelković, I., Özden, T., Dekić, M., & Castronova, E. (2017). Results of a massive experiment on virtual currency endowments and money demand. *PLoS ONE*, *12*(10), 1–14. *https://doi.org/10.1371/journal.pone.0186407*