Digital games as experiment stimulus

Simo Järvelä
School of Economics, Aalto University
P.O. Box 21255
00076 Aalto, Finland
+358405062714
simo.jarvela2@aalto.fi

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

J. Matias Kivikangas
School of Economics, Aalto University
P.O. Box 21255
00076 Aalto, Finland
matias.kivikangas@aalto.fi

Niklas Ravaja
Department of Social Research, University of Helsinki
P.O. Box 9
00014 University of Helsinki, Finland

School of Economics, Aalto University
P. O. Box 21220
00076 Aalto, Finland

HIIT, Aalto University
P.O. Box 19215
00076 Aalto, Finland

ABSTRACT

Digital games offer rich media content and engaging action, accessible individually or in groups collaborating or competing against each other. This makes them promising for use as stimulus in research settings. This paper examines the advantages and challenges of using games in experimental research with particular focus on strict stimulus control through the following four areas: (1) matching and regulating task type, (2) data segmentation and event coding, (3) compatibility between participants and (4) planning and conducting data collection. This contribution provides a breakdown of the steps necessary for using a digital game in experimental studies, along with a checklist for researchers illustrating variables that potentially affect the reliability and validity of experiments. We also offer a practical study example. Ideally, the identification of the methodological and practical considerations of employing games in empirical research

Proceedings of DiGRA Nordic 2012 Conference: Local and Global – Games in Culture and Society.

will also provide useful in interpreting and evaluating experimental work utilizing games as stimulus.

Keywords
digital games, stimulus, experimental psychology, methodology

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 to decision making. Game-like procedure setups are classic to experimental psychology: for example, Deutsch & Krauss's *Trucking Game* (1960) and the *Prisoner's Dilemma* (Jones et al. 1968). Contemporary psychology research is also utilizing 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 utilizing digital games in experiment setups: using games as stimulus to study other forms of behavior, using games to manipulate variables, games providing education and instruction, and gaming as a performance metric. In addition to using games as stimulus for psychological experiments, there is a body of research striving to understand games per se, evaluating design decisions, and measuring the effects of playing. While these studies are not the focus of this paper, these lines of research similarly require certain rigor in the setup of experiments involving games as stimulus.

As of yet, there exists little instruction how to choose digital games for psychological experiments. Also, the field lacks guidelines regarding the experiment setup with games, and work relies on accumulated know-how. This presents challenges especially when comparing findings between studies and in generalizing the results. These difficulties will likely become even more pertinent as interest towards games are spreading 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 stimulus. They present criteria for game selection and game mode selection, and mention the importance of controlling participant demography. They also briefly comment on the importance of managing confounds during gameplay, but consider 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 stimulus in our lab during nearly a decade, in combination with psychophysiological methods (Cacioppo et al. 2007) to study the gaming experience itself (Kivikangas et al., 2011; Ravaja et al. 2004, 2006a, 2006b, 2008), but also used games to access other processes, such as learning (Cowley et al. 2012; Cowley & Ravaja, 2012), 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 from the course of these experiments and the considerations we have found, sometimes by trial and error, to be pertinent for using games as 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, approaches to controlling relevant experimental variables, and describe the practical issues involved in setting up an experiment utilizing a commercially available game title. The detailed discussion will be valuable both to researchers who wish to utilize games in similar strictly controlled studies as well as for more forgiving setups. In addition, readers interested in the results of game-related research will find this paper useful when evaluating published studies, the possible pitfalls in the experimental setup, and ultimately the generalizability and relevance of the results.

GENERAL CONSIDERATIONS FOR CHOOSING DIGITAL GAMES AS STIMULI

Digital games are a natural choice for stimulus, not only when studying gaming and gaming experience, but also for other research questions calling for an engaging, yet challenging activity (Washburn, 2003). Digital games, modern games especially, are very complex stimuli and they are in many ways a unique form of media. There is a huge amount 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 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. A study in Finland (where we recruit most of our test subjects) shows that 54% of respondents were active video gamers. Non-digital games included, 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 them more approachable than abstract psychological tasks, which helps in recruiting participants. Second, the high familiarity allows using more complex tasks that would be very difficult if used as abstract psychological tasks. 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 to cause certain reactions within the player, they are very ecologically valid instruments for eliciting emotions for various purposes. Different game genres typically address different emotions, e.g. horror games aim for quite a different emotional reactions and mood than e.g. racing or educational games. With the proper selection of games, a broad scale of emotions can be elicited. However, as games most often do not focus on a single emotion, experimental design and testing the stimulus beforehand should be done carefully to determine what games and which parts of them can be used, and how, to produce the wanted stimulus but not others.

Furthermore, digital games provide safe virtual environment to conduct studies on topics which could otherwise be deemed as unethical or not safe for the participants. The level of realism in games is high enough that already games and virtual environments are 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 similar experiments can be conducted using virtual game-like activity (Slater et al. 2003). In addition, as McMahan

and colleagues (2011) state, using off-the-shelf games provides benefits of quick implementation, avoiding some researcher bias, and 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. Also, with commercial games adequate data capture might be challenging.

Similarity of stimulus

A major challenge with games is that various factors affect what the actual content of the game is, while in experimental research it would be preferable to use as identical stimulus across the participants as possible. Interactive stimulus is never the same, but changes according to participant actions. 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 reflect in changes in the *content* of the game, for example as the game progresses, a skilled player will likely use more diverse and effective playing styles, or have an access to more advanced game items than a less experienced player. In addition, narrative games bring another consideration, as player reactions are widely different whether they have played the particular game beforehand and are familiar with the story or not.

Therefore it is of utmost importance that the researcher is well aware of the dependent variable and how it might be influenced by elements that do not hold from participant to participant. The choice of what game is used must be done so that the stimulus is identical *in the relevant aspects to the dependent variable*. After that, other variance in the game can be considered irrelevant for the experiment, but it is good to note that they still contribute to the attractiveness of the game for the participant. It would be a mistake on part of the researcher to strip the game from all irrelevant variance, as this makes the game just another psychological task without the positive qualities games can offer. Furthermore, it is important to acknowledge that as game research is still a young field and therefore the researchers are not likely to know what are all the relevant aspects, acquiring a large enough sample size mitigates at least some of the problems that may arise.

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, for example, removing some unsuitable elements. Developer tools and mod kits make some adjustments possible, but major changes come with a risk of compromising game quality. The closed system of many commercial games can also make it difficult in some cases 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 programming level). In some cases open source alternatives are practical for this particular reason. Game logs can be later used on event based analysis, segmentation, performance appraisals etc. This is

not only a question of convenience, as some game manipulations can be impossible to spot from recordings of game play. Still, at least a screen capture recording of the game play should be recorded.

Despite the rich stimuli offered by commercial games, it is not uncommon for researchers to develop their own game for the experiment so that they 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 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 stimuli: when the games are engaging, the participants focus deeply on the task at hand and are more likely to act more as they would outside an experiment, and feel less distracted by the experimental setup. Thus, more engaging stimuli can produce better data.

PRACTICAL AND METHODOLOGICAL CONSIDERATIONS

Besides general considerations why to use digital games as 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 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 games with suitable challenge and action for the research in question is one of the first steps to designing the study. Gameplay consist of various tasks that define what type of stimulus the game actually is. What cognitive tasks are involved is one way of approaching the question; 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 differently on various cognitive tasks. Game tasks need to be considered in relation to the context — the same task, but e.g. with different time limitations will produce vastly different reactions. Intense repetition and extended task times can also change the nature of a task significantly from how they are perceived in shorter durations of play. For example, both Tetris and a modern first-person shooter game might be appropriate stimuli 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 about 3D spatial processing, emotional content from the narrative, and so on.

Naturally the game should be chosen according to what type of stimulus is preferable. In fact, choosing a game title is only part of the task of determining experiment stimulus. The choice of stimulus goes down into choosing levels, playing modes, and narrowing out tasks that are conductive to the intended research. The task structure and difficulty is an important part of defining the game's uses as a stimulus in an experiment and influence the whole procedure setup, as well as data analysis.

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). It would be of utmost importance to make sure that games compared in such study would not differ in quality, the pace of the game, or how engaging the game was, for instance, which might all affect arousal as well. Sadly, this has not always been taken care of. As an example how task and game type and research questions can be structured, see the example study provided below.

Reviews, ratings and genre classifications (for online reviews and rankings, see for example: GameSpot, GameZone, IGN, Metacritic, GameRankings)[2] can be helpful in choosing the game. The ratings give an overall assessment on the quality of the game, which, while not objective, is not dependant on the 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. 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, controls etc. 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, is there enough or too much repetition), how static the action is (is the difficulty level static or does it, for example, increase). 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 may undermine the ecological validity of the gaming scenario). In games that have narration it is also important to consider whether the chosen part of the game serves its function without experiencing the narration before it.

- Define your tasks, find out what can be expected to affect it, 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 influence your task inadvertently.
- Use available reviews to pinpoint effects, challenges and possible shortfalls in the game design, and 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 in creating desired variation.

Determining Data Quantification and Event Coding

To be able to analyze effects while gaming, researchers need to come up with ways to quantify the data. One possibility is of course, to use a block design, for example comparing different levels, games, or game modes against each other. However, sometimes we are interested in smaller events, such as particular actions. The choice of event coding is based not only on the game's available actions, but also on how isolated they occur during gameplay. With many different elements affecting the subject at the same time, it can be impossible to say which of the elements caused a certain reaction or behavior (e.g., in a combat during an action game). On the other hand, if events are too

unique, the sample might not be adequate for statistical analysis unless it is compensated with a high number of participants.

Also, the same repeating event can occur in different contexts thus framing it differently and so having a different meaning within the game. 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 have large enough sample of the same event so that the effect of random noise is balanced out. Naturally these considerations affect also game choice, as games where the same type of event takes place several times are more suitable stimuli as it is easier to have a satisfying sample size of various events. The optimal time scale needed for events has to be balanced in relation to the metrics used in the experiment. As an example, the psychophysiological method (Cacioppo et al. 2007) allows accessing precise events, as the data is gathered continuously, although there is still variance between different measures: some physiological reactions take place immediately (e.g. facial muscles activation that can be measured with EMG) and others more slowly (e.g. electrodermal activity where reactions can be seen a few seconds after the event and lasting several seconds). The choice of method to analyze the data can to some extent mitigate the challenge provided by concurrent and overlapping events. The Linear mixed model incorporates both fixed effects and random effects, and is particularly suitable to repeated measurements produced by psychophysiology. Also known as Hierarchical linear models, this statistical method is necessary if the data is hierarchical (e.g., events within conditions within participants) or the number of samples is not fixed within the level (e.g., if a particular event occurs only once for some, and several times for other players). For other methods than psychophysiology, a different event structure and analysis methods are in order.

When deciding the event coding, it is useful to remember that one can always go from specific to general, but rarely the other way (without recoding the data). Individual events can always later be considered either as separate data points, or transformed into block-level values. Event logs also offer an additional way to evaluate how similar gaming sessions were between players, and thus provide useful for evaluating the reliability of the study.

- 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.
- Code too much rather than too little detail. Extra coding can always be disregarded later, but uncoded is much more difficult to code afterwards.

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

Choosing only subjects that are experienced enough can ensure deeper skill levels during the experiment than including a practice session or giving instructions prior to the test session. Usually at least some experience with digital games within the same game type is often preferable as otherwise learning the basic skills would require too much time and effort in the experiment situation. If no such time is given, the lack of basic gaming skills would make the results non-comparable to others. Importantly, gaming skills are not necessarily transferable across genre borders, and even within a certain genre small changes in e.g. controller behavior can have a major impact on play performance (albeit some of them can be remapped to match participant's preferences). It is often advisable to gather comprehensive background data on gaming experience, including genre preference and even game titles, if the chosen game is known to deviate from the genre norms in some relevant ways (such as controller scheme). On the other hand, experience might also diminish the value of the game in the study: if a participant has played a game to the point of getting tired of it, benefits of engaging play might be lost.

Theoretically, a large enough random sample of males and females is the best for generalizing the results over the general population and avoiding a gender bias. However, in practice this is often problematic. Although there are many women that play digital games, gaming is still much more common among the male population (ESA, 2011; Karvinen & Mäyrä, 2011), and therefore acquiring equal numbers of both sexes with good sample size might be very 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. In turn, these have to be taken into account when analyzing the data, interpreting the results and generalizing them.

Comparable stimulus

In practice it is impossible to ensure that the stimulus is equivalent to all participants. It is important to identify the critical factors affecting the dependent variable, and control those as well as possible. A major factor to observe is game difficulty, which is as much a function of the player as of the game. Some games have built-in automatically balancing difficulty adjustments that change the difficulty of the game according to player's performance and choices within the game. Such a system can be extremely useful in creating relatively equally challenging gaming experience to players of varying skill levels, however, it can also be detrimental to the idea of using the same content for all participants, if that is critical for the experiment. Self-adjusting difficulty levels can, depending on the context and what is being studied, either escalate this problem or for some part counterbalance it. Also, in many cases it is difficult or impossible to know if such a system exists in a game, or how exactly the system works. If there is no information available concerning this from the developer etc., detecting it requires considerable familiarity with digital games. However, in some cases variation in game content is not a problem, for example when measuring 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. 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 game 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 - is 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

The information what happened in the game and when must be obtained somehow so that segmentation or event based analyses are possible. In such cases log files of gameplay or other ways to segment the data with sufficient temporal accuracy are crucial. Game logs are without doubt the easiest way to determine what was actually happening in the game at a precise moment, providing the optimal source of data for event-based analysis. Most games do not employ sufficient logging of game events available for the researcher, and they have to be coded afterwards by reviewing recorded game-play (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 from 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. If available, game developer kits are particularly useful for setting up experiments. Many games also have mod kits, either commercially produced or fan-made that can be used to add event logging. If a built-in logging system is not feasible, some logs can be collected externally. Keyloggers, screen capture videos, mouse-click recorders and such can be helpful. Most games have one or more innate performance metrics in them, available and visible to the player. 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.

When planning logging of data, and especially if the analysis will operate on event data instead of whole blocks, it is imperative to calibrate the timestamps of different data sources. Whereas some game events can be matched manually afterwards, other data sets — like psychophysiological signals — are nearly useless to the analysis if the timestamps do not correspond. Timestamps can be anchored across devices by, for example, filming the moment of turning on measurement devices.

- 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 timestamps across data sources.

Checklist of Questions

The following checklist lists elements that call for special attention when using a digital game as research stimulus. It is not exhaustive, but considers the key questions typically addressed in the beginning of an experiment intending to use games as stimuli. For each question, respectively, we addresses the part(s) of the experiment workflow that is most influenced by the said question. This does not imply there is no influence to other parts of the work as well, instead it points out the work tasks calling 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 when choosing the game.	x	x		x	
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.	x	x		x	
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.	x		x	x	
What game events repeat themselves?	Frequently repeating game events provide larger sample size for event-based analysis and is necessary for within-subject		x	x	x	x

	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.			x				x
How similar as a stimulus is the game across participants?	Identical stimulus across participants is often desirable, but not always necessary.		x			x		x
How much does the player's skill level influence gameplay?	Different backgrounds can result in both factually and subjectively disparate experiences across participants.		x			x		x
What methods of data collection are available?	The research question may be addressed through various different combinations of event coding and data collection.		x		x		x	x
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 events you		x		x			x

	want to examine, the more extensive data logging is required and the higher are demands for temporal acquity.		
How reliably can events be decoded from e.g. video recordings, keylogs, etc.?		x	x
	Manually coding can be laborious, but may also affect data precision.		

STUDY EXAMPLE AND CONSIDERATIONS

In this section we present an example study to illustrate the use of a game as stimulus in a psychophysiological experiment. By detailing the rationale behind the choices we made regarding choice of stimulus, event logging, data analysis etc., we provide an example of how the previously discussed theoretical considerations are applied in practical work. We hope that this example will provide the reader with a better estimate of the actual process, and the preparatory work required for using games as a stimuli. The example is not intended as a canonical solution, but as a grounding example. Indeed, several alternatives exist besides those presented here.

Our research unit conducted a commissioned study to examine the mental effects of consuming a health drink. The health 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 assesses 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 an health drink, an activity that would require intense concentration, alertness and coping with elevated stress levels over an extended period of time was needed. Some built-in performance metrics, to provide internally consistent way to assess the effects was strongly preferred. A realistic racing game fills out all the 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 is 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 has received multiple awards and has Metacritic Metascore at 90/100. Next we will go through the various aspects of this choice, and using a racing game as stimulus in a experiment in general, reflecting the issues discussed on sections 2 and 3.

Planning data collection

GTR 2, especially with additional Motec i2 Pro data acquisition system (http://www.motec.com/i2/i2overview/; a performance testing software also actually used by real world racing teams), provides an extensive array of different metrics that can be used to evaluate player performance. Very few commercially available games provide this much performance data of the game play in an easily accessible way. Those metrics combined with self-report questionnaires and psychophysiological measurements enabled us to thoroughly investigate the players' emotional and physiological state during playing, and to evaluate the test drinks effect on performance and experience. No external logging systems, custom made solutions or laborious video recording analysis was necessary as everything was logged by the stock game and Motec i2 Pro.

Event coding, data segmentation and analysis

In a racing game, each playing session consists of several repeating events, i.e. laps, which then can be compared to each other and see the possible improvement over time. This repetition of similar events in a very predictable manner, while typical for racing games, is not prevalent in vast majority of games. Therefore, a lot of repetition and relatively few random factors make racing games potentially good candidates for stimuli in general and ideal for this type of study.

In this study, the relative lack of random factors is also crucial as the change in performance is one of the main factors under scrutiny and 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 studying reaction times using a digital game, random factors in content are acceptable as long as the key events are repeated to the required extent.

Difficulty, 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 some amount of mistakes will be made during game play and those mistakes by the nature of the game instantly affect 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 therefore perform better.

As an activity, playing games is engaging and strongly focuses player's concentration to the game and playing in natural manner. This is desirable in experimental setting as it pushes the participants to the sector where they are really doing their best and trying to perform as well as they can which brings out the differences between conditions more easily. This is especially true for any sports game that has built-in competition structure and therefore racing as an activity was suitable 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. So, 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, to increase the comparability among subjects. In general, this means that often the more experienced players will perform better, and in competitive situation their emotions will vary according to their performance. If the studied effects would have been different (say, if the test drink would have made claims about the emotional state), the choice would

have been to control the performance by using the difficulty settings to even out skill differences.

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 in question and the controls is often needed. If performance is measured, training sessions can be also used to even out the differences between participants beforehand. As with all stimuli, randomizing playing order helps to avoid systematic errors.

The effect of confounding variables affecting performance (such as learning effect, according to 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 circuits, and incorporating a training session into pre-experiment procedures. In racing games circuits are of different length and one lap can take considerably longer on one circuit than on another. As lap times are one performance metric and laps are repeating events that could be analyzed separately, we chose four different circuits of roughly equal length.

Practice and qualifying sessions were skipped and participants started the race from the back of the grid so that they would all start the race under the same conditions. The race length was adjusted to 25 minutes, difficulty level to novice, and opponent strength to 90%. This was approximated to be the suitable average challenge level across the participants. All participants raced using the same car, with identical car and game settings. Automatic gears were used to avoid amplifying the skill level differences between subjects. In *GTR 2* there are huge number of settings you can adjust both regarding the 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 the similarity of the stimulus across the participants to make analyzing of the results easy by cutting down the number of variables. This makes the experience less ecologically valid (McMahan, 2011), but as we were not investigating the experience per se, it was more important to control the further advantage more experienced players would have had with preferred settings.

CONCLUSIONS

Games have already proved useful beyond their function as entertainment. Among others, they serve a great resource for various types of research, by providing realistic, familiar, and yet, relatively complex and diverse form of stimulus. The same features that make them very promising as stimulus make them particularly challenging to use in controlled experiments. The challenges can be overcome by taking into account the special nature of digital games when designing the test setup, procedure, and data analysis.

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 are examined for the potential effects they impose on the experiment validity and reliability, along with examples of how these considerations reflect in practice when preparing and conducting studies. The fact that the work is based on a very rigorous form of study 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 both game research (and research in other fields using games as stimulus) the multitude of procedures make 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 compared meaningfully. 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 stimuli. This work takes steps towards a more systematical and better documented procedure how to conduct studies using games. The discussion presented in this paper can be used as a practical guide for those planning such experiments. Finally, the same knowledge concrete points of scrutiny for those who try to evaluate the work done by others.

ENDNOTES

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

2 http://www.gamespot.com/, http://www.gamezone.com/, http://www.ign.com/, http://www.metacritic.com/, http://www.gamerankings.com/

BIBLIOGRAPHY

Bushman, B.J. & Anderson, C.A., 2002. Violent video games and hostile expectations: a test of the general aggression model. *Personality and Social Psychology Bulletin, 28 (12)*, 1679-1686.

Cacioppo, J.T., Tassinary, L.G., & Berntson, G.G., 20007. *Handbook of psychophysiology 3rd ed*. New York: Cambridge University Press.

Cowley, B., Heikura, T., & Ravaja, N., 2012 (manuscript submitted for publication). A Study of Learning Effects in a Serious Game Activity. *Computers & Education*.

Cowley, B., & Ravaja, N., 2012 (manuscript submitted for publication). Cardiovascular Physiology Predicts Learning Effects in a Serious Game Activity. *Computers & Education*.

Deutsch, M. & Krauss, R.M., 1960. The effect of threat upon interpersonal bargaining. J*ournal of Abnormal and Social Psychology*, 61, 181-189.

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

ESA - Entertainment Software Association, 2011. *Essential facts about the computer and video game industry*. Available online: http://www.theesa.com/facts/pdfs/ESA_EF_2011.pdf [Accessed 19.2.2012]

Fehr, E. & Gächter, S., 2002. Altruistic punishment in humans. *Nature* 415, 137-140.

Frey, A.,Hartig, J., Ketzel, A., Zinkernagel, A.& Moosbrugger, H., 2007. The use of virtual environments based on a modification of the computer game Quake III Arena in psychological experimenting. *Computers in Human Behavavior* 23(4), 2026-2039.

Jones B., Steele M., Gahagan J., Tedeschi J., 1968. Matrix values and cooperative behavior in the Prisoner's Dilemma game. *Journal of Personality and Social Psychology*, 8(2), 148-53.
Karvinen, J. & Mäyrä, F., 2011. Pelaajabarometri 2011 – Pelaamisen muutos. TRIM Research Reports 6. Tampere: University of Tampere.
Kivikangas, J.M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., Ravaja, N., 2011. Review on psychophysiological methods in game research. *Journal of Gaming and Virtual Worlds*, 3(3), 181-199.
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*.
Milgram, S., 1963. Behavioral Study of Obedience. *Journal of Abnormal and Social Psychology* 67, 371–378.
Ravaja, N., Saari, T., Turpeinen, M., Laarni, J., Salminen, M., & Kivikangas, M., 2006. Spatial presence and emotions during video game playing: Does it matter with whom you play? *Presence, 15*, 381–392.
Ravaja, N., Laarni, J., Saari, T., Kallinen, K., Salminen, M., Holopainen, J., & Järvinen, A., 2004. Spatial presence and emotional responses to success in a video game: A psychophysiological study. In M. Alcañaz Raya & B. Rey Solaz (Eds.), *Proceedings of the PRESENCE 2004* (Vol. 2004, pp. 112–116). Valencia, Spain: Editorial de la UPV.
Ravaja, N., Saari, T., Salminen, M., Laarni, J., & Kallinen, K., 2006. Phasic Emotional Reactions to Video Game Events: A Psychophysiological Investigation. *Media Psychology*, *8*(4), 343-367.
Ravaja, N., Turpeinen, M., Saari, T., Puttonen, S., & Keltikangas-Järvinen, L., 2008. The psychophysiology of James Bond: phasic emotional responses to violent video game events. Emotion, 8(1), 114-20.
Schofield, D., 2011. Playing with evidence: Using video games in the courtroom. *Entertainment Computing, Vol 2* (1), 47-58.
SimBin Development Team AB, 2006. *GTR 2 – FIA Racing Game*. PC. SimBin Development Team AB.
Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., Pistrang, N., Sanchez-Vives, M.V., 2003. A virtual reprise of the stanley milgram obedience experiments. PLoS ONE 1.
Washburn, D., 2003. The games psychologists play (and the data they provide). *Behavior Research Methods*, 35, 185-193.