A Practical Guide to Using Digital Games as an Experiment Stimulus

Simo Järvelä
School of Business, Aalto University
Department of Information and Service Economy
simo.jarvela2@aalto.fi

Inger Ekman
University of Tampere
inger.ekman@uta.fi

J. Matias Kivikangas
School of Business, Aalto University
matias.kivikangas@aalto.fi

Niklas Ravaja
Department of Social Research and Helsinki Institute for Information Technology, University of Helsinki
School of Business, Aalto University
niklas.ravaja@aalto.fi

Introduction

Digital games[1] engage the player in complex behavior, which—depending on the game design—can call upon various types of cognitive and emotional processes. As such, games provide an excellent vessel for examining a multitude of concepts central to psychology, from memory encoding, to social skills and decision making. Game-like task setups are classic to experimental psychology: early examples include e.g. Deutsch & Krauss's Trucking Game (1960) and The Prisoner's Dilemma (Jones et al. 1968). Contemporary psychological research has also begun to utilize digital games (e.g. Fehr & Gächter 2002; Frey

et al. 2007; Slater et al. 2003). In a summary on the use of games in psychological research, Washburn (2003) distinguishes four distinct manners of using digital games in experimental setups: utilizing games as stimulus to study other forms of behavior; involving games to manipulate variables; using games to provide education and instruction; and employing gaming as a performance metric. In addition to psychological studies, games are central stimuli to any research striving to understand games and gaming as a phenomenon, evaluating design decisions, and measuring the effects of playing or the gaming experience itself.

As of yet, there exists little instruction on how to choose digital games for experiments, including research directly focused on the gaming experience, or the short and long term effects of gaming. The field also lacks guidelines regarding the experiment setup with games, and the work relies on accumulated know-how. This presents challenges to both researchers themselves and for those who are interested in the published results. It is especially challenging to compare findings between various studies or to generalize the results across different experimental setups. These difficulties will likely become even more pertinent as interest towards games spreads to new disciplines, as suggested by the use of games, for example, to present forensic evidence in the courtroom (Schofield 2011), or to study animal cognition (ludusanimalis.blogspot.se).

In addressing the use of games in experimental setups, the recent work by McMahan et al. (2011) is a rare exception, as it tackles the relative merits and drawbacks of using commercial video games as a stimulus. The authors also present criteria for game selection and game mode selection, and mention the importance of controlling participant demography. However, they offer only brief discussion on the importance of managing confounds during gameplay and the experiment they present considers only very straightforward gaming tasks where

play affects the scenario minimally. This paper takes up the discussion, extending the level of detail.

We have employed games as stimuli in our lab at the Aalto University School of Business for a decade now, using them and psychophysiological methods (Cacioppo et al. 2007) to study the gaming experience (e.g. Kivikangas et al., 2011; Ravaja et al. 2004, 2006a, 2006b, 2008), but we have also used games to access other processes, such as learning (Cowley et al. 2012), social dynamics and physiological linkage (Järvelä et al. 2013) and multimodal information processing (Ekman et al. 2010). Altogether, this body of work covers all four functions identified by Washburn. This contribution draws upon practical know-how gathered during the course of these experiments and the considerations we have found, sometimes by trial and error, to be pertinent for using games as a stimulus.

Different research methods place different demands on how digital games are best utilized, and also on what has to be taken into account when designing the experiment and analyzing the data. We consider motives for game choice, use of metrics, and approaches to controlling relevant experimental variables. We also describe the practical issues involved in setting up an experiment utilizing a commercially available game title. While the focus of this paper is on digital games, various virtual environments provide similar possibilities and challenges when used as a stimulus in experiments. The following discussion considers uses of games in very strictly controlled studies. Therefore, the work will be valuable both to researchers who wish to utilize games in similar studies, but also provide relevant considerations to those working with more forgiving setups. In addition, readers interested in the results of game-related research may find this paper useful when evaluating published studies, considering the possible pitfalls in experimental setups, reconciling conflicting data and assessing the generalizability and relevance of individual results.

GENERAL CONSIDERATIONS FOR CHOOSING DIGITAL GAMES AS STIMULI

Digital games are a natural choice for a stimulus, not only when studying gaming and the gaming experience, but also for other research questions calling for an engaging, yet challenging activity (Washburn 2003). Digital games, and modern games especially, are very complex stimuli and they are in many ways a unique form of media. There are a large number of readily available commercial games that could potentially be used in an experiment, but the choice has to be made carefully.

Advantages of Using Digital Games in Experiments

According to the Electronic Software Association (ESA 2011), 72% of American households play digital games. Gaming is not limited to a certain age group, and 29% of the gamers are above 50 years old. A study in Finland (where we recruit most of the test subjects) showedthat 54% of the respondents were active video gamers. Non-digital games included, as many as 89% reported playing games at least once a month (Karvinen & Mäyrä 2011). This confers three specific benefits. First, the high penetration in the population serves to make games more approachable than abstract psychological tasks, which helps in recruiting participants. Second, the high familiarity with games allow the use of more complex tasks, that engage subjects in ways that would be very difficult to grasp if framed as abstract psychological assignments. Third, with proper screening, test procedures can rely on previously gathered exposure, which allows addressing, for example, accumulated skills and domain expertise. With experienced players detailed instructions are not needed unless it is desirable that the participants play the game in a specific manner.

As digital games are designed to address a range of emotions and with specific intent to cause certain reactions within the player, successful titles can be considered highly ecologically valid[2] instruments for

eliciting emotions for various purposes. Different game genres typically address different emotions, e.g. horror games aim for quite different emotional reactions and mood than racing games or educational games. Meta-genres such as casual games or social games introduce yet further dimensions to the emotional spectrum of playing. With the proper selection of games, a broad scale of emotions can be elicited in a relatively targeted fashion. However, as games most often do not focus on a single emotion, genres and styles are not guaranteed to provide any specific experience.

Furthermore, digital games provide safe virtual environments to conduct studies on topics and situations which might present either practical or ethical challenges in a non-digital form. Yet the level of realism in games and virtual environments is high enough that they can potentially be used to simulate and draw conclusions about real-world events. For example Milgram's classic study (1963) is considered unethical by today's standards, but Slater and colleagues (2003) were able to replicate the study using a virtual game-like activity. In addition, as McMahan and colleagues (2011) state, using off-the-shelf games provides benefits of quick implementation, avoids some researcher bias and enhances study reproducibility.

Challenges

The distinctive qualities of games have to be well acknowledged if they are to be used in an experiment. Particularly the variation inherent in gaming will call for extra care in choosing the game title(s) for the experiment and defining experiment procedure. Furthermore, adequate data collection might prove challenging when using commercial games due to limited logging capabilities.

Similarity of stimulus

A major challenge with games is that the actual content of the game is defined and shaped by various factors. This creates a challenge for

experimental research, where it would often be preferable to use as identical stimuli as possible across all study participants. Instead, with games the interactive stimulus is never the same, but changes according to participant actions. In virtual environments and MMOGs (massively multiplayer online games) this is even more prevalent as they are influenced by a large number of players at the same time. In addition, game settings, random elements within the game and AI operation all affect how the game proceeds. While the fact that games are widely played ensures target group familiarity, the disparate skill levels of players can also considerably affect how they play and experience a game. Since games are interactive, this skill difference tends to cause not only different experiences, but often leads to changes in the actual content of the game. For example, a skilled player will likely progress further in a given time, use more diverse and effective playing styles, or have an access to more advanced game items than a less experienced player.

Therefore it is of utmost importance that the researcher is well aware of the dependent variables and how they might be influenced by the stimulus properties that vary between participants. The choice of what game is used must be done so that the stimulus is sufficiently identical between participants in the aspects relevant to the dependent variable. After that, any additional variance in the game can be considered irrelevant for the experiment, but it is good to note that the variation still contributes to the attractiveness of the game for the participant. It would be a mistake on part of the researcher to seek to strip a game from all variance, and risk making the game into just another psychological task without the positive qualities games can offer.

Furthermore, it is important to acknowledge that since game research is still a young field, there is little agreed upon theory on precisely which are the relevant aspects for a particular effect or game quality, or how to systematically describe them. Thus, even seemingly simple

decisions will likely be based on assumptions about aspects that are not yet fully understood. As is common in debates around new media forms, the discussion on digital games has its dystopian and utopian visions, which introduce a number of personal and political agendas into research. Particularly for researchers that are personally less familiar with games there is a significant risk of overlooking how seemingly separate game features combine and influence the playing experience, that is, failing to identify game-specific features that confound the main effect (c.f. Adachi & Willoughby [2011] discussing the possibility that it is competition, not violent content, that accounts for game-induced aggression). An agreement on desirable procedure can help mitigate these issues and make work more accessible and comparable across discipline borders.

Off-the-shelf vs. custom games

In general, the closed code of commercial games limits the possibility of modifying the game to suit the experiment. Developer tools and mod kits make some adjustments possible. For example Staude-Müller et al. (2008) used mod kits for *Unreal Tournament 2003* (Epic Games & Digital Extremes 2002) to modify the game to suit their experimental setup and also controlled the stimulus and documented it in exemplary manner. However, it is worth noticing that any major changes come with a risk of compromising game quality. The closed system of most commercial games can also make it difficult to ensure what the program actually does. Adaptive difficulty adjustments, randomly spawning adversaries and minute modifications to auditory and visual stimuli can be hard to spot without extensive game analysis prior to the experiment, but still affect the results.

A common disadvantage with commercial games is also the lack of logging capabilities (i.e. saving the data about what exactly happens in the game on code level). In some cases open source alternatives are practical for this particular reason. If available, log files are immensely

useful, as they can be used in e.g. event based analysis, segmentation, performance appraisals and to spot game manipulations not evident from video recordings.

It is not uncommon for researchers to develop their own games to ensure that they target the desired effects and have a full control over the stimulus. With custom-made games the researchers have an opportunity to modify every detail of the stimulus and tailor the task to suit whatever the experiment might need. However, in addition to requiring considerable amount of work and time, custom-developed games may introduce experimenter bias. Games developed by small-budget research teams also are less likely to be as well-balanced, rich in content and engaging as commercial titles designed and developed by professionals. Employing less engaging games for research undermines one of the biggest advantages of using games as stimulus: when the games are engaging, the participants focus deeply on the task at hand and are more likely to act as they would outside an experiment and feel less distracted by the experimental setup. Thus, more engaging stimulus can produce better data.

PRACTICAL AND METHODOLOGICAL CONSIDERATIONS

Besides general considerations on why to use digital games as a stimulus in the first place, there are several more practical and study specific questions that are relevant when designing an experiment. In this chapter we will discuss issues that are tightly connected to the methodology used. In our experience, there are four main considerations when preparing a study using games as a stimulus: (1) matching and regulating task type, (2) determining data segmentation and event coding, (3) ensuring compatibility between participants and (4) planning and conducting data collection.

Matching and Regulating Task Type

Finding a suitable game is one of the first steps in designing a study.

Gameplay consists of various tasks that define what type of a stimulus the game actually is. One way of approaching the question is to examine the kinds of cognitive tasks that are necessary to overcome the challenges presented in the game: concentration, problem solving, using memory, quickly focusing attention, fast reflexes, planning ahead, spatial awareness, etc., are all tasks that are common in games, but disparate game genres generally weigh the importance of various cognitive tasks differently. Furthermore, all game tasks need to be considered in relation to the context they are presented in—the same task, but e.g. with different time limitations will produce vastly different reactions. Intense repetition and extended task times can also significantly change the nature of a task compared to less taxing options. For example, both *Tetris* (Spectrum Holobyte 1985) and a modern first-person shooter game might be an appropriate stimulus for a performance-based stressor task, but while the first is designed to be constant and increasing stress, the second might have wildly varying arousal levels (depending on the game, level, and play style), not to mention the added efforts of 3D spatial processing, emotional content from the narrative, and so on.

Naturally the game should be chosen according to what type of a stimulus is preferable. There are no general rules applicable for how to make this selection. Games differ widely even within the same genre, and yet—depending on the research questions—comparable effects may be found in games of very different styles. In fact, choosing a game title is only part of the task of determining the experiment stimulus. The choice of stimulus goes down into choosing levels and playing modes, and narrowing down tasks that are conducive to the intended research. For example, a study examining the effects of violent digital games might be based on General Aggression Model, which posits that violence in games elicits arousal and that contributes to resulting aggressive behavior (Bushman & Anderson 2002). In order to make such claims, it would be of utmost importance to make

sure that the compared games would not differ in quality, that the pace of the game is similar in both cases and that the overall gaming experience is equally engaging in both cases, as all these factors might affect arousal levels (cf. Adachi & Willoughby 2011). Often this has proved to be a challenging requirement to meet. For example, Ballard & Wiest (1996) conducted a study where the classic fighting game *Mortal Kombat*™ (Probe Entertainment 1993) was compared to a no-name billiards game to find out the effects of violence to hostility and cardiovascular reactivity. However, in addition to the amount of violence, the two games are so remarkably different on a number of factors (e.g. pace, characters, and type of challenge) that the differences in reactions can hardly be pinpointed to be the result of an increase in violence. Yet, the same experiment also provides a positive example of stimulus control by comparing two modes of *Mortal Kombat*™—with or without blood. In doing so all other factors remained the same, which creates a strong setup for examining the effects of increased violence-related content.

When available, game taxonomies provide helpful sources for making informed game choices. Lindley (2003) slightly modifies Caillois' (1961) classical four elements (*competition, chance, simulation,* and *vertigo)* identifying three primary descriptors (*narrative, ludology,* and *simulation*), upon which operate additional dimensions differentiating the level of *chance vs. skill, fiction vs. non-fiction*, and *physical vs. virtual*. Elverdam & Aarseth (2007) provide a higher level of detail with their 17-dimension taxonomy. Their taxonomy bears a strong link to game design, indeed, they specifically point out the relation to the component framework in Björk & Holopainen's (2004) *Patterns in Game Design*. Finally, Whitton (2009) provides a breakdown of game choice for education, in which she details the expected cognitive and emotional engagement within certain genres. Beyond these, less general taxonomies abound, for example differentiating games particularly based on interaction style (Lundgren & Björk 2003; Mueller,

Gibbs & Veter 2008), or the forms of social interaction they provide (Manninen 2004).

Reviews and ratings (for online reviews and rankings, see GameSpot, GameZone, IGN, Metacritic, or GameRankings)[3] can also be helpful in choosing the game. The ratings give an overall assessment on the quality of the game, which—while not objective—is not influenced by researchers' own views and preferences. Ratings are especially helpful when selecting multiple games to be used in the same experiment, as similar ratings lessen the risk that observed differences are simply due to comparing games of diverse quality. For example, Shu-Fang Lin (2011) studied the effects of shooting either human or monster opponents in a digital game. The study was conducted using *Left 4 Dead* (Turtle Rock Studios 2008) and *25 to Life* (Avalanche Software & Ritual Entertainment 2006) as stimuli. This study completely overlooks the significant difference in quality between the two games (*Left 4 Dead* has received a Metacritic metascore of 89/100 while *25 to Life* scores 39/100), and also ignores the impact of genre (survival horror vs. gangsta shooter) and the player character's portrayed motivations for killing opponents (survival vs. lifestyle), which all introduce confounds to the reported effects.

Commercial games commonly have large number of adjustable features which can be utilized in the experiment setup. Visual settings, sounds, game preferences, difficulty levels, number of opponents, play time, and controls can all be used in controlling the stimulus and creating the necessary manipulations. Finally, task choice (the game actions) involves considering the length of task (can the task be extended, how long does it take, how much does the length vary between participants, and is there enough or too much repetition?), how static the action is (is the difficulty level static or does it vary?). For any extended play scenarios it is necessary to consider how well the intended playing time matches the game in question, so as not to create

untypical scenarios which would undermine the ecological validity[2] of the gaming scenario.

- Define your tasks and find out what can be expected to affect them to get an understanding what kind of games could be suitable and which could not.
- Play the potential game to get a feel for the tasks involved and to spot factors that might inadvertently influence your task.
- Use available reviews to pinpoint effects, challenges, and possible shortfalls in the game design. Compare those with your understanding of relevant aspects of the task.
- Use available ratings to ensure the quality level of the game meets the study requirements.
- Utilize game levels and game control features to create desired variation.

Determining Data Quantification and Event Coding

To be able to analyze effects associated with gaming, researchers typically need a strategy to quantify the gaming data. One possibility is of course to use a block design, for example to compare different games, levels, or game modes against each other. However, sometimes block designs are inadequate. For example, the focus of interest may be smaller events, such as particular actions (e.g. finishing the race, killing an opponent in a first-person shooter [FPS], or picking up a mushroom). For these cases, event-based analysis allows researchers to gain data on the events of interest, and minimize the confounding data from actions occurring before and after the moment of interest. Event-based designs, however, introduce some additional considerations for the researcher. The choice of event coding is based not only on the game's available actions, but also on how isolated these actions occur during gameplay. Often there are over-lapping events that are hard to differentiate from each other. With multiple elements affecting the subject at the same time, it can be impossible to say which of the elements caused a certain reaction or behavior (and to say, for

example, whether the reactions during a combat FPS game were due to shooting at the enemy, to being shot at, or to both). On the other hand, if events are too unique, the sample size might not be adequate for statistical analysis unless it is compensated with a high number of participants. The easiest events to study are those that appear frequently, and in sufficient isolation from everything else.

The same repeating event can occur in different contexts within the game thus framing it differently and so having a different meaning. Whereas some of this diversity can be controlled by fixing game parameters, the level of control varies greatly between games. The common solution is to gather a large enough sample of similar events so that the effect of random noise (e.g. slightly varying framing of the same event) is balanced out. Naturally these considerations should also affect game choice, as games where the same type of event occurs repeatedly are more suitable stimuli as it is easier to have a satisfying sample size of events under scrutiny.

The optimal time scale needed for events has to be balanced in relation to the metrics used in the experiment. Various methods have different time resolutions. This often limits the size of events that can be examined. The necessary resolution influences the temporal accuracy needed for timestamps and also for data synchronization; these should all be in accordance with the research method used. The aim is to select a resolution for event coding that does not limit what can be analyzed from the data. Therefore, even longer duration events should preferably be coded with very accurate starting and ending times. As an example, the psychophysiological method (Cacioppo et al. 2007) allows accessing precise events, as the data is gathered continuously with millisecond precision. To benefit from this level of accuracy, game events must also be coded with millisecond precision. The nature of the effect under scrutiny also determines the necessary duration of events and how event response times are matched to metrics.

The choice of method for analyzing the data can to some extent mitigate the challenge provided by concurrent and overlapping events. For example, the Linear mixed model (Hierarchical linear models) incorporates both fixed effects and random effects, and is particularly suitable to repeated measurements, where the effect is simultaneously influenced by many factors. This statistical method is necessary if the data is hierarchical (e.g., events within conditions within participants) or the number of samples varies within the unit of analysis (e.g., if a particular event occurs a different number of times for diverse players). Simpler data structures may offer the possibility to use other analysis methods.

While typical events in digital games are quite clearly separable from others, in some cases it is not self-evident how events should be defined. They might take over a longer undefined period of time (e.g. in a horror game, how long exactly does the suspense before release last?), or larger events may consist of a number of smaller events in ways that are difficult to precisely define for coding purposes. In these cases data driven approaches may be utilized to explore what clusters of events occur in the material, for example applying machine learning algorithms to find repeating patterns and connections in the event data (see e.g. Kosunen 2012). Data driven approaches may also be applied in order to provide complementary perspective to, or even to test the validity of, coding strategies done by other means.

When deciding on the event coding, it is useful to remember that one can always go from specific to general, but rarely the other way around without recoding the data. Finally, event coding is closely related to data acquisition and how you plan your experiment. It is advisable to have a clear idea of what events will be used in analysis and how they are to be processed, and plan the experiment accordingly. Options are often quite limited afterwards if enough data was not collected in the first place.

- Choose a game where the desired events occur often enough, preferably in isolation.
- Critically consider the various contexts in which events occur. In case of suspected effect, keep track of the context (log it) for each event occurrence.
- Ensure that the event of interest and metrics operate on similar time scales.
- Mitigate overlap and simultaneity by choice of statistical method. Take care that the hierarchical nature of data is accounted for.
- Consider data driven approaches if applicable.
- Code too much rather than too little detail. Extra coding can always be disregarded later, but accessing uncoded material is difficult.

Ensuring Compatibility Between Participants

Fundamental to a successful experiment is ensuring compatible test conditions between multiple participants. Since the game as stimulus changes depending on the participants' choices, skill level, and preferences, this requires a balance between stimulus design (see Matching and regulating task type) and careful participant selection.

Recruiting participants

Unless the research specifically addresses learning, some experience with digital games is usually preferable, as learning basic skills can take up significant time and effort, and any time spent on training sessions are away from the actual experiment tasks. Choosing only subjects that are experienced enough with the task at hand can ensure deeper skill levels during the experiment than what could be achieved by including a practice session or by giving instructions prior to the test session. In contrast, if novices are given too little time to get acquainted with the game, the lack of basic gaming skills is likely to influence the quality of the data. Importantly, gaming skills do not necessarily transfer across genre borders, and even within a certain genre small changes in e.g. controller behavior can have a major impact on play performance.

Theoretically, a large enough random sample of males and females provides the best basis for generalizing results over the general population and avoiding a gender bias. However, in practice this goal is often problematic to achieve. Although many women play digital games, gaming is still much more common among the male population (ESA 2011; Karvinen & Mäyrä 2011), and therefore acquiring comparable numbers of experiment participants of both genders with good sample size can sometimes be difficult—particularly so if comparable gaming experience is a prerequisite. Similarly, it is virtually impossible to conduct an experimental study that would have enough participants in each age group to provide statistically significant results without limiting the amount of relevant variables through participant selection. Instead, these factors have to be taken into account when analyzing the data, interpreting the results, and generalizing them.

Comparable stimuli

It is impossible to create gaming stimuli that is identical for all participants. Instead of aiming for similarity, the researcher should focus on what makes or breaks the experience of interest, and devise strategies for handling variation within this perspective. To ensure stimuli are comparable, and to minimize the impact of variation on results, the imperative is to identify the critical factors that affect the dependent variable(s), and control those as well as possible. Indeed, some variations may be necessary to ensure the overall gaming experience is compatible between participants. Moreover, in some cases individual variation in actual game content is not a problem, for example if measurement concerns general-level experiences such as overall performance and stress levels. Also, if both events and measurements can be narrowed down to a shorter time frame, these shorter spans of gameplay can be comparable between participants even when the whole game sessions are not.

One common aspect which requires consideration is game difficulty. Some games have built-in difficulty adjustments that automatically balance and change the difficulty of the game according to the player's performance and choices within the game. Depending on the context and what is being studied, self-adjusting difficulty levels may either escalate or counterbalance the challenges of using a stimulus with inherent variability. When the aim is to ensure similar experiences across players, automatic adjustment can be useful in creating relatively equally challenging gaming experiences to players of varying skill levels. In contrast, if using the same content for all participants is critical for the experiment, automatic difficulty adjustments can be detrimental to the process. Furthermore, automatic difficulty adjustment is often hard to detect. In the absence of reliable information (e.g. from the developer) to confirm or rule out automatic difficulty adjustment, identifying it generally requires considerable familiarity with digital games. Moreover, even knowing that a game has difficulty adjustment, a researcher may struggle to determine precisely how the system works and how it impacts content.

If performance, and processes related to it (such as general arousal and feelings of frustration), are not relevant for the dependent variable, the difficulty of the game might not be relevant either. In such cases, difficulty level could even be left to participants to choose for themselves. However, this might necessitate using other ways to ensure comparability between trials, for example, by assessing subjective difficulty by a post questionnaire.

- Be selective with your participants, but cautious about generalizing results.
- Pay special attention to gaming experience already when recruiting participants.
- Evaluate gaming experience for the specific genre, game type and title used in the experiment.
- Decide if it is more important to ensure identical tasks/events, or identical difficulty level—if not possible to control both. If possible, include a metric to capture the dimension you do not control (subjective difficulty, counting the number of adversaries, etc.).

Planning and Conducting Data Collection

Depending on the research method used a varying amount of data is needed but all data segmentation and event based analysis require information on what happened in the game. When available, automatically logging gameplay provides a superior method for segmenting system data with sufficient temporal accuracy. Most games do not employ sufficient logging of game events, or alternatively, logs are not available to the researcher. In this case, events have to be marked afterwards by reviewing recorded gameplay (e.g. from video recordings), which can be very laborious. Furthermore, it is often the case that not all player actions can be identified and differentiated based on mere recordings—in modern games with lots of different objects on the screen, it is not clear from the game video alone where the attention of the player is focused at a given moment, for example (though eye trackers can be used for that). Mod kits often provide extended logging capabilities, if available.

If a built-in logging system is not feasible, some logs can be collected externally. Key loggers, screen capture videos, and mouse-click recorders can provide helpful material both for analysis and preprocessing data before manual coding. At least a screen capture video of

the game play should be recorded. Be sure to include good quality sound, as audio cues may be used to differentiate between visually similar-looking actions or inform about off-screen events. Most games have one or more innate performance metrics in them. High scores, achievements, goals, kills, repetitions, accuracy, lap times, duration, rewards, new items, levels, etc., can be used as dependent variables or as covariates, complementing and validating external performance metrics.

It is imperative to calibrate the timestamps of different data sources. This is especially important if the analysis will operate on event data instead of whole blocks. Whereas some game events can be matched manually afterwards, other data sets—like psychophysiological signals—contain no unambiguous handles for time-synchronizing data post hoc, and data will be practically useless to the analysis if the timestamps do not correspond. Depending on the setup there are several methods for anchoring timestamps across devices, for example, sending markers across devices, synchronizing device clocks or using video cameras. The precision of synchronization needed is naturally dependent on the research question, the measurements, and choice of method.

- Utilize game logs whenever available.
- Consider using external logging to capture game data.
- Take advantage of the game's performance metrics when possible.
- Use the game's internal performance metrics to check external performance metrics.
- Be extra careful to calibrate and synchronize timestamps across data sources.

Checklist of Questions

The following is a checklist of elements that call for special attention when using a digital game as a stimulus. It is not exhaustive but considers the key questions typically addressed in the beginning of an experiment. For each question, respectively, we address the parts of the experiment work flow that are most influenced by the decision. These pointers do not imply there is no influence to other parts of the work as well, but merely single out the work tasks that call for extra critical attention.

Checklist question	Why is this important?	Main influence on: GAME CHOICE	EVENT CODING	PARTICIPANT SELECTION	PROCEDURE	ANALYSIS
What tasks does the game play require?	Match research questions and tasks required by the game.	✓	✓		✓	
Is the task represented as game action that is separate from other task types?	Very complex and overlapping events may not allow distinguishing one event from another.	✓	✓		✓	
How does task difficulty influence play? Can task difficulty be balanced?	The difficulty level should be suitable for all participants whether by choosing it properly for the target group, selective recruitment of participants, or by adjusting it case by case.	✓		✓	✓	
What game events repeat themselves?	Frequently repeating game events provide larger sample size for event-based analysis and is necessary for within-subject methods (such as psychophysiology).		✓	✓	✓	✓
Do repeating events occur in a similar context, or does context change?	Adding poorly comparable events only introduces more noise, which blurs results.		✓			✓
How similar as a stimulus is the game across participants?	An identical stimulus across participants is often desirable, but not always necessary.	✓		✓		✓
How much does the player's skill level influence gameplay?	Different backgrounds can result in both factually and subjectively disparate experiences across participants.	✓		✓		✓
What methods of data collection are available?	The research question may be addressed through various different combinations of event coding and data collection.	✓	✓		✓	✓
Does the game provide logs or is external recording needed? Are there developer tools or mod kits that can be customized for data acquisition?	Game logs are extremely useful, if available. The smaller the events you want to examine, the more extensive data logging is required and the higher are the demands for temporal acuity.	✓	✓			✓
How reliably can events be decoded from, e.g., video recordings, keylogs, etc.?	Manually coding can be laborious and may also affect data precision.		✓			✓

Table 1. Checklist

STUDY EXAMPLE AND CONSIDERATIONS

In this section we present an example study to illustrate the use of a game as a stimulus in a psychophysiological experiment. By detailing the rationale behind the choices we made regarding choice of stimulus, event logging, data analysis, etc., we demonstrate how the previously discussed theoretical considerations may be applied in practice. The example is not intended as a canonical solution; the aim of presenting this work is solely to provide the reader with a better estimate of the actual process and the preparatory work required for using games as a stimulus. Indeed, several alternatives exist besides those presented here. Our research unit conducted a commissioned study to examine the benefits of a health drink. The drink is designed to enhance performance during long term performances that call for intense concentration and heavy physical activity. The experiment was conducted to empirically assess whether the test substance would measurably affect performance and concentration, emotional reactions, alertness and stress reactions.

The Choice of Game

To test the effects of a health drink, an activity was needed that would require intense concentration, alertness, and the ability to cope with elevated stress levels over an extended period of time. Some form of built-in performance metric was preferable, as it was considered as the best internally consistent way to assess the task performance. A realistic racing game fills out all these criteria. Playing a challenging racing game consists of several cognitive tasks: fine motor controls and quick reflexes are mandatory, and attention and the ability to quickly change focus are also needed. Longer races require maintaining constant concentration and steady performance throughout the race—the key variables to examine the effects of the test drink.

The game chosen for the experiment was *GTR 2 – FIA GT Racing Game* developed and published by SimBin. *GTR 2* is a realistic sports

car racing simulator for the PC platform (http://www.gtr-game.com). The game is of excellent quality: it has received multiple awards and scores 90/100 on Metacritic. The following sections will detail how we handled the decisions discussed in previous sections, including how we planned for data collection and analysis, our considerations regarding task choice and game settings, and a detailed description of the experiment procedure.

Planning data collection

GTR 2 provides an extensive array of different metrics that can be used to evaluate player performance, which was crucial for this study. For the test, we utilized the *MoTeC i2 Pro* data acquisition system (http://www.motec.com/i2/i2overview/), which is fully compatible with *GTR 2* and also used by real world racing teams. Very few commercially available games provide this much performance data of the game play in an easily accessible way. These metrics logs were combined with self-report questionnaires and psychophysiological measurements. Altogether, these data sources enabled us to thoroughly investigate the players' emotional and physiological state during playing, and to evaluate the test drink's effect on performance and experience. As everything was logged by the stock game and *MoTeC i2 Pro*, no custom made solutions were necessary. We settled for using the computer's clock to synchronize the game logs and psychophysiology as its precision was sufficient though not optimal.

Event coding, data segmentation, and analysis

The high amount of repetition and the relatively low number of random factors in racing games make them good candidates for stimuli in general, and ideal for the type of study we were conducting. Each playing session consists of series of repeating laps, which are clearly demarcated by start and end events. This allowed us to make comparisons between laps and, for example, to monitor the improvement over time. Had we been interested in studying the reactions to various

gaming events, instead of overall levels between conditions, it would have been possible to utilize the exceptionally detailed log files provided by the *MoTeC i2 Pro* system.

The repetition of similar events in a very predictable manner—while typical for racing games—is not prevalent in the vast majority of games. Since we wanted to use the change in performance as a dependent variable, the relative lack of random factors was also of crucial importance. A substantial amount of randomness would make comparisons difficult. In other type of experiments where performance as such is not under scrutiny, randomness might not be as prohibitive. For example, if one were to study reaction times using a digital game, random factors would be acceptable as long as key events repeat often enough.

Difficulty and ensuring similarity

Racing in *GTR 2* is quite demanding. While the difficulty level can be adjusted to suit the skill level of the player, it is still very likely that players will make a number of mistakes that are reflected on the overall lap time. Hypothetically then, if the health drink increases the participants' capability to concentrate over extended periods of time, they should make less mistakes and perform measurably better.

For studying effects on performance, a highly engaging activity was desirable, as an extreme setting was more likely to bring out the differences between conditions. As an activity, playing games is engaging and strongly focuses the players' concentration on the game and playing in a natural manner. A good racing game pushes the participants to the sector where they are really doing their best and trying to perform as well as they can. This is especially true for any sports game that has a built-in competition structure. Therefore a racing game was quite appropriate for this particular experiment. The participants were also motivated to perform as well as they could by rewarding the top three fastest drivers of all participants. In effect, they were not only

racing against the computer, but against other participants, and for a considerable reward.

We decided to control the difficulty level so that all participants used the same settings. In general, this gave an advantage to experienced players. Since the situation was framed as competition, players' emotions would likely relate to their skill level, as responses would vary according to the level of performance. In this case, we chose to prioritize task similarity, to increase the comparability of tasks among subjects. If the studied effects had been something other than performance (say, whether the test drink affected emotional states), then the choice would have been to rather control performance by evening out skill differences with appropriate difficulty settings to suit each player's skill level.

Experiment Procedure Considerations

The experimental procedure must be adjusted to accommodate the unique features of digital games. Incorporating a training phase to get participants acquainted with the game and the controls is often needed. If performance is measured, training sessions can also be used to even out minor skill differences between participants beforehand. As with all stimuli, randomizing playing order helps avoiding systematic errors.

Circuits in racing games are of different length and a lap can take considerably longer on one circuit than on another. In the example study, we chose four different circuits of roughly equal length. Within each circuit, laps form the repeating events that are analyzed using lap times as a central performance metric. Confounding effects on performance (such as learning effect, in which players learn and play better at the end of the experiment than in the beginning) were mitigated by employing a within-subject design, randomizing the playing order of various race circuits, and incorporating a training session into pre-ex-

periment procedures.

To enforce similar starting conditions for the race across all participants, in-game practice and qualifying sessions provided in-game were skipped and participants started the race from the back of the grid. The race length was adjusted to 25 minutes, difficulty level to novice and opponent strength to 90%. This configuration was estimated prior to the experiments as providing a suitable average challenge level across the recruited participants. All participants drove the same car, with identical car and game settings. Automatic gears were used to avoid amplifying the skill level differences between subjects. *GTR 2* offers numerous settings for adjusting both game play and the car. We decided to control all of these and not let participants adjust anything. By enforcing certain settings we aimed at maximizing stimulus similarity across the participants and simplifying analysis by cutting down the number of variables. While this makes the experience less ecologically valid (McMahan 2011), we were not investigating the experience per se but were using the game to create a high-performance challenge. In this case, the tradeoff in ecological validity is both acceptable and necessary in order to control the further advantage more experienced players would have gained, had they been allowed to play with their preferred settings.

Conclusions

Games have already proved useful beyond their function as entertainment. Among others, they serve as a great resource for research by providing realistic, familiar, and yet relatively complex and diverse stimuli for experiments. However, the same features that makegames promising stimuli also make them particularly challenging to use in controlled experiments. Many of these challenges can be overcome by taking into account the special nature of digital games when designing the test setup, procedure, and data analysis. Nevertheless, the use of games calls for methodological balancing acts such as making complex

decisions regarding benefits and tradeoffs of practical decisions, and anticipating the effects of potential confounding factors. The added complexity to the experimental setting calls for particular care whenever games are used as stimulus. High attention to detail is also recommended when analyzing, communicating, and interpreting study results.

This work is primarily based on practical experience and documented know-how on experiment design accumulated in our lab over the last 10 years. We identify the following four key steps in the process of preparing a study using digital games as stimuli: (1) matching and regulating task type, (2) determining data segmentation and event coding, (3) ensuring compatibility between participants and (4) planning and conducting data collection. Each of these factors has potential effects on experiment validity and reliability that should be considered carefully when designing and conducting the study. The ideas presented here are based on a very rigorous type of study design but that does not limit its utility for less controlled experiments. On the contrary, scholars preparing studies with more flexible design will find the checklist useful for deciding which elements they will want to control, even if they decide to leave some other variables open. Currently in game research—and also in other fields using games as a stimulus—the multitude of procedures makes it difficult to draw conclusions from research conducted by others. If the studies use vastly different procedures or very dissimilar levels of stimulus control, results cannot be reasonably compared. This not only slows down the accumulation of knowledge, but may confuse readers less familiar with games and the pitfalls involved in using games as a stimulus. The present work takes steps towards a more systematic and better documented procedure for how to conduct studies using games. The discussion presented in this paper is primarily directed as a practical guide for planning and conducting experiments. Nonetheless, the information provided here also offers material for readers wishing to interpret or

evaluate the works of others.

Endnotes

[1] Digital games means all games played on digital devices, from game consoles to desktop computers and modern mobile devices.

[2] Ecological validity refers to how closely various aspects of an experimental setup such as stimulus, task, setting etc. correspond to real life context.

[3] http://www.gamespot.com/, http://www.gamezone.com/, http://www.ign.com/, http://www.metacritic.com/, http://www.gamerankings.com/

References

Adachi, P.J.C. and T. Willoughby. "The Effect of Video Game Competition and Violence on Aggressive Behavior: Which Characteristic Has the Greatest Influence?" In *Psychology of Violence*, vol. 1, no. 4 (2011): 259-274.

Avalanche Software & Ritual Entertainment. (2006). 25 to Life. Eidos Interactive.

Ballard, M. and J. Wiest. "Mortal Kombat (tm): The Effects of Violent Videogame Play on Males' Hostility and Cardiovascular Responding." In *Journal of Applied Social Psychology*, vol. 26, no. 8 (1996): 717–730.

Björk, S. and J. Holopainen. *Patterns In Game Design*. Cengage Learning, 2005.

Bushman, B.J. and C.A. Anderson. "Violent video games and hostile expectations: a test of the general aggression model." In *Personality and Social Psychology Bulletin*, vol. 28, no. 12 (2002): 1679-1686.

Cacioppo, J.T., L.G. Tassinary and G.G. Berntson. *Handbook of psychophysiology.* 3rd ed. New York, NY: Cambridge University Press, 2007.

Caillois, R. *Man, Play and Games*. University of Illinois Press, 2001.

Cowley, B., T. Heikura, and N. Ravaja. “A Study of Learning Effects in a Serious Game Activity.” In *Computers & Education*, 2012.

Deutsch, M. and R.M. Krauss. “The effect of threat upon interpersonal bargaining.” In *Journal of Abnormal and Social Psychology,* 61 (1960): 181-189.

Ekman, I., K. Kallinen, and N. Ravaja. “Detection and identification of vibrotactile stimulation in stressful conditions.” In *European Workshop on Imagery and Cognition* (EWIC2010), Helsinki, Finland, 2010.

Elverdam, C. and E. Aarseth. “Game Classification and Game Design Construction Through Critical Analysis.” In *Games and Culture* vol. 2 no. 1 (2007): 3–22.

Epic Games & Digital Extremes (2002). *Unreal Tournament 2003*. Atari Inc.

ESA - Entertainment Software Association. “Essential facts about the computer and video game industry.” 2011.

Fehr, E. and S. Gächter. “Altruistic punishment in humans.” In *Nature* Vol. 415 (2002): 137-140.

Frey, A., J. Hartig., A. Ketzel, A. Zinkernagel. and H. Moosbrugger. “The use of virtual environments based on a modification of the computer game Quake III Arena in psychological experimenting.” In *Computers in Human Behavior* vol. 23, no. 4 (2007): 2026-2039.

Jones B., M. Steele, J. Gahagan and J. Tedeschi. “Matrix values and cooperative behavior in the Prisoner's Dilemma game.” In *Journal of Personality and Social Psychology* vol. 8, no. 2 (1968): 148-53.

Järvelä, S., J.M. Kivikangas, J. Kätsyri, and N. Ravaja. "Physiological linkage of dyadic gaming experience." In *Simulation & Gaming* (n.d.). Manuscript accepted for publication.

Karvinen, J. and F. Mäyrä, “Pelaajabarometri 2011 – Pelaamisen Muutos.” TRIM Research Reports 6. Tampere, Finland: University of Tampere, 2011.

Kivikangas, J.M., G. Chanel, B. Cowley, I. Ekman, M. Salminen, S. Järvelä. and N. Ravaja. "Review on psychophysiological methods in game research." In *Journal of Gaming and Virtual Worlds*, vol. 3, no. 3 (2011): 181-199.

Kosunen, I. "Clustering psychophysiological data with mixtures of generalized linear models." Master's thesis, 2011.

Lin, S.-F., "Effect of Opponent Type on Moral Emotions and Responses to Video Game Play." In *Cyberpsychology, Behavior, and Social Networking*, vol. 14, no. 11 (2011).

Lindley, C. A. "Game Taxonomies: A High Level Framework for Game Analysis and Design." In *Gamasutra*, 2003.

Lundgren, S. and S. Björk. "Game Mechanics: Describing Computer-augmented Games in Terms of Interaction." In *Proceedings of TIDSE*, 3, 2003.

Manninen, T. "Rich Interaction Model for Game and Virtual Environment Design." University of Oulu, 2004.

McMahan, R.P., E.D. Ragan, A. Leal, R.J. Beaton and D.A. Bowman. "Considerations for the use of commercial video games in controlled experiments." In *Entertainment Computing* (2011).

Milgram, S. "Behavioral Study of Obedience." In *Journal of Abnormal and Social Psychology* vol. 67 (1963): 371–378.

Mueller, F.F., M.R. Gibbs, and F. Veter. "Taxonomy of Exertion Games." In *Proceedings of the 20th Australasian Conference on Computer-Human Interaction: Designing for Habitus and Habitat,* 263–266. OZCHI '08. New York, NY, USA:ACM 2008.

Pazhitnov, A. *Tetris*. Spectrum Holobyte, 1985.

Probe Entertainment. *Mortal Kombat* ™. [Sega Genesis]. Acclaim, 1993.

Ravaja, N., T. Saari, M. Turpeinen, J. Laarni, M. Salminen, and M. Kivikangas. "Spatial presence and emotions during video game playing: Does it matter with whom you play?" In *Presence* vol. 15 (2006): 381–392.

Ravaja, N., J. Laarni, T. Saari K. Kallinen, M. Salminen, J. Holopain-

en, and A. Järvinen. ”Spatial presence and emotional responses to success in a video game: A psychophysiological study” in M. Alcañaz Raya and B. Rey Solaz (eds), *Proceedings of the PRESENCE 2004* (2004) pp. 112–116. Valencia, Spain: Editorial de la UPV.

Ravaja, N., T. Saari, M. Salminen, J. Laarni and K. Kallinen. “Phasic Emotional Reactions to Video Game Events: A Psychophysiological Investigation.” In *Media Psychology,* vol. 8 no. 4 (2006): 343-367.

Ravaja, N., M. Turpeinen, T. Saari, S. Puttonen, and L. Keltikangas-Järvinen. “The psychophysiology of James Bond: phasic emotional responses to violent video game events.” In *Emotion*, vol. 8 no.1 (2008): 114-120.

Schofield, D. “Playing with evidence: Using video games in the courtroom.” In *Entertainment Computing*, vol. 2, no. 1 (2011): 47-58.

Staude-Müller, F., T. Bliesener, T. and S. Luthman. "Hostile and hardened? An experimental study on (de-) sensitization to violence and suffering through playing video games." In *Swiss Journal of Psychology* vol. 67 no. 1 (2008): 41–50.

SimBin Development Team AB. *GTR 2 – FIA Racing Game* [PC] SimBin Development Team AB, 2006.

Slater, M., A. Antley, A. Davison, D. Swapp, C. Guger, C. Barker, N. Pistrang and M.V. Sanchez-Vives. “A Virtual Reprise of the Stanley Milgram Obedience Experiments” *PLoS ONE 1*, (2003).

Turtle Rock Studios. *Left 4 Dead*. Valve Corporation, 2008.

Washburn, D. “The games psychologists play (and the data they provide).” In *Behavior Research Methods* vol. 35 (2003): 185-193.

Whitton, N. *Learning with Digital Games: A Practical Guide to Engage Students in Higher Education*. Taylor & Francis, 2009.