Empirical Data and Results of the Construct Vadility Studies

Ravaja, N.; de Kort, Y.A.W.; IJsselsteijn, W.A.; Klimmt, C.; Lindley, C.; Mathiak, K.; Turpeinen, M.

Published: 01/01/2007

Document Version
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the author's version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 29. Dec. 2018
028765
FUGA
The fun of gaming:
Measuring the human experience of media enjoyment

STREP / NEST-PATH
Deliverable D5.1:
EMPIRICAL DATA AND RESULTS OF THE CONSTRUCT
VALIDITY STUDIES

<table>
<thead>
<tr>
<th>Title of Contract</th>
<th>FUGA - The fun of gaming: Measuring the human experience of media enjoyment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acronym</td>
<td>FUGA</td>
</tr>
<tr>
<td>Contract Number</td>
<td>NEST-PATH-028765</td>
</tr>
<tr>
<td>Start date of the project</td>
<td>01.05.2006</td>
</tr>
<tr>
<td>Duration</td>
<td>36 months, until 30.4.2009</td>
</tr>
<tr>
<td>Reporting period</td>
<td>January 2007 – February 2008</td>
</tr>
<tr>
<td>Date of preparation</td>
<td>28.2.2008</td>
</tr>
<tr>
<td>Project coordinator</td>
<td>Name: Niklas Ravaja</td>
</tr>
<tr>
<td></td>
<td>Organization: Helsinki School of Economics</td>
</tr>
<tr>
<td></td>
<td>Phone: +358 40 411 4565</td>
</tr>
<tr>
<td></td>
<td>Fax: +358 9 431 38391</td>
</tr>
<tr>
<td></td>
<td>E-Mail: <a href="mailto:niklas.ravaja@hse.fi">niklas.ravaja@hse.fi</a></td>
</tr>
<tr>
<td></td>
<td>Project web site: <a href="http://www.hse.fi/fuga">http://www.hse.fi/fuga</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Partners:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partner 1 (CKIR)</td>
</tr>
<tr>
<td>Partner 2 (TKK)</td>
</tr>
<tr>
<td>Partner 3 (HGO)</td>
</tr>
<tr>
<td>Partner 4 (HMTH)</td>
</tr>
<tr>
<td>Partner 5 (UKA)</td>
</tr>
<tr>
<td>Partner 6 (TUE)</td>
</tr>
</tbody>
</table>

Project funded by the European Community under the FP6 New and Emerging Science and Technology (NEST) programme
# Table of Contents

1 General introduction

2 Results of construct validity studies by partner

  2.1 CKIR

    2.1.1 Construct validation of phasic psychophysiological responses
    2.1.2 Construct validation in multiplayer games and real-life contexts
    2.1.3 Conclusions from the validity studies

  2.2 HMTH

    2.2.1 1st Construct validity study of identification
    2.2.2 2nd Construct validity study of identification
    2.2.3 1st Construct validity study of self-esteem
    2.2.4 2nd Construct validity study of self-esteem
    2.2.5 General conclusion / outlook

  2.3 UKA

    2.3.1 fMRI of digital gaming and Think-Aloud
    2.3.2 Conclusion

  2.4 HGO/BTH

    2.4.1 Mixed measures construct validity pilot study of combat game experience
    2.4.2 Conclusion

  2.5 TUE

    2.5.1 Construct validation of behavioral indicators of flow, frustration, and boredom
    2.5.2 Construct validity of the Game Experience Questionnaire
    2.5.3 Conclusion

3 General discussion

4 References

  4.1 References for section 2.1 (CKIR)
  4.2 References for section 2.2 (HMTH)
  4.3 References for section 2.3 (UKA)
  4.4 References for section 2.4 (HGO/BTH)
  4.5 References for section 2.5 (TUE)
1 GENERAL INTRODUCTION

In this deliverable, the WP5 partners present the results of the FUGA construct validity studies. The objective of the WP5 was to establish the construct validity of the different potential measures of game experience based on the different measurement techniques. This was the first step to ensure that the game experience measures meet the psychometric criteria that are commonly applied to self-report and behavioral measures. Given that the aim was to thoroughly understand and describe emotion and experience during gaming, the measurements addressed experience at its various response levels, i.e., subjective (cognitive and emotional), behavioral, psychophysiological, and neural. Different measures were construct-validated against the self-report and think-aloud methods and the convergent and discriminant relations between the different measures and between the different types of game episodes were examined.

In order to validate and calibrate the objective measures, the GEQ (see Deliverable 3.3) was previously developed. The theoretical work performed by all partners in the FUGA project (Deliverable 2.1) provided a firm and broad basis for the development of the questionnaire. The GEQ has a modular structure, consisting of the core questionnaire (GEQ-Core), probing multiple components of players’ experience while gaming, the post-game questionnaire (PGQ), probing gamers’ experience after the gaming session and any after effects, and the social presence module (SPGQ), probing players’ experience of and involvement with their co-player(s). Additionally, a short in-game version of the GEQ was developed, the iGEQ, for probing in-game experience multiple times during a gaming session. Although questionnaires are sensitive to significant biases and experimental demand characteristics, self-report data provides valuable insights into the subjective aspects of an experience when used in combination with objective corroborative indicators.

The GEQ measurements were used by all partners as one accepted representation of game experience, to serve as a comparison to other measures to be construct-validated. Although the GEQ played a central role in validating the objective game experience measures, the project partners used also other self-report methods for this purpose. Where appropriate, the validity studies also involved the use of multiple measures so that convergent and discriminant relations could be identified. For example, several different psychophysiological, behavioral, and implicit measures were included that could be validated against each other (for example, HMTH used Lexical Decision Task, Affect Misattribution Procedure, Self-Description Task, and Implicit Association Test to examine the same constructs; see section 2.2). Think-aloud method described in Deliverable 3.4 was also used in construct-validating fMRI measures, thus offering an additional approach to measure game experience.

To further facilitate comparisons between the different game experience measures, the plan was to use the same game by different partners, where possible. In WP4, the partner HGO/BTH created a modification of a commercial game Half-Life 2 named FUGAmod, designed to provide an ecologically valid digital game that would also support the experimental designs described in Deliverable 3.1. (more information on stimulus game production is available in section 3 of D3.1). However, not all experiments allowed the use of FUGAmod. For example, a mobile psychophysiological experiment had to be performed with a handheld console not able to run the game engine the mod is based on, and experiments using implicit measures had to use clear-cut roles (such as soldier, racer) which the mod couldn't provide. Nevertheless, the project goals benefit from the fact that a common stimulus game was used in the construct validity studies of the different measures.
In section 2.1, CKIR presents the results of the construct validation of psychophysiological measures of game experience in the laboratory and real-life contexts, where facial electromyography (EMG), electrodermal activity (EDA), electroencephalography (EEG), physical movement (acceleration) sensors, and cardiac measures were used. In section 2.2, HMTH presents their validation process of seven different implicit methods focusing on the concepts of identification and self-esteem, putatively important facets of game experience. In section 2.3, UKA focused on flow and its neurophysiological correlates in the brain, providing evidence for validity of the fMRI method in studying game experience. HGO/BTH used a multimethod approach to look into the effect of difficulty of a game level, reported in section 2.4. And finally, in section 2.5, TUE presents the results of their work focusing on behavioral measures, Gaming Experience Questionnaire (GEQ), and their cross-validation. Because not all analyses could be done within the timeframe of WP5, more detailed and comprehensive results will be reported in the publications of the project.

In the Annex I, it was stated that the construct validity studies in WP5 are critical for the success of the project. Although some of the expected relationships did not gain support from the present studies, mostly the results increased our confidence that many of the measures examined can be used to index game experience. Therefore, the reliability and predictive validity studies (in WP6 and WP7) have a solid base to start from and can be carried out mostly as planned. The changes are explained in the partners' respective sections.

2 RESULTS OF CONSTRUCT VALIDITY STUDIES BY PARTNER

2.1 CKIR

The objective of the construct validity studies was to validate different measures based on laboratory and mobile psychophysiological recordings as measures of game experience and investigate how they reflect the different dimensions of game experience. In all the experiments conducted by CKIR several well-researched and intercomparable psychophysiological methods are used: electrocardiography (ECG), facial electromyography (EMG), measures of electrodermal activity (EDA), and electroencephalography (EEG). In addition to these basic measurements, others are included as supplemental methods in only some of the studies, such as measurement of respiration and bodily movement. Many of the psychophysiological methods can be cross-validated, due to their converging theoretical backgrounds. Each study also makes use of Game Experience Questionnaire (see FUGA Deliverable 3.3; and section 2.5 for more information) as a means to provide common measure that links the results of all the partners together.

As it has been stated before, many psychological phenomena, such as attention, information processing, emotion, and arousal, that are important aspects of game experience, have psychophysiological components or concomitants (e.g., autonomic nervous system activation associated with emotions). Particularly emotions are an integral part of game experience, as the activity itself is voluntarily entered and intrinsically motivated, and therefore primarily the emotions, and not conscious cognitive assessments, are in critical position when the experience is valued (see e.g., Grodal, 2000). However, no single psychophysiological measure is likely to be sufficient when studying game experience, but it is advisable to use multiple measures so that differential response patterns (or profiles) can be identified. This is because, in psychophysiology, there is often a many-to-one relation between psychological processes and psychophysiological measures (e.g., Cacioppo & Tassinary, & Berntson, 2000). That is, a given psychophysiological measure can potentially be linked to several psychological constructs (such as emotion and...
attention), and the interpretation of psychophysiological measures is highly dependent on the context and research paradigm. Also, the strength of the association between a physiological event and a psychological event is typically not very high. The use of response profiles of different physiological variables, instead of single variables, results in much higher specificity and sensitivity in the relationship under study (Cacioppo et al., 2000).

According to the dimensional theory of emotion, all emotions can be located in a two-dimensional space, as coordinates of valence and arousal (or bodily activation; e.g., Lang, 1995; Larsen & Diener, 1992). The valence dimension reflects the degree to which an affective experience is negative (unpleasant) or positive (pleasant). The arousal dimension indicates the level of activation associated with the emotional experience, and ranges from very excited or energized at one extreme to very calm or sleepy at the other. The different psychophysiological recordings chart these dimensions relying on partly overlapping physiological processes, and together they provide a well-established body of data which can be validated in the context of digital games.

The facial EMG is a direct measure of the electrical activity associated with facial muscle contractions that are an important form of emotional expression (Tassinary & Cacioppo, 2000), and are considered as the primary psychophysiological index of hedonic valence (e.g., Lang et al., 1993; Ekman, Davidson & Friesen, 1990). That is, facial EMG has been shown to be a successful method in discriminating positive emotions from negative emotions; increased activity at the zygomaticus major (cheek) and orbicularis oculi (periocular) muscle regions has been associated with positive emotions and activity at corrugator supercilii (brow) with negative emotions.

Frontal alpha asymmetry is probably the most studied electroencephalographic (EEG) measure of emotion and motivation. Studies that have examined relationships between different individual difference measures and resting EEG activity have shown that asymmetrical activation of the anterior cortical regions seems to influence emotional responding (Allen, Harmon-Jones & Cavender, 2001). According to Davidson (1998) left frontal activity, either as a state or a trait, indicates a propensity to approach a stimulus, whereas relatively greater right frontal activity indicates a propensity to withdraw from a stimulus. The underlying assumption in frontal asymmetry studies is that activity in the alpha range (8-13 Hz) is inversely related to underlying cortical processing, it is typical that alpha power decreases when the underlying cortical systems engage in active processing (Coan & Allen 2004). It must be emphasized that these frontal asymmetries are not measures of positive or negative affects per se, but they tap a broader motivational tendency towards approach-related or withdrawal-related behaviors and emotions (Allen, Harmon-Jones & Cavender, 2001; Davidson, 1998)

Electrodermal activity (EDA), commonly known as skin conductance, is considered an important psychophysiological index of arousal (Lang et al., 1993). EDA gives direct information about the electrical conductance of the skin that is related to the level of sweat in the eccrine sweat glands. The neural control of the sweat glands is entirely under control of the sympathetic nervous system (SNS), and as people experience arousal their SNS is activated, resulting in increased sweat gland activity and skin conductance (Dawson, Schell & Filion, 2000). Several studies examining psychophysiological reactivity to stress have also shown that digital games elicit a pronounced emotional arousal- or stress-related increase in heart rate (HR) and blood pressure (e.g., Llabre, Spitzer, Saab, Ironson, & Schneiderman, 1991; Matthews & Jennings, 1984; Murphy, Stoney, Alpert, & Walker, 1995). While HR has been suggested as a simple index of arousal (Lang et al., 1993; Witvliet & Vrana, 1995), it is influenced by both the SNS and PNS (parasympathetic nervous system; Papillo & Shapiro, 1990). Where increased cardiac sympathetic activity causes the heart to speed up and is associated with emotional arousal, general preparation for action, and mobilization of various types of resources (Obrist, 1981), increased cardiac parasympathetic (i.e., vagal) activity
causes the heart to slow down when attention is paid to an external stimulus (Porges, 1995). On the other hand, the approach motivation and other EEG findings can be used for studying the attentional component of the game experience. Numerous studies have documented the relationship between reduced EEG alpha activation and an increase in visual attention and interest (e.g., during television viewing; Reeves, Thorson, Rothschild, McDonald, Hirsch & Goldstein, 1985; Simons, Detenber, Cuthbert, Schwartz, & Reiss, 2003). Because playing a digital game is an active task, employing attentional resources and evoking emotional arousal are both likely to occur during the activity. Depending on the game, it may also elicit negative and positive emotions alike (e.g., horror, frustration, joy of victory, humor). How these different facets can be best tapped into with the psychophysiological methods is tracked in these studies.

The acceleration data, providing a measurement of overall movement of the participant, was included in the study as a new potentially useful measure of something that might be related to game experience. It is a commonly known phenomenon that people tend to move when they play an engaging digital game, for example when trying to lean to get a better view behind an obstacle, or when tilting and moving a game controller, instinctively trying to control the game in a way the game does not support. This engagement in a game is an enjoyable experience (e.g., Ravaja, Saari, Turpeinen, Laarni, Salminen & Kivikangas, 2006), and it could be measured with properly located accelerometers.

Two experimental studies with extensive psychophysiological recordings were conducted. Study 1 mainly aimed to construct validate the phasic psychophysiological measurements of specific game events, by comparing the results with each other, but also with self-reported experiences of those events. The essential benefit of psychophysiological methods, covert recordings that do not confound the experience itself, was maintained by recording the whole audiovisual output of the game, and using automatized software to pick the preselected events for a self-report assessment afterwards. Study 2 was aimed at expanding the scope of the methodology to multiplayer games and mobile gaming, in addition to executing the essential comparison between laboratory and natural playing environments. The validation of the methodology in both studies is done by comparing the physiological responses with self-reports of relevant dimensions and by using experimental manipulations designed to reveal the differences that are already known from previous research. However, due to strict time constraints of the project plan, only basic analyses could be completed, and the results represent this unfinished state. More analyses are conducted afterwards, and the final results will be reported in the future deliverables.

2.1.1 Construct validation of phasic psychophysiological responses

Introduction

In this study, the goal was to construct validate phasic psychophysiological methods in the context of digital games. The main validation method was to examine the relationship of the various psychophysiological measures with self-reported game experience as measured by the GEQ and other questionnaires.

This was done in two ways. The self-report forms filled in after the playing period were compared to the physiological data recorded from the entire playing periods (i.e., tonic physiological data). This is the most common way to compare data from self-reports and continuous psychophysiological data (e.g., Ravaja, Saari, Turpeinen et al., 2006; see Ravaja, 2004). In some studies the researchers have resorted to interrupting the game experience in order to fill out a questionnaire in particular intervals or at specific pre-defined points (e.g., van Reekum, Johnstone,
Banse, Etter, Wehrle & Scherer; see also Ravaja, 2004). However, it is very likely that this unnatural pausing affects the game experience and thus may taint the self-report measurements. This is why for this study a novel technique was developed: predetermined events from the video recording of the game are automatically picked and shown to the participant after the game, thus enabling in-game GEQ self-reports of the without affecting the game experience in any way. Those in-game GEQ scores are compared to high temporal resolution (i.e., phasic) physiological data at the time of the event, providing information about a certain event type and how the participant experienced it on subjective (self-report) and objective (physiological reactions) level. Nevertheless, it is notable that the in-game variation of GEQ was designed for the interruption technique, which aims at probing the game experience not in response to certain events, but the game experience in general (see D3.3). For this reason, some additional measures must be used to achieve satisfactory convergent validity.

Self-Assessment Manikin (SAM; Lang, 1980) has been found a valid and reliable method to provide self-report information about valence and arousal, and in several studies it has been successfully used with physiological methods (e.g., Schneider, Lang, Shin & Bradley, 2004). It is used to assess valence and arousal at a point, and due to the basic nature of the dimensions, it can easily be used in conjunction with specific events. The Positive and Negative Affect Scale (PANAS; Watson, Clark & Tellegen, 1988) is another short inventory that has been employed with physiological methods, assessing discrete emotions such as joy and anger which can also be connected to specific event.

To further ensure the validity of these methods, only events that have been previously studied were used as events that were automatically reviewed during the experiment. There is already research on what kind of subjective and objective responses should events such as death of a game character and victory elicit (see Ravaja, Turpeinen, Saari, Puttonen & Keltikangas-Järvinen, 2008; Kivikangas & Ravaja, n.d.), and use of these events in this study allows for further validation of the event-related self-reports and psychophysiological measurements.

Another type of events potentially providing evidence for convergent validity concerns the use of non-player character's (NPC) facial expression. It is well established that seeing the facial expression representing an emotion elicits a similar expression in the viewer (e.g., Dimberg, 1982). Furthermore, it has been shown that while stimuli can affect the emotions, more complex stimuli of congruent (emotions of similar hedonic valence) or incongruent (emotions of opposite hedonic valence) emotions either intensify or damp down the elicited emotions, respectively (Branscombe, 1985; Neumann, Seibt & Strack, 2001). By using NPCs with different emotional facial expressions in the game it should be possible to replicate these effects, further providing evidence for validity of psychophysiological measurements during a digital game. However, these effects are expected to be much more subtle than main effects, and should be only considered to provide corroborating evidence if both the main effects and the specific event comparisons show significant differences.

Methods

Participants

The participants were 40 (21 male and 19 female) volunteers from various backgrounds, between ages of 18 to 31 (mean = 22.7) years, and who played digital games at least 4 hours per month. They were recruited by advertisements in gaming-related websites and contacting student mailing lists and student organizations. The participants received three movie tickets for participation.
**Game**

The game used was FUGAmod, a custom-made mod based on popular FPS game Half-Life 2 (Valve Corporation, Bellevue, WA, 2004), played on a powerful desktop computer. The original game is a FPS (first-person shooter) with a heavy emphasis on story and realistic character depiction. FUGAmod was produced within FUGA project by Cognition, Simulation, Interaction, and Game Science Group at Blekinge Institute of Technology, and it had several different game levels. Of those, one level (Church Walk) was used to familiarize the participants with the game controls and environment, and two other levels (Rescue Action and Secret Corridors) were used as experiment stimuli. All three levels follow the typical FPS conventions of character control, environment interaction, and combat (see e.g., http://en.wikipedia.org/wiki/First_person_shooters). The mod was created to give maximum information about in-game events, both by saving it to a text file and by sending it as event markers directly to the psychophysiological data acquisition system via a serial cable.

Secret Corridors is made in a typical FPS game fashion to provide a game experience similar to one of a commercial game. The goal of the mod is simply to reach the end of the level, navigating through the labyrinthine and zombie-infested town. Both ammo and health packs are available but not overly abundant (some of them hidden). In addition to breakable wooden barriers, some visibly distinguishable barrels would explode when shot at, providing some complexity to the environment and momentarily offering a significant increase in firepower if used properly. Thus, Secret Corridors is mainly a combat oriented game experience with small aspects of tactics and resource management.

Rescue Action includes also some combat, but the main content is a simple task incorporating interaction with the NPCs. The story conveyed to the players beforehand is that the PC is a freedom fighter in a dystopian future that needs to rescue some people, one by one, from an occupied building, and distinguish these from enemy spies. The player needs to go to one room, press a button, and decide what to do with the appearing NPC. After the button is pressed, the game zooms to the face of the appearing NPC for two seconds and shows two things: a facial expression (either a happy smile or an angry frown) and a password that gives away if the NPC is an ally or not. Regardless of the facial expression, an ally NPC should be led to a lobby with a lift, where he or she thanks the player, gives a reward (a health pack), and exits through a door. A spy NPC should be simply shot, although the NPC starts shooting back if he or she does not die right away.

More elaborately, the two types of NPCs (ally and spy) in the level come in four subtypes consisting of the combinations of the types and facial expressions: a) a smiling NPC that should be rescued, b) a smiling NPC that should be killed, c) a frowning NPC that should be rescued, and d) a frowning NPC that should be killed. Each of these NPC subtypes occurred three times during the task, to a total of 12 NPC encounters in the entire play period. There are also twelve different NPC models, which represent both sexes (six men and six women) and different races. The encounters were randomized separately by NPC types, by NPC models, and by expressions, so that any given encounter could incorporate any NPC type, any NPC model, and either of the expressions.

The smiling and frowning expressions were created for the twelve NPC models with the Source SDK Face Poser (included in Half-Life 2 Source tools), using Ekman & Friesen's (1976) happy and angry faces as example, and could be attached to any NPC model. The participants assessed the pictures of the expressions (see below), and the expressions could be told apart by their valence ($F(911.039) = 2366.095, p < .001$).
Experimental design and procedure

All the measurements were conducted in an electrically shielded room which could be monitored from the observation room next to it by a camera to ensure the safety of the participant. The game was also monitored on a clone screen in the observation room. All instructions were provided on the computer screen, and the most important parts were also repeated by the experimenter on the intercom. The proceeding of the experiment was controlled by custom software.

Before starting the experiment, the participant confirmed he or she had filled in the background and trait questionnaires via internet. After a brief description of the experiment, the participant filled out an informed consent form. Electrodes were then attached to the participant and they were seated in comfortable armchairs. The participant had an introductory to the game and a brief practice period using the Church Walk level, during which the experimenter explained the controls and interaction with game environment. After that, the experimenter left the room and only re-entered at the end of the experiment or in the occasion of technical problems. A rest period of five minutes, during which baseline physiological measurements were performed, followed. After the rest period, and before and after both playing periods the state-related self-report measures were filled in.

The main part of the experiment consisted of two playing periods which were recorded on video, and two review periods (one following each playing period) in which the participant reviewed short clips of the videos and answered a few items concerning each clip. The order of two playing periods, one using the level Secret Corridors and one using the level Rescue Action, was randomized. Both levels were played until the end of the level, or, in case of Secret Corridors, alternatively until 10 minutes was full, whichever was sooner. Rescue Action was not stopped at any time limit due to the small amount of different NPC repetitions per participant. Therefore, whereas after Secret Corridors the amount of video clips shown was dependent on how far the participant could proceed in the level, after Rescue Action necessarily all 24 events had occurred.

Finally, the validity of facial expressions of the NPC models was examined. The software showed the participant all the 12 models with both facial expressions (smile and frown, totaling in 24 pictures), and with each of them the participant assessed the emotional status of the expression. After that the electrodes were removed, and the participants were debriefed and they had a possibility to present questions. The participants were thanked for their participation and handed the compensation for their time. In total, the experiment took about two and a half hours.

Events and reviewing

Before each video clip, a pop-up described the event depicted in the video clip and what the participant was supposed to pay attention to (e.g., "in the next video clip, pay attention to: the character and the password"). After the video clip the participant was instructed to self-report the experience during the game, not the current experience (i.e., the video clips served only as a reminder of the actual game event).

Secret Corridors level included three main kinds of enemies that are of interest here. Human-like walking zombies slowly wobble towards the PC without presenting much of a challenge, but they'd normally occur in larger packs. Fast, screaming dog-like beasts on four legs are located so that they are always encountered close up. Smaller creatures called "headcrabs" that resemble a fat spider typically attack by leaping longer distances and are harder to hit due to their small size, and of which a poison headcrab is a more dangerous variation, immediately lowering PC's health to minimum on impact and thus giving other enemies a better chance of killing the PC before the health recovers to normal level. There are further variations of these, such as faster headcrabs and
more rugged zombies that carry poison headcrabs with them, but events concerning these weren't analyzed.

These game events were recorded during the game play and showed to the participant in the review period: (a) Zombie Killed, where the participant shot or otherwise killed a basic variation of a zombie-type enemy, (b) Beast Killed, where the participant shot or otherwise killed a fast dog-like type of enemy, (c) Headcrab Killed, where the participant shot or otherwise killed a poison variation of a headcrab, and (d) PC Died, where the player character was killed. Each of these events was reviewed for six seconds, three seconds before the event and three seconds after.

Rescue Action events were divided to two classes, faces and actions, and video clips of both classes were shown 12 times, totaling in 24 reviews in the review period. Of the first class, the events were all the same: (e) Face Seen, where a new NPC first appears and the game zooms to the face, to make sure that the participant noted the expression (the controls are disabled so the participant couldn't turn away or shoot the NPC too early). Action class was further divided to three subclasses: (f) Ally Rescued, where the participant succeeded in escorting an ally NPC to the lobby and the NPC thanked the PC and gave a health kit as a reward, (g) Spy Killed, where a NPC designated as spy was killed, and (h) Ally Killed, where a NPC designated as ally was killed. It is worth noting that in the game the participant could shoot an ally NPC, or the ally NPC could be killed by enemies en route to lift, so the ratio of killed/rescued NPCs could be skewed from the default 6 rescues / 6 kills, despite the original balance between the two event subclasses. During the review period after Secret Corridors level, up to 12 video clips could be shown depending on which events actually occurred during the preceding free play.

In order to minimize the overlap between the Face Seen and Spy Killed events, the both were reviewed for five seconds, three before the event and two after. The Face Seen event video started a second before the game started zooming to the face, the zooming and showing the password took two seconds, and the two seconds after started when the participant again gains the controls of the PC. NPC Rescued events were reviewed for six seconds, three seconds before the event (from the moment where the NPC turns to the PC and thanks for the help) and three seconds after.

Data collection

Physiological data acquisition

Facial EMG activity was recorded from the left corrugator supercili, zygomaticus major, and orbicularis oculi muscle regions as recommended by Fridlund and Cacioppo (1986), using surface Ag/AgCl electrodes with a contact area of 4 mm diameter (Med Assoc. Inc., St. Albans, VT). Electrodes were filled with TD-240 electrode gel (Med Assoc. Inc). The raw EMG signal was amplified, and frequencies below 30 Hz and above 400 Hz were filtered out, using the Psylab Model EEG8 amplifier (Contact Precision Instruments, London, UK). The raw signal was rectified and integrated over 1000 ms using the RFP language of Psylab8 software.

Electrodermal activity (EDA) was recorded with Psylab Model SC5 24-bit digital skin conductance amplifier that applied a constant 0.5 V across Ag/AgCl electrodes with a contact area of 8 mm diameter (Med Assoc. Inc.). Electrodes were filled with TD-246 skin conductance electrode paste (Med Assoc. Inc.) and attached to the middle phalanges of the ring and little fingers of the subject’s left hand after hands were washed with soap and water (the ring and little fingers were used to reduce the interference between gaming and EDA recording).

The digital data collection was controlled by Psylab Stand-Alone Monitor and Psylab8 software, and all physiological signals were sampled at a rate of 1000 Hz.
Electroencephalography (EEG) and cardiac activity were also recorded.

**Game event recording and reviewing system**

FUGAmod was specifically created for FUGA so that it could save data about the game events without the need of laborious manual scoring, e.g., from recorded video. This also enabled the automated review system of the played game. Whenever a predefined event of interest occurred in the game, it was logged on the computer's hard disk and sent as an event code to the Psylab BIN8 data acquisition system with a precision of milliseconds (limited only by operating system and the game itself running on the computer).

The game was recorded in real-time with a commercially available frame grabber software Fraps (Beepa P/L, Melbourne, Australia, 2007) with capture rate set to 25 frames per second, and the resolution equalling the game's screen resolution. Using a custom-made software the events could then be extracted from the video file and shown in a quality identical to the image seen by the participant during the game. Thus, each review of a game event was identical to the corresponding game event in audio and video modalities, leaving out only the effect of interaction between the participant and the stimulus.

**Questionnaire data**

Both trait and state questionnaires were used in the experiment. The trait questionnaires were available in the internet so that the participants could complete them at any time before the experiment. The trait questionnaires used were:

- Trait aggression was measured using the Aggression Questionnaire (Buss & Perry, 1992)
- Dispositional behavioral inhibition system (BIS) and behavioral activation system (BAS) sensitivities were measured using the BIS/BAS Scales (Carver & White, 1994)
- Balanced emotional empathy was measured using the Balanced Emotional Empathy Scale (Mehrabian, 2000)
- Perfectionism was measured using the Striving for Excellence subscale of Perfectionism Inventory (Hill, 2004).
- Impulsivity / sensation seeking and sociability were measured using the respective scales from ZKPQ (Zuckerman, Kuhlman, Joiremann, Teta, & Kraft, 1993)

Game experience was assessed by self-report measurements before and after both playing periods, and after each review of a game event. The scales used were:

- Both the in-game and core versions of the game experience questionnaire (GEQ), developed by research partner TUE (see FUGA Deliverable 3.3; also see section 2.5 for more information). The short in-game GEQ was used in case of the reviews, the core GEQ after the playing periods.
- The pleasure and arousal scales of visual Self-Assessment Manikins (Lang, 1980; for more information, see section 2.5.)
- A modified version of translated Positive and Negative Affect Scale (PANAS; Watson, Clark & Tellegen, 1988). The scales used were: Positive Affect, Negative Affect (Anger), Negative Affect (Fear), Relaxation, Depression.
Before the playing periods, an item assessing the expected stressfulness, threat, and getting along in the game, and after the playing periods the corresponding items assessing the same dimensions actually experienced.

In addition the SAM assessments were used after the rest period, and with each picture of a NPC expression during facial expression validity check, together with items describing the six basic emotions (joy, anger, fear, sadness, disgust, surprise).

All the self-report measurements were completed on another computer next to the game computer to minimize their impact on the game experience.

Data reduction and analysis

Mean values for each physiological channel were derived for one second before (baseline; second 1) and six seconds after the event (seconds 2-7). Logarithmic transformations were conducted for skin conductance and EMG data to normalize the distributions. All data were analyzed by the Linear Mixed Model procedure with restricted maximum likelihood estimation and a first-order autoregressive covariance structure for the residuals.

Results

In this deliverable only results from Rescue Action are reported. Due to technical problems experienced when using the SC level some additional work is required for the data to reach quality standards necessary for the analyses.

Self-reported emotional responses to game events

Self-reported pleasure was higher when killing a frowning NPC compared to killing a smiling NPC, $M_s = 5.1$ and 4.8, respectively, $F(1, 165.27) = 5.73, p = .018$. Likewise, killing a frowning NPC elicited higher arousal compared to killing a smiling NPC, $M_s = 6.0$ and 5.5, respectively, $F(1, 164.90) = 13.83, p < .001$. Also, killing a frowning NPC elicited higher Positive Affect compared to killing a smiling NPC, $M_s = 2.9$ and 2.7, respectively, $F(1, 164.42) = 4.68, p = .032$. The differences between the conditions in Relaxation, Negative Affect (Anger), Negative Affect (Fear), and Depression were nonsignificant, $p_s > .10$.

Self-reported pleasure was also dependent on the type of NPC (ally or spy); that is, interacting with an ally elicited greater self-reported pleasure compared to interacting with a spy (enemy), $M_s = 6.4$ and 4.8, respectively, $F(1, 778.663) = 140.03, p < .001$. In addition, there was a significant Facial Expression × NPC Type interaction in predicting pleasure, $F(1, 778.75) = 4.55, p = .033$. Figure 1 shows that pleasure was higher when there was a congruence between the type of a NPC and facial expression of a NPC (i.e., smiling ally, frowning spy) compared to an incongruent situation (i.e., frowning ally, smiling spy).

The results also showed that interacting with a frowning NPC elicited greater self-reported arousal compared to interacting with a smiling NPC, $M_s = 5.4$ and 5.1, respectively, $F(1, 776.13) = 8.78, p = .003$. Likewise, interacting with a spy elicited greater self-reported arousal compared to interacting with an ally, $M_s = 5.5$ and 5.1, respectively, $F(1, 776.37) = 11.15, p = .001$. 
Physiological responses to game events

The Congruent NPC versus Incongruent NPC × Quadratic Trend across Seconds 1-5 interaction contrast was significant when predicting zygomaticus major EMG activity elicited by the Face Seen event, $t(2798.61) = -1.96, p = .049$. That is, when the player saw the face of a NPC for the first time and the type of a NPC (ally or spy) was revealed to the player, a congruency between NPC type and facial expression (i.e., smiling ally and frowning spy) elicited an increase in zygomatic activity, whereas an incongruence of NPC type with facial expression (i.e., smiling spy and frowning ally) elicited a decrease in zygomatic activity (see Figure 2). There was also a significant Spy versus Ally × Linear Trend across Seconds 1-7 interaction contrast when predicting orbicularis oculi EMG responses, $t(2840.59) = -3.52, p < .001$. That is, the emergence of a spy elicited a less pronounced decrease in orbicularis EMG activity compared to the emergence of an ally.
When comparing facial EMG responses elicited by killing a spy and receiving thanks and a health kit from a rescued ally, it was found that killing a spy elicited an increase in zygomatic and orbicularis oculi EMG activity, whereas the reverse tended to be true for rescuing an ally, for the Spy Killed vs. Ally Rescued × Linear Trend across Seconds 1-7 contrast, $t(2712.46) = -4.92$, $p < .001$ (for zygomatic EMG) and $t(2810.05) = -6.87$, $p < .001$ (for orbicularis oculi EMG). Figure 3 shows that killing a frowning spy tended to elicit a greater increase in orbicularis oculi EMG activity compared to killing a smiling spy, although the Smiling Spy versus Frowning Spy × Linear Trend across Seconds 1-4 interaction contrast narrowly failed to reach statistical significance, $t(1353.67) = -1.92$, $p = .056$. It was also found that killing the spy NPC elicited a greater increase in SCL compared to rescuing an ally NPC (see Figure 4), for the Spy Killed vs. Ally Rescued × Linear Trend across Seconds 1-7 interaction contrast, $t(3079.41) = -4.53$, $p < .001$. 

![Graph showing zygomaticus major EMG responses to the Face Seen event for congruent and incongruent NPCs.](image)

Figure 2. Zygomaticus major EMG responses to the Face Seen event for congruent (smiling ally and frowning spy) and incongruent (smiling spy and frowning ally) NPCs.
Orbicularis Oculi EMG (ln[ȝV])

Figure 3. Orbicularis oculi EMG responses elicited by killing a spy NPC for smiling spies and frowning spies.

SCL (lg[ȝS])

Figure 4. Skin conductance level (SCL) responses to killing a spy and receiving a reward from a rescued ally.
The relationship of phasic physiological responses with self-reported emotions

Table 1 shows the results of the analyses of the relationship of phasic psychophysiological responses to different game events with self-reported valence and arousal during RA. It was found that the linear trend across Seconds 1-3 in zygomatic EMG activity (contrast score) was positively associated with self-reported pleasure, \( p = .047 \). Likewise, the linear trend across Seconds 1-3 in orbicularis oculi EMG activity was positively related to self-reported pleasure, \( p = .010 \). In addition, the linear trend across Seconds 1-3 in corrugator supercilii EMG activity was negatively related to self-reported pleasure, \( p = .003 \).

Table 1. Results of Linear Mixed Models Analyses: Relationship of Self-reported Valence and Arousal with Phasic Physiological Responses to Game Events

<table>
<thead>
<tr>
<th>Variable Source</th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
</table>
| **ZM EMG**
| Intercept       | 1, 181.06 | 2.480 | .117  |
| Event type      | 7, 720.71 | .837  | .556  |
| Event number \(^b\) | 2, 721.57 | 1.983 | .138  |
| Valence \(^c\)  | 1, 303.77 | 3.985 | .047  |
| Arousal \(^d\)  | 1, 58.58  | .077  | .782  |
| **CS EMG**
| Intercept       | 1, 328.43 | 4.336 | .038  |
| Event type      | 7, 747.56 | .648  | .716  |
| Event number \(^b\) | 2, 748.52 | .647  | .524  |
| Valence \(^c\)  | 1, 644.82 | 8.966 | .003  |
| Arousal \(^d\)  | 1, 139.02 | .005  | .945  |
| **OO EMG**
| Intercept       | 1, 231.57 | 3.887 | .050  |
| Event type      | 7, 719.93 | 1.225 | .286  |
| Event number \(^b\) | 2, 725.48 | .516  | .597  |
| Valence \(^c\)  | 1, 396.83 | 6.705 | .010  |
| Arousal \(^d\)  | 1, 82.74  | .165  | .686  |

Note. EMG = electromyography; CS = corrugator supercilii; OO = orbicularis oculi; ZM = zygomaticus major.

\(^a\)Linear trend across Seconds 1 through 3 (contrast score)

\(^b\)Sequence number of an event

\(^c\)Self-reported valence

\(^d\)Self-reported arousal
Discussion

The aim of this study was to construct-validate the psychophysiological methods as measurements of the game experience. We used the psychophysiological measures of media research previously proven to be useful in the field: facial electromyography, electroencephalography, and a measure of electrodermal activity (skin conductance). The experiment was designed to enable direct comparisons between the measurements under scrutiny and self-reports, but also to provide convergent validity by focusing on variables already known from previous research. A novel technique where quick reviews of previously played digital game events could be assessed without interrupting the game play was employed.

The comparisons between the self-report data and facial EMG measurements showed encouraging results, as the game events were systematically assessed more positively when the activity on zygomaticus major (ZM) and orbicularis oculi (OO) muscle areas were higher and the activity on corrugator supercilii (CS) muscle area was lower, and vice versa. Furthermore, the discriminant validity of the measurements was also found to be strong, as expectedly they were not connected to other constructs such as arousal. This is completely in line with the dimensional model of emotion, where the emotions can be located in the space made of two dimensions, valence and arousal (Lang, 1995; Larsen & Diener, 1992). Previous research on pictures (Lang et al., 1992; Lang, 1995), emotional video clips of varying length (Simons, Detenber, Roedema & Reiss, 1999; Hubert & de Jong-Meyer, 1990), and emotional audio clips (Bolls et al. 2001) have shown the consistency of these results, which have also been repeated with more subtle emotional subtext in natural media messages, such as news clips or advertisements (Ravaja, 2004b; Ravaja, Kallinen, Saari & Keltikangas-Järvinen, 2004; Hazlett & Hazlett, 1999). Similar results have also been reported in the context of digital games (e.g., Ravaja, Saari, Turpeinen, et al., 2006), but the relationship has not been this systematically examined before the present study.

Unexpectedly the results did not support the relationship between electrodermal activity and arousal. Although the biological basis of skin conductance as index the index of arousal is the most justifiable, and the use of the measurement with other kinds of media is well-grounded, we did not find statistically significant relationship between self-reported arousal and skin conductance level when examining phasic events. This is also contrary to previous findings: several studies (e.g., Ravaja et al., 2004, 2005, 2006; Schneider et al, 2004) have successfully used EDA as an index of arousal, in connection with both tonic and phasic data. However, due to the time constraints of the project the data analysis was quickly done, and it is entirely possible that there are some decisive hidden variables that are affecting these results. For example, the events reviewed might not be ideal to bring out enough variation in self-reports, reducing the power of the statistical test. Furthermore, there is some strong convergent evidence (see below) from a comparison between specific events that corroborate with the original assumptions about the connection between arousal and EDA.

At this point the data also did not support the assumption that the facial expression of a NPC, smile or frown, could be discriminated with psychophysiological measurements: the Face Seen event did not elicit different physiological responses dependent on the facial expression of the NPC. It could be argued that the effect of the facial expression is too weak in comparison to the variation due to other sources (e.g., the task) in which case the connection should be examined indirectly. Examining the congruency between the NPC's expression and task is one way to circumvent the problem and see if interactions with other variables and facial expressions can be found. This approach provided a significant difference between the congruent and incongruent face during Face Seen event in EMG activity over ZM muscle area, denoting a more positive response for a congruent expression-task combination. However, no other measures showed significant effects.
Furthermore, the Ally Rescued event did not reveal a significant difference between the expressions, and although the Spy Killed event showed an effect to the expected direction on OO activity, the difference between the expressions narrowly failed to reach statistical significance. The activity increased more in connection to killing a frowning spy as compared to a smiling spy, which could be interpreted that the incongruency between the NPC’s expression and task might suppress the positive reaction to a "victory" where the PC defeated a spy (see Ravaja et al., 2004), or simply the joy of active participation after passively monitoring for the affiliation of the NPC (cf. Ravaja et al., 2005, 2006). The lack of meaningful connections on part of any other measurements might indicate that OO, tapping specifically into high arousal positive (Ekman et al., 1990), is more sensitive to the reaction of this kind. Another possibility, of course, is that the facial expressions simply did not cause genuine response differences discernable with these methods.

At the time of writing the post-test assessments of the facial expressions of NPC models were not yet completely analyzed. The preliminary analyses however revealed that the quality of NPC facial expressions varied greatly, and it is possible that some individual NPC model-expression combinations did not qualify as satisfactorily depicting the expression in question. Therefore it is also possible that the results concerning the facial expressions are contaminated with sub-par stimuli and might present stronger relationships if cleaned up thoroughly. The data will be inspected and purified when necessary, and more comprehensive analyses will be done in the future.

When comparing the spy and ally NPCs of the Face Seen event, it was revealed that linear OO activity was reduced significantly less when the emerging NPC was designated as a spy than when it was designated as an ally. This could be related to the same kind of joy of active participation as mentioned in case of Spy killed event – the anticipation of action, a positive high arousal affect, is reduced more when the participant knows that the NPC is not an enemy.

A strong linear effect was found when comparing the event types Spy Killed and Ally Rescued, as the former elicited an increase in ZM and OO, denoting a positive affect, whereas the latter elicited a decrease in the same muscle areas. Also, skin conductance had a significant linear increase in case of Spy Killed event as compared to Ally Rescued event. These results could be explained by the nature of the events: in the Spy Killed event the participant is an active agent and personally accomplishing the victory over an enemy NPC, whereas in the Ally Rescued event the participant mainly waits for the NPC to get in the lift so that he or she could proceed to meet the next NPC. Thus, the participation in the action elicits rises in positive valence and in arousal, when compared to decreases or no effect in case of passive waiting, which is completely in line with previous research (Ravaja et al., 2005, 2006), and strongly supports the use of EMG and EDA measurements in relation to phasic events in a digital game.

2.1.2 Construct validation in multiplayer games and real-life contexts

Introduction

While the basic validation of psychophysiological methods as an instrument in measuring game experience was the main purpose of this study, it had two additional important goals: to find out if the measurements of the game experience work differently if, first, the game is played with other people, and second, the game is played in real-life context (i.e., home) as compared to laboratory setting.

According to the previous research there is reason to suspect that the player setup – is it a single-player or a multiplayer game, are the players located in the same room or connected via internet, are they playing cooperatively or against each other – may have a substantial effect on the game
experience. (Massively multiplayer online, or MMO, games are outside the scope of this study.) Studies aimed at finding out why people play digital games report that players of the single-player and multiplayer games may have different motivations for playing (e.g., Raney, Smith & Baker, 2006). A study by Sherry and others (Sherry, Lucas, Greenberg & Lachlan, 2006) listing the motivations found by focus groups shows two of the six dimensions, competition and social interaction, that emphasize other people as essential for how the game is experienced. Indeed, this has also been shown with psychophysiological measures: in a study by Ravaja and coworkers both self-reported and psychophysically measured pleasure (as EMG activity) differed significantly depending on whether the game was played alone, against a stranger, or against a friend (Ravaja, Saari, Turpeinen et al., 2006; Kivikangas & Ravaja, n.d.).

However, there are important differences in multiplayer games which may be as important as the single-player/multiplayer distinction itself. The most obvious is the team division or "sides": who are in the same side, and who are their opponents? In a cooperative game, there are many human players in the same team, with or without some extra computer-controlled characters, against the game (in form of monsters, enemies, obstacles, etc.). In competitive games the players progress independent of each other (such as reaching the goal first or getting more points), or they directly oppose the other player or players (e.g., the players attempt to kill the other PCs). While a cooperative game cannot be completed only by playing against other players, competitive games are designed exactly for that. They have multiplayer set-ups ranging from everyone-against-everyone to equally balanced teams to one where a single player has a highly favorable position but plays against all others. As it is often possible to replace some of the team members with computer control, from the experimenter's point of view this allows for examining the effect of different playing modes during a multiplayer game.

The aforementioned study by Ravaja and others (Ravaja, Saari et al., 2006), which found differences in the game experience depending on who the opponent was, employed a competitive action game. While it gives important information about the game experience concerning the different opponents, the study only starts to cover the many possible variations of multiplayer gaming. The important distinction between cooperative and competitive modes of play raises a question how the mode affects the game experience. For example, competition has been seen as detrimental to intrinsic motivation (see, e.g., Tauer & Harackiewicz, 1999), an important part of gaming motivation (Sellers, 2006; Grodal, 2000), which suggests possible differences in the game experience between cooperative and competitive modes. In an effort to clear this up, we employed both the cooperative and competitive modes of play in our experiment. Although including also the single-player mode as a control condition would have been ideal, the duration of the experiment rose over 4 hours with all the manipulations, and the single-player mode was ruled out for practical reasons. Additionally, we chose a game which is as similar as possible in both modes, thus avoiding the effect of game play itself changing the measured game experience.

The other specific goal for this study is more practical. Virtually all experimental game studies are conducted in a laboratory, which is an unfamiliar environment for most participants, possibly presenting nervousness, anxiety, and other confounding variables to the measured and assessed game experience. While some of these unwanted effects cannot be removed as long as the participant is aware of the experimental situation, simply moving the experiment to a familiar environment, the participant's home, may help to alleviate the problem. To see if the place where the experiment is conducted indeed has some effect on the game experience, we repeated the experimental procedure both in a laboratory and at a participant's home. This allows the comparison of all the measurements used in the study as a function of the experimental environment.
Methods

Participants
Participants were 48 (30 male and 18 female) volunteering Finnish young adults, ranging from 18 to 34 years of age (mean = 24.0), who played digital games at least 4 hours per month. They were recruited in pairs (two same-sex friends) by advertisements in gaming-related websites and contacting student mailing lists and student organizations. All participants received three movie tickets for participation.

Design
A 2 (Game mode: competition, collaboration) × 2 (Playing Context: laboratory, home) within-subjects design was employed.

Game
The game used was Bomberman (Hudson Entertainment, Inc., Redwood Shores, CA, 2006), played on PlayStation Portable (PSP) handheld game console (Sony Computer Entertainment, Inc., Tokyo, Japan). Bomberman is a remake of a classic action game where two or more players are situated in a small maze with partly breakable walls, and the goal is to use bombs to clear new routes in the maze, reveal bonuses, and ultimately, blast the other player(s). It employs small-sized abstract graphics from isometric view with cartoonish characters and happy music and sounds. There is no realistic violence included in the game.

The game was played with the two participants assigned either to the same or different team. In the collaboration mode, a team of two human-controlled player characters (PCs) competed against a team of two computer-controlled non-player characters (NPCs), whereas, in the competition mode, a team of one PC and one NPC competed against a team of another PC and another NPC. With the exception of team arrangements, the game is identical in both modes. In the game, each player starts from one corner of the maze, blows up breakable walls with bombs, collects bonuses when they appear, and tries to blast both characters of the other team. By positioning the bombs appropriately the player can force the opponent character to a situation where avoiding all the blast waves is difficult or impossible (see picture 1). The play becomes progressively more intensive as the maze gets more and more open and the players receive random bonuses that grant the characters more speed, more simultaneous bombs to use, stronger blasts, and various other helpful or hindering effects.
The game was played with three matches per battle (if the same team won the first two matches, the third match was not played). The duration of each match was 3 min maximum, and there were one or two battles per condition (8-min play session). Sudden death was set off, player starting position shuffle was set on, and skulls (bonus items that hinder rather than benefit the collector) were set as blast-proof. Revenge mode was turned on, so that when a character was defeated in the maze, the player could move the character around the maze and throw bombs inside to help his or her team. If the player of the defeated character succeeded to blow up another team’s character before either team won, the player got his or her character back to the maze.

**Procedure**

In the beginning of the experiment in the first context (laboratory or participant’s home; see below), the participants confirmed they had filled in the background and trait questionnaires via the Internet. After a brief description of the experiment, they filled out an informed consent form. The participants were seated in comfortable armchairs (or equivalent in the participant’s home) located next to each other, electrodes were attached to both participants, and the physiological data recorders were time synchronized with each other. While the electrodes were attached, the participants played the game “Untold Legends: The Warrior’s Code” (Ubisoft Entertainment, Montreuil, France, 2006) to get used to the game device. This was followed by a 5-min practice session during which the participants played the game “Bomberman” without team allegiances (everyone against everyone).

The experimental play sessions took place in two different playing contexts (locations): (a) laboratory and (b) home of one of the participants. The order of the playing contexts was randomized for each pair of participants. In each playing context, there was a rest period of 5 min (during which baseline physiological measurements were performed), followed by two 8-min play sessions, one in competitive mode and one in collaborative mode. Across the experiment, the two game modes were played in one of two orders (BA AB or AB BA), the order being chosen randomly for each pair of participants. The laboratory room was dimly illuminated during the rest period and when playing the game. When playing at the participant’s home, there was normal illumination adjusted by the participant. When arranging the experiment, the experimenter ensured
that the participant’s home would be peaceful (no distractions like children or loud music) and suitable chairs would be available for both participants. During the game sessions, the experimenter was in the adjacent room. After the rest period and before and after each play session, different self-report measures were obtained. After answering the last post-game questions in the first location, the participants moved to the other location with the experimenter (electrodes were not removed). When the experiment was completed in the second location, the electrodes were removed and the participants were debriefed and thanked for their participation.

Data collection

Physiological data acquisition

The physiological signals and environmental data were recorded from both players with the Varioport-B portable recorder systems (Becker Meditec, Karlsruhe, Germany).

Facial EMG activity was recorded from the left corrugator supercili, zygomaticus major, and orbicularis oculi muscle regions as recommended by Fridlund and Cacioppo (1986), using surface Ag/AgCl electrodes with a contact area of 4 mm diameter (Becker Meditec, Karlsruhe, Germany). Electrodes were filled with Synapse conductive electrode cream (Med-Tek/Synapse, Arcadia, CA). The raw EMG signal was sampled at 1024 Hz, amplified (amplification factor = 4899) and frequencies below 57 Hz and above 390 Hz were filtered out, using the analog filter built in the Varioport device. The raw signal was rectified and smoothed implementing a linear phase FIR filter using the Kaiser window method (101 coefficients, low-pass cutoff frequency 40 Hz).

Electrodermal activity (EDA) was recorded with Varioport 16-bit digital skin conductance an amplifier (input range = 0–70 μS) that applied a constant 0.5 V across Ag/AgCl electrodes with a contact area of 4 mm diameter (Becker Meditec), sampling at 32 Hz. Electrodes were filled with TD-246 skin conductance electrode paste (Med Assoc. Inc.) and attached to the middle phalanges of the ring and little fingers of the subject’s left hand after hands were washed with soap and water (the ring and little fingers were used to reduce the interference between gaming and EDA recording).

Electrocardiogram (ECG) was recorded using Varioport isolated AC amplifier, together with three EKG leads in a modified Lead 2 placement, sampling at 512 Hz.

Two accelerometers were used to record global level of body and hand-held console movement. Varioport acceleration sensors measure all three axes accurately (2 %) to an acceleration equivalent to 2 g, and with reduced accuracy (10 %) to 8 g on each axis. The accelerometer recording the body movements was inside the Varioport device, firmly held in place by the carrying belt and positioning the device in upright position on participant’s chest (to keep the arms free and back to the chair). The accelerometer recording the hand-held console movements was attached with velcro tape to the bottom of the console.

In addition, we recorded respiration, ambient illumination, temperature, and noise, but the results were not yet available at the time of writing. Respiration was measured with the respiration belt, which consists of a disposable belt with an internal sinus wire. The belt acts as the inductance of an oscillator, so stretching the belt results in frequency changes which represent the respiration curve after amplification and filtering process. The respiration belt was applied on thorax and calibrated individually for each participant to ensure that the maximum and minimum of the breathing cycle do not exceed the recordable range. The sampling rate was 32 Hz. Ambient illumination was measured with a light sensor consisting of a photo-sensitive semiconductor
element. Ambient temperature was measured with a temperature sensor (accuracy = 0.1 °C between 20 and 40 °C).

All data were stored on a CompactFlash memory card (1 GB) after digitizing with a 16-bit A/D converter. Sampling rate was 512 Hz for ECG, 1024 Hz for facial EMG, 32 Hz for EDA, 32 Hz for acceleration, 128 Hz for light, and 1 Hz for temperature.

**Electroencephalography (EEG) data acquisition**

**EEG recordings.** Electroencephalography (EEG) was recorded continuously from 24 subject pairs with VarioPort EEG-amplifier using 0.9-Hz high-pass filter and 70-Hz low-pass filter. The input range was ± 500 μV. Recordings were done using an electrode cap. Data was collected from F3, F4, C3, C4, P3, and P4 electrodes, according to the international 10/20-system. The electrodes were referred to linked ears and the ground electrode was located at the mid-forehead (AFz). To facilitate artifact detection ocular movements were recorded with two EOG channels. For vertical eye-movements, the electrodes were placed below and above the right eye; for horizontal eye-movements the electrodes were placed to outer canthi of right and left eye. Impedance at each electrode site was maintained below 5 kΩ. In this deliverable only results for frontal alpha (8-12 Hz) asymmetry are reported.

**Questionnaire data**

Both trait and state questionnaires were used in the experiment. The trait questionnaires were available in the internet so that the participants could complete them at any time before the experiment. The trait questionnaires used were the same that were used in the Study 1.

Game experience was assessed by self-report measurements before and after both playing periods. The scales used were:

- The core, post-game and social presence versions of the game experience questionnaire (GEQ), developed by research partner TUE (see FUGA Deliverable 3.3).
- The pleasure and arousal scales of visual Self-Assessment Manikins (Lang, 1980; for more information, see section 2.5.).
- A modified version of translated Positive and Negative Affect Scale (PANAS; Watson, Clark & Tellegen, 1988). The scales used were: Positive Valence, Negative Valence, Positive Affect, Negative Affect (Anger), Negative Affect (Fear), Relaxation, Depression.
- A shortened version of ITC-Sense of Presence Inventory (ITC-SOPI; Lessiter, Freeman, Keogh & Davidoff, 2001). The scales used were Engagement and Spatial Presence.
- An item assessing the expected stressfulness, threat, and getting along in the coming game and corresponding items assessing the same dimensions actually experienced after the game.
- In addition, the modified PANAS assessments were used after the rest period.

**Physiological data reduction and analysis**

Mean values of the psychophysiological measures were derived for each of the sixteen 30-s epochs during the games. Acceleration data were integrated over one second and 3-dimensional axes were added together and rectified by taking a square root from the sum of second powers of the axes. Logarithmic transformations were conducted for facial EMG, EDA, and acceleration data to normalize the distributions.
The data were analyzed by the Linear Mixed Models procedure in SPSS with restricted maximum likelihood estimation and a first-order autoregressive covariance structure for the residuals. Participant pair (1 to 24) and individual within a pair (1 or 2) were specified as the subject variables, and playing context (laboratory or home), game mode (collaboration or competition), and time (Epochs 1 to 15) were specified as the repeated variables. Playing context, game mode, time, order of playing context, order of game mode, and gender were selected as factors, and age, baseline physiological value, and different self-report dimensions of emotions and game experience were selected as covariates (GEQ and SAM dimensions in separate analyses). A fixed-effects model that included the main effects of the aforementioned variables and the Playing Context × Game Mode and Game Mode × Gender interactions were specified.

EEG data reduction and analysis

After the recordings the EEG data were segmented into 30-s epochs for each 8-min gaming session. The first 30-s segment from each gaming session was excluded from further analysis. For artifact removal each 30-s segment was further segmented to 2-s segments. Each 2-s segment was automatically screened to exclude those in which amplitudes exceeded ± 50 μV. Thereafter, the epochs were additionally visually inspected for artifacts and those found to contain artifacts were rejected on all channels from further analyses. Then, for each artifact-free (2-s) epoch, power estimates (μV²) were derived with the fast Fourier transform (FFT) for the broad alpha band 8-12 Hz. A Hanning window with no overlap was used. The spectral estimates were finally In-transformed to normalize the distributions. Finally, using the power values of 2-s segments, average power values in the 8-12 Hz alpha band were calculated for each 30-s epoch.

The band power estimate data were analyzed by the Linear Mixed Models procedure in SPSS with restricted maximum likelihood estimation and a first-order autoregressive covariance structure for the residuals. To examine the interaction of frontal alpha asymmetry and playing mode or context a model where context (two levels: home, laboratory), playing mode (two levels: competition, co-operation) sequence number of an event (ranging between 1 and 15), order of context (two levels: home first, laboratory first), order of playing mode (two levels: competition first, co-operation first), and hemisphere (two levels: right, left) were selected as factors, and a fixed-effects model that included the main effects of these variables and the interactions playing mode × hemi and context × hemi was specified.

Another model was specified to study the interactions of frontal alpha asymmetry and self-reported valence and arousal ratings given after the playing session. In this model context (two levels: home, laboratory), playing mode (two levels: competition, co-operation), order of context (two levels: home, laboratory), order of playing mode (two levels: competition first, co-operation first), sex (two levels: female, male), hemisphere (two levels: right, left), sequential count of event (ranging between 1 and 15), valence (ranging from 1 to 9), and arousal (ranging from 1 to 9) were selected as factors, and a fixed-effects model that included the main effects of these variables and the interactions valence × hemisphere and arousal × hemisphere was specified.

Three more models were specified to test interactions between frontal alpha asymmetry and different dimensions of the FUGA Game Experience Questionnaire (GEQ). For GEQ-Core dimensions a model was specified where context (two levels: home, laboratory), playing mode (two levels: competition, co-operation), order of context (two levels: home, laboratory), order of playing mode (two levels: competition first, co-operation first), sex (two levels: female, male), hemisphere (two levels: right, left), sequential count of event (ranging between 1 and 15), seven different GEQ-Core dimensions, and the interactions of these seven dimensions and hemisphere were selected as factors. The scores of these seven dimensions were dichotomized thus there were two levels (high...
and low: 1, 2) for each GEQ-Core dimensions: Positive Affect, Competence, Sensory and Imaginative Immersion, Flow, Tension, Negative Affect, and Challenge.

For GEQ-Social dimensions a model was specified where context (two levels: home, laboratory), playing mode (two levels: competition, co-operation), order of context (two levels: home, laboratory), order of playing mode (two levels: competition first, co-operation first), sex (two levels: female, male), hemisphere (two levels: right, left), sequential count of event (ranging between 1 and 15), four different GEQ-Social dimensions, and the interactions of these four dimensions and hemisphere were selected as factors. The scores of these seven dimensions were dichotomized thus there were two levels (high and low: 1, 2) for each GEQ-Social dimensions: Psychological Involvement – Empathy, Psychological Involvement – Negative Feelings, and Behavioural Involvement.

For GEQ-Post-Game dimensions a model was specified where context (two levels: home, laboratory), playing mode (two levels: competition, co-operation), order of context (two levels: home, laboratory), order of playing mode (two levels: competition first, co-operation first), sex (two levels: female, male), hemisphere (two levels: right, left), sequential count of event (ranging between 1 and 15), three different GEQ-Post-Game dimensions, and the interactions of these three dimensions and hemisphere were selected as factors. The scores of these dimensions were dichotomized thus there were two levels (high and low: 1, 2) for each GEQ-Post-Game dimensions: Returning to Reality, Negative Experiences, Positive Experiences, and Tiredness.

Results

Results from external acceleration sensor data are not reported, because the data could not be adequately normalized. With most participants the data distribution presented two peaks, representing the fact that either the participant held the console very still, or moved it quickly, but these were clearly two different activity categories with no mid-area of slow movement in between.

The effects of the experimental manipulations on psychophysiological measurements are presented below in the Table 2. Due to the significance of sex in these tests a post-hoc test comparing the sex and Bomberman background variable (representing the previous experience of the participant with the digital game used in the experiment) was run. It was found that males were significantly (two-tailed $t(42) = 2.936$, $p = .006$) more experienced with the game than the females ($M_s = 3.62$ and 2.00, respectively, $SD_s = 1.96$ and 1.53, respectively).

Table 2. Means of psychophysiological measurements by experimental manipulations. The differences are analyzed using Linear Mixed Models.

<table>
<thead>
<tr>
<th>Variable Source</th>
<th>1 (and SE) across conditions</th>
<th>2</th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ZM EMG</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Context</td>
<td>3.093 (0.039)</td>
<td>3.163 (0.039)</td>
<td>1, 512.848</td>
<td>1.639</td>
<td>.201</td>
</tr>
<tr>
<td>Playing mode</td>
<td>3.027 (0.037)</td>
<td>3.229 (0.037)</td>
<td>1, 934.443</td>
<td>18.669</td>
<td>.000***</td>
</tr>
<tr>
<td>Order of context</td>
<td>3.123 (0.039)</td>
<td>3.133 (0.039)</td>
<td>1, 577.427</td>
<td>0.034</td>
<td>.853</td>
</tr>
<tr>
<td>Order of playing mode</td>
<td>3.069 (0.037)</td>
<td>3.187 (0.037)</td>
<td>1, 802.959</td>
<td>6.121</td>
<td>.014*</td>
</tr>
<tr>
<td>Sex</td>
<td>3.228 (0.035)</td>
<td>3.028 (0.045)</td>
<td>1, 394.410</td>
<td>12.412</td>
<td>.000***</td>
</tr>
<tr>
<td>Interaction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Playing mode ×</td>
<td>3.000 (0.053)</td>
<td>3.053 (0.051)</td>
<td>1, 797.225</td>
<td>0.126</td>
<td>.723</td>
</tr>
<tr>
<td>Context</td>
<td>3.186 (0.051)</td>
<td>3.273 (0.053)</td>
<td>1, 1060.933</td>
<td>22.065</td>
<td>.000***</td>
</tr>
<tr>
<td>Sex ×</td>
<td>3.018 (0.045)</td>
<td>3.438 (0.045)</td>
<td>1, 1060.933</td>
<td>22.065</td>
<td>.000***</td>
</tr>
<tr>
<td>Playing mode</td>
<td>3.036 (0.058)</td>
<td>3.021 (0.058)</td>
<td>1, 1060.933</td>
<td>22.065</td>
<td>.000***</td>
</tr>
<tr>
<td><strong>CS EMG</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

23
### Context
- **Playing mode**: 2.938 (0.057) 2.951 (0.057) 1,311.428 0.146 .702
- **Order of context**: 2.907 (0.057) 2.982 (0.057) 1,253.375 4.556 .033*
- **Sex**: 2.899 (0.065) 2.990 (0.087) 1,275.044 4.355 .039
- **Interaction**: 2.904 (0.072) 2.971 (0.061) 1,2767.999 3.939 .047*

### Playing mode
- **Context**: 3.341 (0.034) 3.344 (0.034) 1,464.906 0.003 .956
- **Order of context**: 3.302 (0.034) 3.383 (0.034) 1,571.265 3.211 .074
- **Sex**: 3.292 (0.030) 3.393 (0.039) 1,337.782 4.169 .042*
- **Interaction**: 3.289 (0.045) 3.283 (0.043) 1,866.703 0.049 .825

### Sex
- **Playing mode**: 3.417 (0.049) 3.369 (0.049) 1,216.132 18.522 .000***

### EMG
- **OO**: 2.907 (0.063) 2.954 (0.064) 1,217.482 0.086 .769
- **Playing mode**: 2.938 (0.057) 2.951 (0.057) 1,311.428 0.146 .702
- **Order of context**: 2.907 (0.063) 2.954 (0.064) 1,217.482 0.086 .769
- **Order of playing mode**: 2.928 (0.056) 2.961 (0.056) 1,2767.908 1.905 .168
- **Sex**: 2.899 (0.065) 2.990 (0.087) 1,275.044 4.355 .039
- **Interaction**: 2.904 (0.072) 2.971 (0.061) 1,2767.999 3.939 .047*

### EMG
- **EDA**: 1.200 (0.085) 1.211 (0.085) 1,995.552 0.289 .591
- **Order of context**: 1.160 (0.085) 1.250 (0.085) 1,2701.888 49.001 .000***

### EDA
- **Playing mode**: 1.200 (0.085) 1.211 (0.085) 1,995.552 0.289 .591
- **Order of context**: 1.160 (0.085) 1.250 (0.085) 1,2701.888 49.001 .000***
- **Sex**: 1.258 (0.100) 1.153 (0.137) 1,45.035 0.381 .540
- **Interaction**: 1.218 (0.090) 1.181 (0.086) 1,2679.322 10.702 .001***

### Acceleration
- **Context**: -4.272 (0.015) -4.260 (0.015) 1,815.618 0.273 .601
- **Playing mode**: -4.284 (0.015) -4.248 (0.015) 1,940.049 3.096 .079
- **Order of context**: -4.245 (0.015) -4.288 (0.015) 1,817.787 3.841 .050*
- **Order of playing mode**: -4.287 (0.014) -4.245 (0.014) 1,878.231 4.421 .036*
- **Sex**: -4.254 (0.013) -4.278 (0.016) 1,761.403 1.271 .260***
- **Interaction**: -4.304 (0.021) -4.265 (0.021) 1,876.838 1.879 .171

### Interaction
- **Playing mode**: -4.240 (0.021) -4.256 (0.021) 1,942.876 8.951 .003***
- **Sex**: -4.265 (0.023) -4.291 (0.023) 1,942.876 8.951 .003***

**Note.** EMG = electromyography; CS = corrugator supercilii; OO = orbicularis oculi; ZM = zygomaticus major; EDA = electrodermal activity. All mean values are natural logarithms of the respective physiological measure. Significance: * p < .050, ** p < .010, *** p < .001.

Conditions 1 and 2 denote the following conditions in the manipulations, respectively: Context, laboratory and home; Playing mode, cooperative and competitive; Order of context, whether the playing period was first or second; Order of playing mode, cooperative first and competitive first; Sex, male and female. In case of interaction tables the column represents the first variable and the row represents the second variable (for example, in Sex × Playing mode interaction the column represents the sex, male in the first row and female in the second row, and the row represents the playing mode, cooperative on the first column and competitive on the second column.)
The relationship of tonic physiological responses with self-reported emotions

Table 3 shows the results of comparisons between psychophysiological measurements and the self-reported emotions measured by modified PANAS. It was found that the EMG activity over zygomaticus major and orbicularis oculi facial muscle areas and overall body movement level (acceleration data) were positively associated with self-reported positive affect, \( p = .011, .043, \) and \(.000, \) respectively. Furthermore, orbicularis oculi EMG activity was also positively associated with self-reported positive valence, \( p = .010. \) The EMG activity over corrugator supercilii muscle area was found to be positively associated with self-reported negative valence. Both electrodermal activity and overall body movement level were negatively associated with self-reported depression, \( p = .006 \) and \(.010, \) respectively.

Table 3. Results of Linear Mixed Models Analyses: Relationship of self-reported emotions with tonic physiological responses over the playing periods.

<table>
<thead>
<tr>
<th>Variable Source</th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ZM EMG</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>1, 172.10</td>
<td>83.427</td>
<td>.000***</td>
</tr>
<tr>
<td>Positive valence</td>
<td>1, 147.60</td>
<td>1.422</td>
<td>.235</td>
</tr>
<tr>
<td>Negative valence</td>
<td>1, 160.82</td>
<td>0.399</td>
<td>.528</td>
</tr>
<tr>
<td>Negative affect: anger</td>
<td>1, 153.84</td>
<td>0.780</td>
<td>.378</td>
</tr>
<tr>
<td>Negative affect: fear</td>
<td>1, 157.71</td>
<td>0.193</td>
<td>.661</td>
</tr>
<tr>
<td>Positive affect</td>
<td>1, 145.24</td>
<td>6.703</td>
<td>.011*</td>
</tr>
<tr>
<td>Relaxation</td>
<td>1, 155.91</td>
<td>0.544</td>
<td>.462</td>
</tr>
<tr>
<td>Depression</td>
<td>1, 148.26</td>
<td>0.000</td>
<td>.995</td>
</tr>
<tr>
<td><strong>CS EMG</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>1, 166.00</td>
<td>112.251</td>
<td>.000***</td>
</tr>
<tr>
<td>Positive valence</td>
<td>1, 139.62</td>
<td>1.318</td>
<td>.253</td>
</tr>
<tr>
<td>Negative valence</td>
<td>1, 150.25</td>
<td>6.308</td>
<td>.009**</td>
</tr>
<tr>
<td>Negative affect: anger</td>
<td>1, 143.65</td>
<td>0.681</td>
<td>.410</td>
</tr>
<tr>
<td>Negative affect: fear</td>
<td>1, 148.36</td>
<td>0.021</td>
<td>.886</td>
</tr>
<tr>
<td>Positive affect</td>
<td>1, 138.04</td>
<td>0.531</td>
<td>.468</td>
</tr>
<tr>
<td>Relaxation</td>
<td>1, 147.09</td>
<td>0.142</td>
<td>.707</td>
</tr>
<tr>
<td>Depression</td>
<td>1, 140.36</td>
<td>0.002</td>
<td>.964</td>
</tr>
<tr>
<td><strong>OO EMG</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>1, 172.02</td>
<td>125.039</td>
<td>.000***</td>
</tr>
<tr>
<td>Positive valence</td>
<td>1, 144.99</td>
<td>6.765</td>
<td>.010**</td>
</tr>
<tr>
<td>Negative valence</td>
<td>1, 156.13</td>
<td>1.171</td>
<td>.281</td>
</tr>
<tr>
<td>Negative affect: anger</td>
<td>1, 150.13</td>
<td>1.454</td>
<td>.230</td>
</tr>
<tr>
<td>Negative affect: fear</td>
<td>1, 153.34</td>
<td>1.168</td>
<td>.282</td>
</tr>
<tr>
<td>Positive affect</td>
<td>1, 143.00</td>
<td>4.164</td>
<td>.043*</td>
</tr>
<tr>
<td>Relaxation</td>
<td>1, 151.87</td>
<td>0.239</td>
<td>.626</td>
</tr>
<tr>
<td>Depression</td>
<td>1, 145.50</td>
<td>0.268</td>
<td>.605</td>
</tr>
<tr>
<td><strong>EDA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>1, 169.85</td>
<td>83.427</td>
<td>.000***</td>
</tr>
<tr>
<td>Positive valence</td>
<td>1, 129.19</td>
<td>0.046</td>
<td>.831</td>
</tr>
<tr>
<td>Negative valence</td>
<td>1, 133.63</td>
<td>0.002</td>
<td>.968</td>
</tr>
<tr>
<td>Negative affect: anger</td>
<td>1, 131.24</td>
<td>1.194</td>
<td>.277</td>
</tr>
<tr>
<td>Negative affect: fear</td>
<td>1, 132.21</td>
<td>0.423</td>
<td>.516</td>
</tr>
<tr>
<td>Positive affect</td>
<td>1, 128.91</td>
<td>2.312</td>
<td>.131</td>
</tr>
<tr>
<td>Relaxation</td>
<td>1, 131.33</td>
<td>0.614</td>
<td>.435</td>
</tr>
<tr>
<td>Depression</td>
<td>1, 129.50</td>
<td>7.972</td>
<td>.006**</td>
</tr>
<tr>
<td><strong>Acceleration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>1, 166.10</td>
<td>83.427</td>
<td>.000***</td>
</tr>
<tr>
<td>Positive valence</td>
<td>1, 152.15</td>
<td>0.420</td>
<td>.518</td>
</tr>
<tr>
<td>Negative valence</td>
<td>1, 166.98</td>
<td>0.376</td>
<td>.540</td>
</tr>
</tbody>
</table>
The fun of gaming: measuring the human experience of media enjoyment

Report D5.1

<table>
<thead>
<tr>
<th>Affect</th>
<th>df 1</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative affect: anger</td>
<td>1, 161.38</td>
<td>0.773</td>
<td>.380</td>
</tr>
<tr>
<td>Negative affect: fear</td>
<td>1, 166.46</td>
<td>2.067</td>
<td>.152</td>
</tr>
<tr>
<td>Positive affect</td>
<td>1, 148.31</td>
<td>15.229</td>
<td>.000***</td>
</tr>
<tr>
<td>Relaxation</td>
<td>1, 165.46</td>
<td>1.076</td>
<td>.301</td>
</tr>
<tr>
<td>Depression</td>
<td>1, 154.52</td>
<td>6.739</td>
<td>.010**</td>
</tr>
</tbody>
</table>

Note. EMG = electromyography; CS = corrugator supercilii; OO = orbicularis oculi; ZM = zygomaticus major; EDA = electrodermal activity. Significance: * p < .050, ** p < .010, *** p < .001.

The relationship of tonic physiological responses with GEQ dimension

GEQ Positive Affect was positively associated with zygomatic EMG activity, F(1, 137.65) = 8.995, p < .003, orbicularis oculi EMG activity, F(1, 133.164) = 7.16, p = .007, and overall body movement level, F(1, 141.84) = 3.337, p = .070, although the last association only approached significance. GEQ Negative Affect was negatively associated with orbicularis oculi EMG activity, F(1, 133.12) = 6.599, p = .011.

EEG frontal alpha asymmetry and playing mode and playing context

Both interactions hemi × playing mode and hemi × context were non-significant, for both p > .5.

EEG frontal alpha asymmetry and self-reported valence and arousal ratings

Both interactions hemi × valence and hemi × arousal were significant, p < .001 for both. There was relatively more left than right hemisphere frontal alpha activity for both those who rated more negative and for those who rated more positive in valence. The hemispheric asymmetry was more prominent for those who gave more positive valence ratings. Similarly for arousal ratings, for those who reported higher arousal values the hemispheric difference was more prominent, although regardless of the arousal rating (high or low) the alpha activity was relatively stronger on left than on the right hemisphere (see Table 3 and Figure 5).

Table 3. Summary of interaction analyses

<table>
<thead>
<tr>
<th>Interaction</th>
<th>df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemi × Valence</td>
<td>4753.048</td>
<td>12.430</td>
<td>.000</td>
</tr>
<tr>
<td>Hemi × Arousal</td>
<td>4751.151</td>
<td>16.305</td>
<td>.000</td>
</tr>
<tr>
<td>Hemi × Imaginary and Sensory Immersion</td>
<td>4977.971</td>
<td>12.551</td>
<td>.000</td>
</tr>
<tr>
<td>Hemi × Competence</td>
<td>4974.814</td>
<td>26.632</td>
<td>.000</td>
</tr>
<tr>
<td>Hemi × Positive Affect</td>
<td>4974.656</td>
<td>76.331</td>
<td>.000</td>
</tr>
<tr>
<td>Hemi × Challenge</td>
<td>4973.517</td>
<td>117.826</td>
<td>.000</td>
</tr>
<tr>
<td>Hemi × Psychological Involvement – Negative Feelings</td>
<td>4982.748</td>
<td>11.170</td>
<td>.001</td>
</tr>
<tr>
<td>Hemi × Positive Experiences</td>
<td>4982.487</td>
<td>6.361</td>
<td>.012</td>
</tr>
<tr>
<td>Hemi × Returning to reality</td>
<td>5045.075</td>
<td>126.682</td>
<td>.000</td>
</tr>
<tr>
<td>Hemi × Tiredness</td>
<td>5044.271</td>
<td>101.964</td>
<td>.000</td>
</tr>
</tbody>
</table>

Note: Only statistically significant results are reported.
Figure 5. Mean EEG power values for interactions valence × hemi and arousal × hemi.

**EEG frontal alpha asymmetry and GEQ-Core dimensions**

Interactions of hemi with Sensory and Imaginative Immersion, Competence, Positive Affect, and Challenge were all significant, \( p < .001 \) for all. For both those who rated low and those who rated high in Sensory and Imaginative Immersion the alpha activation was relatively stronger on left than on right frontal lead. Similarly for both those who scored higher or lower in Competence the left frontal alpha activity was more prominent than the right frontal alpha activity. For dimension Positive Affect high self-reported scores were related to more dominant left frontal alpha activation, whereas low scores were related with somewhat larger right frontal alpha activation. For the Challenge dimension high self-reported scores were associated with relatively larger right hemisphere frontal alpha activation, and the low scores were associated with relatively larger left hemisphere frontal alpha activation. In all these interactions the frontal alpha activity was relatively stronger on left than on the right hemisphere (see Table 3 and Figure 6.).
Figure 6. Mean EEG power values for interactions of hemi and different GEQ Core dimensions: Sensory and Imaginative Immersion × hemi, Competence × hemi, Positive affect × hemi, and Challenge × hemi.

**EEG frontal alpha asymmetry and GEQ-Social dimensions**

Interaction hemi × Psychological Involvement – Negative Feelings was significant, \( p = .001 \). For both those who rated low and those who rated high in this dimension the alpha activation was relatively stronger on left than on right frontal hemisphere. This asymmetry was smaller for those who reported lower values in this dimension (see Table 3 and Figure 7).
EEG frontal alpha asymmetry and GEQ-Post-Game dimensions

Interactions of hemi with Positive Experiences, Returning to reality, and Tiredness were all significant, $p = .012$, $p < .001$, and $p < .001$, respectively. For both those who rated low and those who rated high in Positive Experiences dimension the alpha activation was relatively stronger on left than on right frontal hemisphere. This asymmetry was smaller for those who reported lower values in this dimension. With Returning to Reality dimension lower self reported values were associated with relatively stronger left frontal alpha power, and higher self reported values were associated with relatively stronger right frontal alpha power. Similarly with Tiredness dimension lower self reported values were associated with relatively stronger left frontal alpha power, and higher self reported values were associated with relatively stronger right frontal alpha power (see Table 3 and Figure 7).

Discussion

Physiological measurements and self-report scales

As expected, physiological activity was associated to self-reports in various levels. Self-reported Positive Affect, which represents a high arousal positive affect (consisting of expressions such as "enthusiastic", "joyful") in valence-arousal model, was positively associated with ZM and OO EMG activity, the well-established indices of positive emotions in psychophysiology (Lang et al, 1993). However, while there was a significant positive relationship between the OO EMG activity and Positive Valence scale, which is supposed to represent the valence dimension independent of the arousal dimension (consisting of expressions "delighted" and "happy"), the same self-report scale did not relate significantly with the ZM EMG activity. This is quite curious, as specifically OO has been reported to index high arousal positive emotions, whereas ZM has been thought to correlate more purely with valence (see e.g., Ravaja, 2004). Both ZM and OO activity was similarly positively associated with GEQ scale Positive Affect (consisting of expressions such as "I enjoyed it" and "I felt good"), which is very close to PANAS scale Positive Affect. Also, OO activity was negatively associated with GEQ scale Negative Affect, which represents boredom and frustration.
(consisting of expressions such as "I felt bored" and "I found it tiresome"), a low arousal negative emotion, the opposite of high arousal positive emotion that OO is indeed assumed to index (Witvliet & Vrana, 1995).

The other GEQ scales did not reveal any relationship with ZM or OO EMG activity, positive or negative. It could be argued that, according to two-dimensional model of emotion (Larsen & Diener, 1992), physiological measures positively associated with positive valence should also be able to capture negative valence by distinct lack of activity in these measures. However, the empirical studies (e.g., Witvliet & Vrana, 1995; Heponiemi, Ravaja, Elovainio, Näätänen & Keltikangas-Järvinen, 2006) have shown that in reality the association is not as clear as scholars would want it to be. There is also an alternate view on the dimensional model of emotion, where the valence dimension is not bipolar but can be divided to separate positive affect and negative affect dimensions (Watson, Wiese, Vaidya & Tellegen, 1999; Watson & Tellegen, 1999).

Negative valence scale (consisting of expressions "unhappy" and "sad") was positively associated with CS EMG activity, as was expected. Self-reported Negative Affect subscales, Fear and Anger, were not found to be in connection with CS activity (or any other physiological measure), although this could be due to the fact that the circumstances in the playing period did not warrant for enough variation in the negative end of valence dimension. The general mood of the game was happy (bright colors, happy music, cartoonish characters), and the participants were voluntarily playing with their friend of their own choosing and only for a short time – not giving much room for real negative emotions. This suspicion is further strengthened when the variations of the self-report scales are examined: on the scale from 1 to 7, the maximum value used on all the positive emotion scales is 7, but varies on negative emotion scales from 4.5 to 5, leaving the high end of the scales completely unused. This might also contribute to the fact that CS EMG activity did not present any relationships with GEQ scales.

Electrodermal activity is supposed to index arousal, the bodily activation component of emotions. While the self-reported emotion scales in PANAS do not employ a specific arousal scale, all but the Positive and Negative Valence scales include the arousal component within them: Positive Affect scale and Negative Affect subscales are supposed to tap into high arousal emotions, whereas Relaxation and Depression are low arousal emotions. However, only Depression, which had a significant negative relationship with EDA, revealed any association with the physiological index. It is possible that also here the lack of variation on the negative emotion scales has affected the results, although this explanation does not answer why the Depression scale did indeed provide a significant relationship. The lack of association with Relaxation scale might in turn be attributed to the very nature of the task: playing the Bomberman game is quite a hectic activity, and while the participant does not really have an opportunity to relax, he or she might get depressed by, for instance, his or her own failure in the game.

The acceleration data, measuring the overall movement of the sensor located on the chest of the participant, was positively associated with self-reported Positive Affect and negatively associated with self-reported Depression. Both relationships fit well in the idea that active engagement in the game is an enjoyable experience. When the participant is engaged in the game, he or she both moves much and assesses the experience with high Positive Affect. On the other hand, if the participant reports high Depression, he or she is not likely to move much during the game.

As for discriminant validity, none of the self-reported emotions yielded unexpected associations with any physiological measures. Overall, the validation of physiological measurements, particularly EMG activity measures, as indices of emotions were satisfactory, although the results are not strong and unanimous enough to disregard the need for further research on the current topic.
By employing proper experimental manipulations the experiment successfully provided corroborating evidence of physiological measurements' convergent validity. The significant difference in self-reported Positive Affect between playing modes was confirmed by ZM and OO EMG activity measures (and acceleration data supported the result, although narrowly failing to reach significance): competitive game mode elicited more positive emotion than cooperative game mode. Interestingly, also playing mode × sex interaction was found to be significant, meaning that while males experienced more positive emotions during the competitive game mode as compared to the cooperative game mode, the effect was reversed in case of females. However, conducting a test to compare sex and participants' previous experience with the game revealed that males had significantly more experience with the game than females. Thus, without further examinations we cannot conclude whether the difference was due to sex or previous gaming experience. The same applies to the effect of sex on physiological responses: more positive emotions were elicited in males than females (as indexed by ZM and OO activity), but the difference might be a product of differences in gaming experience.

No significant differences were found between the playing contexts (home or laboratory); this is also in line with our previous research (unpublished data, 2005). This is reassuring in that it supports the validity of laboratory studies. However, the interaction between the playing mode and playing context was found to be significant for corrugator EMG activity and EDA, suggesting that, when in a cooperative mode, the laboratory context elicited higher electrodermal response and lower corrugator activity than the home context; in the competitive mode, the effect was reversed for both physiological measures. This calls for additional studies, as the reason for the finding is not readily apparent.

**EEG frontal alpha asymmetry**

In the discussion concerning electroencephalography below we take the assumptions (presented in introduction) that 1) relatively larger right frontal activation indicates withdrawal-related emotions and behaviour, 2) relatively larger left frontal activation indicates approach-related emotions and behaviour, and that 3) alpha activation is inversely related to cortical activation.

There were no statistically significant effects to the frontal asymmetry by the playing mode or the context. Both valence and arousal ratings were similarly related to the frontal asymmetry: withdrawal-related EEG activation was observed for those who gave more positive valence ratings and for those who gave higher arousal ratings. It is suggested that during the playing the players encountered challenging events and possibly experienced these as causing withdrawal. However after the playing session the experience was rated positively, as the experience of excitement is one of the reasons for playing digital games and often more exciting games are considered better than those that evoke milder emotions.

Low ratings in Sensory and Imaginative Immersion were related to EEG response suggesting withdrawal, which seems understandable. If the player is not fully enough immersed to the game, he/she might feel like an outsider and possibly unmotivated to continue playing. In this sense it seems somewhat surprising that also the higher ratings in this dimension evoked similar, withdrawal-related, responses. This issue needs to be studied more carefully in future studies and also with other psychophysiological signals, for example facial electromyography. Similarly low level in dimension Competence was related to withdrawal related EEG responses, but the same held also for the higher ratings in this dimension, although with less prominent asymmetry.

Low levels in dimension Positive Affect were accompanied with EEG asymmetry suggesting approach-related emotions, and high levels were accompanied with EEG asymmetry suggesting
withdrawal-related emotions. It could be that the high values in SAM-rated valence and high values in dimension Positive affect are related. It is suggested that the withdrawal-related EEG asymmetries that accompanied these both is due to the excitement felt during the game, which afterwards was rated as a positive experience.

For dimension Challenge this was opposite: for low values the EEG asymmetry suggested withdrawal-related emotions, and for high values the EEG asymmetry suggested approach-related emotions. It seems logical that moderately high challenge is desirable in a game that was used in the experiment; the game had a clear mission and a set time frame. Setting the right level of challenge for a game is one of the most important tasks in a game development process and psychophysiological recordings could be useful in tasks like that.

EEG asymmetries for both dimensions Psychological Involvement – Negative Feelings, and Positive Experiences were similar. High values in both indexes were followed with EEG asymmetries indicating withdrawal-related emotions. For the dimension Psychological Involvement – Negative Feelings this is intuitive as this GEQ-Social dimension contained items like “I felt revengeful” and “I felt schadenfreude”. For the GEQ-Post-Game dimension Positive Experiences it may be that the withdrawal-related EEG asymmetry that accompanied high rating values was due to threat felt during the game, which afterwards was rated as an exciting positive experience.

Higher ratings in Returning to Reality dimension were associated with EEG asymmetries suggesting approach-related emotions and vice versa. Similarly with dimension Tiredness higher ratings were associated with EEG asymmetry suggesting approach-related emotions. It is hypothesized that it was felt positively that the game was captivating and somewhat fatiguing.

2.1.3 Conclusions from the validity studies

The objective of these studies was to validate the different psychophysiological measures as measures of game experience both in the laboratory and in a real life context (home). We examined the relationship of self-reported game experience with both tonic psychophysiological responses during the game session and phasic psychophysiological responses to specific in-game events. The physiological measurements used here were facial EMG, EEG, EDA, and body movement level.

It was found that facial EMG activity over zygomaticus major and especially orbicularis oculi muscle areas was strongly associated with different self-report measures of positive affect, and the EMG measurements also differentiated the experimental conditions used. While some additional research may be needed, the present results suggest that zygomatic and orbicularis oculi EMG activities are valid indices of positive affect during digital game playing.

Facial EMG activity over corrugator supercilii muscle area was also confirmed to be associated with different self-report measures of negative affect, but the results related to the experimental manipulations to which CS activity was expected to respond were less than ideal. However, it should be acknowledged that, in these two studies, the experimental manipulations may not have elicited particularly strong negative emotions, which may explain the present finding. Apparently, to examine the negative emotions relevant to game experience, such as frustration, anger, and depression, the experimental manipulations should be specifically designed to result in failures and feelings of inadequacy. A longer playing period than merely ten minutes could also help in wearing down the potential indifference associated with the first failures, although great care should be taken to ensure that the experiment is kept ethically acceptable. Nevertheless, although the validity of corrugator EMG activity cannot be considered as well established, it can be considered as a
potentially useful index of game experience. Thus, we continue to examine the benefits and weaknesses of corrugator EMG activity in the future studies.

Although the validity of EDA as a measure of emotional arousal during game playing was not unanimously supported by the data from these experiments, the results showed that EDA is sensitive to certain aspects of game experience. It can also be concluded that frontal EEG asymmetry is a potentially valuable measure in studying digital game playing-related experiences.

The overall level of body movements, as measured by an accelerometer attached on the participant's upper torso, yielded promising associations with positive emotions during digital gaming. This is most likely predicated on increased body movements when engagement is high. Preliminary analyses of the data from an accelerometer attached to the handheld console suggest that movement of the game controller may also be associated with game experience. However, in the experiments carried out by the FUGA partner TUE, TUE found no interesting associations of acceleration data with their other measurements. The differences in the experimental designs may have contributed to this discrepancy, but ultimately, the usefulness of acceleration measurements will be decided by future research.

In sum, the psychophysiological measures examined in these studies can be used to measure different aspects of game experience, although more research is needed to map more comprehensively the limits and strengths of the particular measures in particular experimental designs and contexts.
2.2 HMTH

The mission of partner HMTH within the FUGA project is to explore and establish useful so-called implicit measures to assess defined cognition-based elements of video game enjoyment. Within the broader conceptual frame of game enjoyment, two dimensions had been theoretically identified as candidates for implicit measurement, because they relate to processes well-researched in social psychology, the discipline that has invented and is using most implicit measures.

One enjoyment dimension is identification with a game character or role. Vicarious or “as-if”-experiences of ‘being’ a person in a game world, such as a war hero, a police officer, or a chief executive officer, have been argued as important driver of video game enjoyment (Klimmt, 2003; Hefner, Klimmt & Vorderer, 2007; see FUGA deliverable D2.1). HMTH has proposed to conceptualize such identification as alteration of players’ self-perception in such a way that in the state of strong identification, players see themselves more like the character they identify with (‘merged identity’: Goldstein & Cialdini, 2007) than they would do in a normal every-day state. Changes in self-perception, in turn, may be measurable by specific implicit methods, especially because such processes are likely to be automatic and non-conscious, which is the major reason why social psychologists are researching implicit methods that are capable to avoid self-report data (that necessarily cannot fully reflect automatic, non conscious cognition). Following extensive pilot testing in the preparation stage and a major problem with a candidate measure (EAST, see D3.2), one construct validity experiment was conducted within WP5. Because results were partially dissatisfying, more time and energy was invested into this line of FUGA research, and a second construct validity experiment was conducted in order to obtain a more clear overall picture. These studies and the main conclusions HMTH is proposing are presented in the following chapters 2.2.1 and 2.2.2.

The second dimension of game enjoyment HMTH is dedicated to measure implicitly is self-esteem processes related to player performance. Ample conceptual and some empirical arguments are available to understand players’ actual achievement, the individual sense of achievement and its impact on self-esteem (with associated emotional consequences such as pride or guilt) as important drivers of video game enjoyment (e.g., Klimmt, 2003; Klimmt & Hartmann, 2006; Behr, Klimmt & Vorderer, 2008; Ryan, Rigby & Przybylski, 2006). Because self-esteem is a construct well-researched in social psychology and has already been studied with implicit measures (Farnham et al., 1999), it was reasoned to conduct experiments in the game enjoyment context to see if implicit measures of self-esteem could be established as important explanatory factor (or mediator) in the player performance – game enjoyment relationship. Following extensive literature research, one construct validity experiment was conducted within WP5. Due to unsatisfying results, an additional experiment was implemented to find out if a more solid measurement solution in the game enjoyment framework could be established. These two studies and the key conclusions HMTH has derived from the findings are presented in chapters 2.2.3 and 2.2.4. A general discussion and outlook on future project work is offered at the end of this report chapter.

2.2.1 1st Construct validity study of identification

Introduction

On the basis of our findings in the three pilot studies, a new measurement was designed in order to assess identification implicitly. In the following paragraphs, the design and procedure of the experiment, as well as results and a discussion are reported.
Design of Implicit Measurement of Identification

Subsequent to the failure of the last pilot study using an “Extrinsic affective Simon Task” (EAST) to assess experimentally induced variation in identification intensity, the measurement approach to identification was redesigned. Additional expertise from social psychologists involved in research on implicit measurement was collected (we thank Tobias Gschwendner, University of Koblenz-Landau, for his friendly support). A completely new approach was then developed that aimed to measure identification with a digital game character as a three-component procedure of cognitive availability of character-related concepts, positive automatic evaluation of character-related visuals, and explicit character-related self-descriptions.

Cognitive availability of character-related concepts

The key characteristic of implicit measures is that they assess very fundamental and simple cognitive structures that do not depend on higher-order elaboration or reflection. One important and basic implicit surrogate of identification processes (as explicated above) is the cognitive salience of concepts that are closely connected to the character or action role with which players identify. Close identification with, for instance, a soldier role in a first-person shooter should come along with ‘thinking in a soldier’s categories’, such as friend or foe, attack and withdraw, weapons, tactics, camouflage, tanks, explosions, etc. To the extent that a player identifies with a character or role, concepts that are related to the character or role should thus be activated frequently in the player’s mind during game play. As a consequence, these concepts should be more salient and more accessible for automatic processing during and immediately after game play. Implicit measures addressing accessibility of concepts should thus be able to contribute to the measurement of identification, as concept accessibility is a fundamental and simple (implicit) cognitive variable, and related measures have been used widely in social psychology. In turn, cognitive accessibility of character-related concepts is proposed as necessary, but not sufficient condition of identification – strong involvement (Wirth, 2006) without identification could also result in higher accessibility of concepts related to the character or game world. Thus, accessibility of character-related concepts was used as one component of implicit identification measurement.

An important implicit measure that is designed to assess concept accessibility is the Lexical Decision Task (Macrae, Bodenhausen & Milne, 1995). This procedure presents letter sequences on a screen and requires participants to quickly decide whether the sequences present a word or are not a real word (non-word). Because the letter sequences appear only for a few milliseconds, decisions have to be made very fast (speeded task, an important element of implicit measurement). Concept accessibility is measured by the time participants need to identify words semantically related to the concept under study in comparison to words unrelated to that concept. For instance, strong accessibility of the concept ‘love’ should result in very fast decision times for words such as kiss, romance, or luck, but not speed up decision times for unrelated words such as train, politics, or tree.

Following the lexical decision task procedure described by Macrae et al.(1995), a lexical decision task was implemented to measure cognitive accessibility of character-related concepts in video game players. Again, a combat game was used as reference game (a shooter game with Word-War-II background). Character-related concepts thus come from military contexts, such as gun, weapon, attack, order, tank, or explosion. Players identifying with the soldier role should display greater accessibility to such concepts than players who do not identify with the soldier role. Thus, the lexical decision task presents words that reflect soldier-related contexts or are unrelated to that context (for the purpose of comparison). Specifically, a set of 16 words with connection to the soldier role and 16 words unrelated to that role is presented to participants in random order. The
same number of non-words (each created by changing one letter in a real word, which renders the non-words quite similar to words and thus increases the difficulty of the task) was added to the list of targets appearing on screen. Each word / non-word appears on screen for 75 milliseconds. Participants are requested to press either ‘d’ if the presented word is a real word or ‘k’ if the word is a non-word. The prediction is that players in the high-identification condition will display greater cognitive accessibility and thus faster decision times if words related to the soldier role appear on screen (compared to words unrelated to the soldier role and in comparison to players in a low-identification condition). Thus, cognitive accessibility is addressed as implicit component of identification with a game character.

Positive Automatic Evaluation of Character-Related Visuals

After assessing the cognitive availability of game-related words, we assessed in a second step the evaluation of game-related concepts. For this task, we adopted the so-called Affect Misattribution Procedure (AMP) for our purpose. This procedure was developed and first introduced in 2005 (Payne, Cheng, Govorun & Stewart, 2005). It combines the logic of projective measuring methods with an affective priming procedure and is thus based on the assumption that an evaluation of an object may be influenced by a preceding object. Thus, the affect following a positive or negative prime should be misattributed to a subsequent target which is then evaluated rather positively or negatively.

The design of the procedure bases on the work of Murphy and Zajonc (1993). They presented positive and negative valenced primes to participants and let the prime make an evaluative judgment of a Chinese pictograph on a 10-point-scale. The duration of the presentation of the prime was varied and they found out that the affective priming had only worked with a subliminally presented prime. Payne et al. doubted that a consciously processed prime does not have any effect of the subsequent evaluation and thus altered the design: The affective prime was presented visibly rather than subliminally and the ambiguity of the judgment was maximized by only allocating two answering possibilities, namely pleasant and unpleasant. The results of the described experiments show significant differences between the judgment of Chinese pictographs following positive (i.e. a baby), neutral (a grey square), and negative affective primes (i.e. a spider), even if the primes are presented visibly and even if the participants are actively told not to consider the primes while evaluating the target (which was assumed to effect the judgment even more because participants are sensitized) (Payne et al., 2003).

We adopted this task following the logic that primes associated with the military game should lead to a more positive evaluation of the subsequent targets for the participants who played this game before the procedure. Participants who played a different game with no military context should evaluate the military primes more negatively compared to the other group. We thus assume that identification with the player character leads to a positive affect towards game-related concepts. We contrasted the game-related primes with unrelated primes. These “neutral” primes were also in their affective value relatively neutral, some probably slightly positive. The assumption regarding the misattribution of those primes is that there should be no difference between participants playing the stimulus game and playing a different game. This implicit measure would thus turn out as valid instrument to assess identification processes (or, more precisely, the component of automatic concept evaluation within the overall identification construct) if subliminal primes related to a target game character are differently processed (i.e., affect AMP target evaluation differently) by players of that game compared to a control group who has not played that game (i.e., played a completely different game), whereas subliminal primes unrelated to the target game character.
should not produce any evaluation differences between players of the game and members of the control group.

**Explicit Character-Related Self-Descriptions**

With the first two parts of the measurement, we assess the cognitive availability of game-related words and the misattributed evaluation of game-related visuals. In the third step, the adoption of character-related attributes (such as heroic or brave in a military-/battle-context) is explicitly assessed through a self-description task (SDT). The logic of this measurement bases on the assumption that perspective-taking with the character while playing a video game leads to identification. This monadic identification should express itself in a temporarily altered self-conception: attributes of the character should be integrated into the self-conception while contrary attributes should be excluded. Results of four experiments conducted by Goldstein and Cialdini (2007) provide support for this assumption: Through encouraging perspective-taking with another person, participants were led to feel a sense of merged identity with introduced actors. Afterwards they filled out a questionnaire concerning their self-concept. Participants effectively ascribed attributes of the observed person stronger to themselves.

The logic of our measurement is similar, since we expect a merged identity as well. In our case, the identity of the player is expected to merge with the identity of the game character while playing. We adopted the design of the questionnaire. Participants are given an attribute list. Half of the attributes are character-related, the other half are attributes contrary to the character. Those attributes had to be rated on a 11-point scale with 1 meaning “the attribute describes me much less than the average person” 11 meaning “the attribute describes me much more than the average person” and 6 meaning “the attribute describes me about the same as the average person.

Following the present logic, participants who played the stimulus game, should rate the character-related attributes higher as describing their selves as the participants who did not play the stimulus game.

**Methods**

**Participants and Procedure**

60 voluntary male university students aged between 20 and 34 years (M = 24.97, SD = 3.32) took part in the experiment. We eliminated five participants due to their performance on the lexical decision task (> 25% of the words wrongly or too quickly/slowly assorted (<300ms or >1400 ms, respectively). 46 of the 55 analyzed participants play at least sometimes video games, almost all of them (51) at least played in the past. They all had at least some experience with the genre of the game used in the experiment, first-person-shooters (FPS).

The participants were invited individually to a quiet room with controlled lighting conditions and were asked to do a testing phase of the speeded tasks. Afterwards, they were randomly assigned to either play the first-person-shooter “Call of Duty” (N=25; see screenshot) or play the game “Tetris” (N=30).
After 10 minutes, the game session was interrupted. Participants were then requested to complete the lexical decision task, which was designed similar to Macrae et al. (1995, see above) and contained 14 words relating to the CoD-environment (such as explosion or gun), 14 neutral words (such as coffee or spring) and 28 non-words (words that can be pronounced in German easily; these words differ from real words only in one letter). These stimuli words and non-words appeared on a Computer screen for 700 milliseconds after a standardized mask (the letter sequence “XXXXXXXXXX” appeared on screen for 550 ms) and a blank screen for 600ms. The mask served to alert participants for the next upcoming decision task in order to prevent disturbing effects of attention shifts or surprise. Responses were collected from the keyboard. Stimulus presentation and reaction time recording were controlled by a real-time computer system. Participants were instructed to press the computer key “k” if they identified the presented strings as a word. If they identified it as a non-word, the key “d” should be pressed. In case of an incorrect response, the word “error” (German: Fehler) was presented on the screen. The task lasted approximately 3 minutes.

After this first speeded task, participants were asked again to play the game for another 5 minutes before the next measurement procedure was executed. The short playing phase served to reinforce the identification process that had been interrupted by the LDT. The second gaming phase was followed by the affect misattribution procedure (AMP). In advance to the study, participants were informed that the study examined “how people make simple but quick judgments”. Their job was to judge the visual pleasantness of Chinese pictograms. In advance to showing the pictograms, either a CoD-related picture (such as a picture of a pistol) respectively a neutral picture (such as an umbrella) was subliminally primed through a very short presentation on the computer screen (83.5 ms), followed by a blank screen for 133.6 ms, and then the Chinese pictograms for 167 ms. Finally, a gray pattern mask appeared until the participant responded (design based on Payne et al.; 2005). Participants were instructed to press the computer key “k”, if they judged the pictogram to be more visually appealing than average and to press the “d” if they found it less pleasing. In addition, they were asked to respond quickly and spontaneous. Primes were 16 CoD-related pictures and 16 neutral pictures that appeared in a random order. The task lasted approximately 5 minutes.

Both the LDT and the AMP tasks were completed on a different notebook from the gaming computer in order to guarantee a quick transition from playing to the task. After those tasks and after five more minutes of playing, participants were asked to fill out a questionnaire regarding their self-description on the given relevant attributes (SDT, see above; Goldstein & Cialdini, 2007). Participants rated themselves for 13 attributes that are associated with the soldier role (such as brave, courageous, and disciplined) and 13 attributes that were in contrast to the soldier role (such
as caring, warm-hearted, and friendly). Furthermore, the questionnaire assessed on a five-point-scale the following dimensions:

- enjoyment experience (3 items like “the game was entertaining; Cronbach's Alpha of mean index = .92, M = 3.59; SD = 1.01)
- perceived competence while playing (3 items on a continuum from e.g. “my performance was…very bad to very good”; Cronbach's Alpha of mean index = .84; M = 2.66; SD = 0.81)
- state of presence during the game (8 items like “I had the feeling of really being in the game”; Cronbach's Alpha of mean index = .9, M = 2.18; SD = 0.8) and

Finally, participants were informed about the research interest and the measurement in more detail. Each person received 5 EUR as compensation.

Data cleansing

For the lexical decision task, responses were counted as invalid on the level of individual trials, if the word was wrongly assorted or if the reaction time was less than 300 ms or more than 1400ms. 300ms is usually used in implicit measurement procedures as lower limit because it is assumed that reaction times below 300ms occur accidentally and do not represent a participant’s attempt to respond to the sorting task. Using 1400 ms as upper limit means to admit response times of up to two standard deviations above the empirical average response time. The exact limit value would then have been 1359ms, so the next higher rounded value was chosen, which is also an established practice for data cleansing in reaction time tasks (Perea & Carreiras, 2003).

On the level of participants, five cases were totally removed from the data due to more than 25% invalid responses (> 14 words).

A third data cleansing procedure was applied at the level of target words and non-words. Several words were eliminated for all participants, because they were assorted too often to the wrong category (more than 15 times → 25% of all observations) and could be thus identified as too difficult. This was the case for the word “Minen” (english: mines), which belonged to the Call of Duty-category and for three non-words that obviously were too similar to real words and were therefore wrongly assorted as correct words frequently (“Geheise”, “tinken” and “Maneral”).

Concerning the AMP, the same minimum and maximum response time rules were applied as with the LDT: We also cut off the trials under 300ms and those higher than the average response time plus two standard deviations, which resulted in the rounded maximum limit of 1600ms. Due to the type of the task, no incorrect answers were possible. Thus, no participants were removed from the data set.

Results

Results of the Lexical Decision Task

Basis of analysis for the LDT was comparing the two experimental groups (playing Call of Duty versus playing Tetris) regarding the mean differences of reaction times in identifying the CoD-connected words as real words. The prediction is that the CoD-players (identification condition) will display greater cognitive accessibility and thus faster decision times if words related to the soldier role appear on screen (compared to Tetris-players in the low-identification condition). In reverse, words unrelated to the soldier role (non words and neutral words) should not produce significant differences in the mean reaction time between the two experimental groups.
A repeated measure ANOVA with the word category (CoD-related vs. neutral) as within subjects factor and the experimental condition (playing CoD vs. Tetris) as between subjects factor was conducted to address the hypothesis and test the effect of the video game played on the dependent measures. As hypothesized, we observed a main effect for the between subjects factor game (F(1/53) =4.23; p<.05; $\eta^2 = .07$) and an interaction effect for the word-category and game (F(1/53) =13.66; p<.01; $\eta^2 = .21$). Reaction times to Call of Duty-related words were significantly faster recognized as neutral words by the players of CoD (M=586,86 ms, SD=76.59, results in milliseconds) than by the players of Tetris (M=655.23; SD=91.57). In contrast, there is only a marginal difference concerning the neutral words between the two groups (M=644,67 ms for the ones having played Tetris and M=622,51 ms for the ones having played CoD).

Results of the Affect Misattribution Procedure

Basis of analysis for the Affect Misattribution Paradigm (AMP) was comparing the two experimental groups (playing Call of Duty versus playing Tetris) regarding the number of Chinese pictograms judged as “visually more pleasing than average” after a subliminal CoD vs. neutral prime. Furthermore, differences in mean reaction time for judging a pictogram following a CoD-prime vs. neutral prime as pleasing were examined. We assumed that participants having played CoD would evaluate more pictograms following a CoD-prime as positively and would react faster whereas the ones having played Tetris would faster evaluate those pictograms following a neutral picture prime as pleasing.

The results cannot support our assumptions. Judgment patterns concerning the number of positively evaluated pictograms correspond to our hypotheses: The participants that played CoD evaluated more pictograms following a CoD-prime positively (M = 6.3 out of 16) than the participants that played Tetris (M = 5.7). Vice versa, they found less pictograms following a neutral prime pleasing (M = 8.1) than the ones having played Tetris (M = 8.9). However, these differences are not significant.

Concerning the reaction times, the results even contradict our assumptions: The mean reaction time for the positive evaluation of Chinese pictograms following neutral pictures was faster for the participants who played CoD (M=615 ms, SD=182ms) compared to the ones that played Tetris (M=702ms, SD=144ms; F(1/54) =3.9; p=.06; $\eta^2 = .07$). Concerning the reaction time for positive evaluations of the pictograms following a CoD-picture, no significant differences were observed (players of CoD: M=667ms, SD=185ms; players of Tetris: M=683ms, SD=158ms).

Due to the obvious failure of the AMP to reveal cognitive-evaluative residuals of identification, the data was not further analyzed.

Results of the Self-Description Task

Basis of analysis for the questionnaire-based assessment of identification with the player character was the self-description on given attributes on a 11-point scale (SDT; Goldstein and Cialdini, 2007). We assumed that participants after playing CoD would describe themselves as more soldier-like and less warn-hearted etc. than participants would do after playing Tetris. A mean indices of the respectively 13 attributes were computed and mean values regarding these indices for the both experimental groups were compared through an ANOVA. Response patterns modestly support our assumptions, but no significant differences could be observed. Players of Tetris showed a higher mean value for the “anti-soldier”-index (M = 6.7 compared to M = 6.3). There was no difference for the soldier-index, though.
Due to the fact that the selection of the attributes was rather intuitively and they probably load on different factors, an explorative factor analysis was conducted. The first factor analyses extracted 8 factors that were too many and difficult to interpret. Thus, a second analysis was conducted and number of the factors was pre-set to five (eigen-value > 1.5) and a varimax rotation was applied. Four of the five resulting factors could be interpreted and were thus transformed into the following new variables:

- **Factor 1 “Anti-soldier”:** This factor contains attributes like helpful, friendly, tolerant, and warm-hearted.
- **Factor 2 “Manful soldier”:** This factor contains attributes typical for soldiers like assertive, aggressive, and brave.
- **Factor 3 “Calmness”:** This factor contains attributes like anxious, patient, and serene on the one side, and loyal and valiant on the other side. It is thus not very easily interpretable. Due to the fact that the three most-loading attributes fit together very well, it is called “calmness” and included into further analyses.
- **Factor 4 “Virtuous soldier”:** This factor contains attributes important to be a good soldier: disciplined, persistent, and tough.

These four factors were introduced into a MANOVA with the game played as factor. Results support our assumptions: While those participants who played Tetris display higher values for the first factor “anti-soldier” and the third one “calmness”, respondents who played CoD show higher values on the two “soldier factors” “manful” and “virtuous” soldier. The differences are significant for the factor “anti-soldier” (F(1/56) = 3.7; p < .05; η² = .08) and marginally significant for the second factor “manful soldier” (F(1/56) = 2.7; p = .09; η² = .05). See Table 1 for results:

<table>
<thead>
<tr>
<th></th>
<th>CoD (n=29)</th>
<th>Tetris (n=28)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Factor 1: Anti-Soldier</strong></td>
<td>-.17</td>
<td>.34</td>
</tr>
<tr>
<td><strong>Factor 2: Manful soldier</strong></td>
<td>.28</td>
<td>-.16</td>
</tr>
<tr>
<td><strong>Factor 3: Calmness</strong></td>
<td>.01</td>
<td>.09</td>
</tr>
<tr>
<td><strong>Factor 4: Virtuous Soldier</strong></td>
<td>.07</td>
<td>-.05</td>
</tr>
</tbody>
</table>

Overall, the SDT based on Goldstein and Cialdini’s (2007) approach turned out to be a useful method to assess identification with the soldier role offered by the shooter game CoD.

**Correlations of the different identification measures**

LDT and SDT

- No correlations between the LDT and self-description were observed (cf. Discussion for discussion of this result).
- Correlations between the AMP results and other identification measures were not examined due to the failure of the AMP as an identification measure.

**Relation of identification and enjoyment**
Based on our theoretical assumptions, we expect a correlation between identification with the player character and fun of gaming. We thus computed a MANOVA with the experimental condition as factor and identification and enjoyment, assessed through self-description (mean indices; see Results of Self-Description Task, above), as dependent variables. In fact, the ones having played Call of Duty, exhibited a modest higher identification with the game world (in terms of feeling of presence and self-location; M = 2.53 compared to M = 2.34, n.s.). In contrast, the players of Tetris (M = 3.7) enjoyed the game more than the players of CoD (M = 3.48), which contradicts our assumptions. This difference also is not significant, though.

We further analyzed correlations between the implicit and measurements of identification (LDT, AMP and self-evaluation of attributes) and enjoyment ratings. Due to the fact, that we explicitly assume for the ones playing CoD to identify with the player character and thus "achieve" enjoyment, only players of CoD were included in all further analyzes.

For the LDT, no significant correlation between enjoyment and number of correctly identified CoD-words and respective reaction time were detected.

Regarding the AMP, a correlation between enjoyment and the number of positively evaluated Chinese pictograms following a CoD-prime and respective reaction time was computed. The results support our assumptions: The analyses exhibit a significant correlation of enjoyment and number of positively evaluated Chinese pictograms following a CoD-prime (r = .41; p < .05) and a marginally significant correlation of respective reaction time and enjoyment (r = .34; p = .07).

Concerning the self-evaluation of the given attributes, correlations between enjoyment and the soldier-index, the anti-soldier-index, and the four extracted factors (see Results of Self-Description Task, above) were computed. For the anti-soldier-index and the first three factors, no significant correlation was detected. There was a marginal significant correlation between enjoyment and the soldier-index (containing all soldier attributes), however (r = .35; p = .06), and a significant correlation between enjoyment and the fourth factor "virtuous soldier" (r = .40; p < .05; attributes "disciplined", "persistent", and "tough" loaded highly on this factor). Evaluating oneself as more "soldierful" after playing CoD thus is associated with more enjoyment, which supports our hypothesis.

Discussion

Findings of this multi-method study are highly informative and complex. We first discuss the implications of results in terms of measurement capacities (i.e., the ability of the tested measures to assess identification as enjoyment sub-process) and then turn to the issue of empirical relationships between identification measures and the game enjoyment rating.

The Lexical Decision Task turns out as a valid instrument to measure a cognitive residual of identification processes, namely the increased accessibility to concepts related to the role or character players have identified with. The experimental differences between shooter players and "Tetris" players indicate that LDT reflect players’ thinking in the categories of their role (i.e., the soldier role in the shooter game CoD). Digital games that attempt to create a strong identification of players with a certain role and/or character can thus be pretested by using an LDT: This method will separate those games who fail to elicit identification (as no response latency differences to a control group will emerge) from those games who achieve to establish identification (because test players of such games will differ in an LDT from a control group). If the 'right' target words are used, that is, if concepts are tested that are really relevant to the player role to be identified with, the LDT may augment explicit measures such as verbal ratings of "I felt like character X during game play". The LDT may provide a less biased and more objective part of the measurement of identification compared to explicit ratings, as social desirability (e.g., adults unwilling to reveal a 'childish' sense
of identification to an interviewer) and incomplete memory of experiences during gameplay may affect explicit measures, but not the LDT.

The AMP did not produce any meaningful experimental differences between shooter and “Tetris” players. Our approach to modify the AMP to address the identification construct has therefore failed. The interesting correlation between AMP values (the number of positively evaluated Chinese pictograms following a CoD picture) and the overall enjoyment rating ($r = .41$) suggests, however, that AMP may be useful as an implicit measure for overall game evaluation. Recent findings from the inventors of AMP (Payne et al., 2008) indicate that implicit liking of alcohol – measured through the AMP – predicts alcohol use and addictive drinking behavior. In future studies with adjusted research designs, the AMP may thus turn out as useful implicit measure of overall game appreciation or as implicit marker of game addiction. To use AMP as a general enjoyment measure would correspond with the explication of media enjoyment as attitude. Nabi and Krcmar (2004) argue for enjoyment as a positive attitude towards a media message and/or towards a media experience. AMP may thus be useful to reveal implicit overall evaluations of media experiences (as opposed to specific enjoyment components such as identification, which have been investigated in the present context). Concerning the assessment of identification with a video game character, however, the AMP should be removed from the list of candidate procedures.

The attribute-based self-description measure that we construed following Goldstein and Cialdini (2007) may serve as a useful tool to assess identification, as it produced interpretable experimental group differences. Interestingly, the measure was sensitive to the players’ experiences although it explicitly asked participants to describe themselves ‘in general’, that is, without reference to the gaming experience they had had before. In this sense, the self-description measure is also implicit, as respondents seem to be not aware of the impact the identification process during gameplay has on their responses. Future studies will have to show the stability of the findings and further substantiate the evidence for the validity of this measurement technique. Moreover, findings suggest to select role-related attributes most carefully, because not all attributes introduced in our study worked equally well.

For construct validation of the implicit measures of identification, the experimental comparison has been the main strategy within the current research design. Additional insight comes from the correlations between the (implicit) identification measures and the explicit enjoyment rating. Our theoretical base is that the stronger players identify with an attractive or interesting character/role, the more should they enjoy the game. The findings from the current study do only partially match this expectation.

- The Lexical Decision Task data do not correlate with explicit enjoyment ratings, although they are compatible with our theoretical expectations on identification (see above).
- The AMP correlates with enjoyment, but does not work effectively as measure of identification (see above).
- The self-description measure (based on Goldstein & Cialdini, 2007) did display some relevant relationships with the enjoyment ratings (between .35 and .40).

In spite of these mixed findings, our interpretation of this aspect of implicit methods’ construct validity is rather optimistic, because there are explanations for the findings that are in line with our theoretical base. First, identification has been conceptualized as one element of game enjoyment, with several other (partially additive, partially competing) factors also contributing to the fun of gaming (see Deliverable D2.1). Therefore, positive correlations between identification values and enjoyment ratings were hypothesized; however, the strength of the relationship does not need to
turn out as very great to fit our multi-causal approach to game enjoyment. In this sense, the findings from the self-description measure are considered to be perfectly in line with our framework.

The zero-correlation between LDT data and game enjoyment ratings is in need of an alternative explanation, however. One perspective is that the theoretical assumption on identification and enjoyment is invalid; however, results from the self-description measure conflict with this possibility (see above). Interestingly, results from LDT and SDT do not correlate, too, although both measures responded to the experimental variation in a meaningful way. Various reasons are possible. Firstly, concept accessibility – as measured by the LDT – may be a necessary but not sufficient condition of identification. In this case, people who did not identify with the player character, could have experienced a greater concept accessibility as well, but did not respond the SDT like the ones who identified with the player character. Furthermore, the sample size of 55 participants (for the LDT analyses) is relatively small for implementing correlation analyses.

In general, the structural difference an implicit cognitive accessibility measure on the one hand and an explicit rating measure on the other hand could be the explanation for the observed zero-correlation. Hofmann et al. (2005) present a list of moderators that affect the statistical correlation between implicit and explicit measures. They expect positive, but modest correlations between implicit and explicit data, however, their review focuses on implicit versus explicit measures of the same construct (e.g., implicit and explicit stereotypes or self-esteem). In the present case regarding the correlation between concept accessibility and enjoyment, we have modeled a partial overlap between variance in identification and variance in game enjoyment (due to the simultaneous effect of other determinants of fun such as performance or curiosity). Implicit identification measures and explicit enjoyment ratings therefore do not address the same construct; rather, they are intended to measure different concepts that we expect to be correlated. From this perspective, a moderately positive correlation between identification and enjoyment would be expectable; this moderate correlation may then be suppressed by the implicit-explicit divergence. That is, an explicit enjoyment rating should correlate more strongly with an explicit identification measure than with an implicit identification measure; in turn, the implicit identification measure should correlate more strongly with an implicit enjoyment measure than with an explicit enjoyment rating. While our research design does not allow to establish a full implicit / explicit x identification / enjoyment measurement matrix to test this argument with sufficient rigor, the interpretation of our findings concerning the LDT is that the zero correlation with the explicit game enjoyment rating does not disqualify the LDT as construct-valid measure; rather, replication studies are needed to better understand the interplays (1) between concept accessibility, identification, and game enjoyment and (2) between implicit and explicit cognition in digital game appreciation.

Our overall reading of the reported study is thus modestly optimistic. The fun of gaming is, as a multidimensional and dynamic phenomenon, a challenge for any scientific method mix. Implicit measures are a challenging class of methods per se; the current experiment allowed to differentiate between useful (LDT, self-description task) and invalid (AMP) identification measures; further work will be required to contextualize implicit measures within the enjoyment framework. Specifically, a new implicit measure to replace the AMP should be tested; this measure should tap one key element of the identification concept that is out of reach for accessibility measures such as the LDT, namely the strength of associations between character-related concepts and players’ self. Moreover, a broader account of explicit game evaluation measures should be pursued than the three-item scale on game enjoyment used in the present experiment, the Game Experience Questionnaire developed by partner TUE is promising to be more informative in this regard. In addition, implicit methods’ usefulness in applied (i.e., industrial product testing) research will have to be evaluated; insights from the present study and further research needs to be discussed in terms
of practicability and efficiency. In sum, the reported study marks a significant step towards the
definition of implicit measurement paradigms in the context of identification with a role or game
character; however, some replication and extension is indicated to achieve fully satisfying results on
construct validity.

2.2.2 2nd Construct validity study of identification

Introduction

The 2nd construct validity study on identification was based on the results of the 1st study of
identification processes. In this last study, the Lexical Decision Task and the Self-Description Task
turned out to be useful measurements of (components of) identification. In contrast, the Affect
Misattribution Procedure failed to assess the evaluation of word-related concepts. Due to these
results, some changes in the design were indicated. In the following paragraphs, the design and
procedure of the experiment, as well as results and a discussion are reported.

Design of Implicit Measurement of Identification

Like in the first study, the measurement-procedure contains three components. The Lexical
Decision (Macrae, Bodenhausen & Milne, 1995, see above for detailed description) task turned out
to be a valid approach to assess the cognitive availability of character-related concepts and was used
again. Furthermore, the third component of the preceding study, the explicit character-related self-
descriptions (following Goldstein & Cialdini, 2007), were kept. Instead of trying to assess positive
automatic evaluations of game-related concepts through an Affect Misattribution Procedure (AMP;
Payne, Cheng, Govorun & Stewart, 2005, see above for detailed description), an Implicit
Association Test (IAT; Greenwald, McGhee & Schwartz, 1998) was applied in order to assess the
strengths of associations between the player’s self and the concepts of the game he was playing.

Methods

Participants and Procedure

The experimental design of this study aimed to offer two different roles to identify with. The
experimental factor “identification role” had two values: “soldier” (manifest in the shooter game
“Call of Duty 2” that had been used before) and “race driver” (manifest in the racing game “Need
for Speed Carbon”). Construct validity of implicit measures was thus to be tested by exploring
whether the measures would produce differences in players’ implicit cognition that could be
interpreted either as residual of identification with a soldier or with a race driver, respectively.

61 male university students aged between 19 and 31 years (M = 22.61, SD = 2.75) volunteered to
take part in the experiment. 59 of them play at least sometimes video games. The participants were
invited individually to a quiet room with controlled lighting conditions were randomly assigned to
either play the first-person-shooter “Call of Duty2” (N=30) or play the game “Need for Speed
Carbon” (N=31).

After playing the game for 10 minutes, participants were kindly interrupted and asked to complete
the first implicit identification measure, i.e. the Lexical Decision Task (see below for a more
detailed description). They then played the game for further five minutes and then completed the
second measure– the Implicit Association Test. After another five minutes of playing, participants
filled out a questionnaire containing the Self-Description Task and several game-related dimensions
(see below). After the procedure, participants were debriefed and dismissed. Each person received 10 EUR as compensation.

**LDT**

Following the lexical decision task procedure described by Macrae et al. (1995), a lexical decision task was implemented to measure cognitive accessibility of character-related concepts in video game players. Character-related concepts come from military contexts, such as gun, crossed wires, or ammunition. Specifically, a set of 7 words with connection to the soldier role (such as gun, crossed wires, or munitions), 7 words with connection to the role of a car racer (such as exhaust or steering wheel) and 7 neutral words (such as paper or spring) unrelated to that role were presented to participants in random order. The same number of non-words – (i.e. 21; each created by changing one letter in a real word, which renders the non-words quite similar to words and thus increases the difficulty of the task) was added to the list of targets appearing on screen. Each word / non-word appears on screen for 75 milliseconds. Participants are requested to press either ‘d’ if the presented word is a real word or ‘k’ if the word is a non-word.

**IAT**

Implicit Association Tests measure the strengths of associations among concepts (Nosek, Greenwald & Banaji, 2007). During the task, participants are to sort words belonging to one of four concepts with two response options on the keyboard, meaning that in each case two categories have to be responded with pushing the same key on the keyboard. To apply the IAT principle to the measurement of identification (as explicated in the theory work package, see D3.1.), the proposition of identification as altered self-perception is translated into an increased (activated) association between game players’ self and attributes that describe the character or role with which players (are presumed to) identify. For instance, if players identify with a soldier, they should display an activated stronger association between the concepts “me” and soldier-related concepts such as “war” or “gun”.

The categories in this study are “me” and “furniture” as target categories and “military” and “car racing” as attribute category. In one turn, participants have to push the key “E”, if a “me-related” word (like “mine” or “self” appears on the screen or if a military-word (like “soldier” or “pistol”) appears on the screen. Vice versa, they have to respond to furniture words (like table or bed) or car racing-related words by pushing the key “I”. A reverse block then combines “me” and “car racing” on the key “E” and “furniture” and “military” on the key “I”. Orders of the blocks are interchanged. Furniture were chosen as the “mirror-category” due to its neutrality. Response times for sorting are recorded and compared for the block that combines “me-related” words and “military” against the block that combines “me-related” blocks and “car-racing”. The interpretation of response latency patterns is that (substantially) faster sorting performance when “me” and “military” are paired is an indicator of a strong association between these categories. This pattern is assumed for participants playing the military game “Call of Duty 2” (CoD) versus the opposite is assumed for those who played the car racing game “Need for Speed Carbon” (NFS).

**SDT**

With the first two parts of the measurement, cognitive availability and the association of game-related concepts with the players’ self is assessed. In the third step, the adoption of character-related attributes (such as patriotic and brave for the CoD-character and fast and ruthless for the NFS-character) is explicitly assessed through a self-description task (SDT, for a detailed description see
The logic of this measurement bases on the assumption that perspective-taking with the character while playing a video game leads to adoption of attributes of the character identified with.

The attribute list for the participants contained 10 traits. Four of them were Soldier-related (CoD), three were car racer-related (NFS) and three were contrary to traits typically ascribed to soldiers and car-racers (i.e. helpful, shy, and tolerant). The list was carefully compiled after analyzing the preceding study on identification. The attributes had to be rated on a 11-point scale with 1 meaning “the attribute describes me much less than the average person” 11 meaning “the attribute describes me much more than the average person” and 6 meaning “the attribute describes me about the same as the average person.

Following the present logic, participants who played the CoD, should rate the soldier-related attributes higher, while the participants who played NFS, should rate the car-racer attributes higher. There should be no difference in rating the “neutral” attributes.

**Questionnaire**

Furthermore, the questionnaire assessed on a five-point-scale the following dimensions:

- enjoyment experience (3 items like “the game was entertaining”; Cronbach's Alpha of mean index = .88; M = 3.94; SD = 0.74)
- explicit identification with the game character (3 items, e.g. “I almost felt as if I was the character”; Cronbach's Alpha of mean index = .74; M = 2.21; SD = 0.85; one item was deleted due to insufficient correlation with the other items)
- state of presence during the game (3 items like “I forgot everything around me while playing the game”; Cronbach's Alpha of mean index = .82 M = 2.85; SD = 1.07) and
- the short version of the GEQ questionnaire (developed by FUGA partner TUE)

**Data cleansing**

An elaborate data-cleansing procedure was applied regarding the implicit measurements. It was defined following previous work with the Implicit Association Test (Greenwald et al., 1998) and the Lexical Decision Task (Macrae et al., 1995) and studies with similar procedures. Data were cleaned for both measures once in advance to the aggregation of the data (1 case = 1 reaction time) and after the aggregation (1 case = 1 participant). Regarding the IAT, all reaction times (N = 136) more than twice the standard deviation above or below the mean reaction time (761.44 ms, SD = 341.67) or less than 300 ms were excluded from the data set. Furthermore, all wrongly assorted trials were eliminated (N = 362).

After aggregating the data, all mean values of compatible and incompatible trials, as well as the IAT-Effect, were excluded per person, if they were lower or higher than twice the standard deviation. Thus, the data-set for the IAT containes 58 participants.

Concerning the LDT, the nonword “Giffel” was excluded as a whole due to its wrong assignment of more than 25% of the participants (35%). All incorrect trials were excluded as well (N=323) and all trials were eliminated with a value lower than 300 ms (29 trials) or higher than twice the standard deviation of the mean (M=863.79, SD=1679.26; 28 values excluded).

After aggregation, all mean values per person were excluded that had more than 25% of the words or nonwords wrongly assorted in the regarding category (CoD-words, NFS-words, neutral words, or nonwords), which was for five mean values the case. Furthermore, all mean reaction times were
excluded that were above twice the standard deviation of the regarding word category (in total 13 values over the four word categories).

Results

Results of the LDT

Basis of analysis for the LDT was comparing the two experimental groups (playing Call of Duty 2 versus playing Need for Speed Carbon) regarding the mean differences of reaction times in identifying the CoD-connected words as real words. The prediction is that the CoD-players will display greater cognitive accessibility of CoD-related concepts and thus faster decision times if words related to the soldier role appear on screen. In contrast, those playing NFS should display faster reaction times for car racing-related words appearing on the screen. Words unrelated to the soldier role (non words and neutral words) should not produce significant differences in the mean reaction time between the two experimental groups. Mean values are displayed in Table 2.

Table 2: mean differences of reaction times for CoD words, NFS words, neutral words and nonwords

<table>
<thead>
<tr>
<th></th>
<th>CoD words</th>
<th>NFS words</th>
<th>Neutral words</th>
<th>Non-words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>630.66</td>
<td>647.74</td>
<td>671.56</td>
<td>707.73</td>
</tr>
<tr>
<td>SD</td>
<td>119.3</td>
<td>132.02</td>
<td>164.11</td>
<td>122.55</td>
</tr>
<tr>
<td>Mean</td>
<td>713.62</td>
<td>689.51</td>
<td>659.57</td>
<td>758.62</td>
</tr>
<tr>
<td>SD</td>
<td>164.11</td>
<td>122.55</td>
<td>123.71</td>
<td>124.25</td>
</tr>
</tbody>
</table>

N = 49

A repeated measure ANOVA with the word category (CoD-related vs. NFS-related) as within subjects factor and the experimental condition (playing CoD vs. NFS) as between subjects factor was conducted to address the hypothesis and test the effect of the video game played on the dependent measures. Mean effects and the interaction effect were not significant. However, the result pattern displays the assumed direction on a descriptive level: While the CoD players are faster at recognizing CoD words as real words than the NFS players, NFS players react faster to the NFS words. These differences are relatively small though.

Results of the IAT

The IAT detects strengths of associations between four pairings of concepts. In our case, these four pairings are “me” + “military”, “me” + “car racing”, “furniture” + “military”, and “furniture” + “car racing”. The trials were named from the view of the CoD-players: “me” + “military” and “furniture” + “car racing” are the compatible trials, while “me” + “car racing” and “furniture” + “military” are incompatible trials. For people who had played CoD, we assume fast reaction times for the compatible trials and slow reaction times for the incompatible trials and for people having played NFS, we hypothesized the opposite result pattern. Key information of the IAT is the so-called IAT effect, which is the difference between incompatible and compatible trials (incompatible minus compatible). High positive IAT-Effects indicate a strong identification with the player character for the CoD-players and low negative IAT-effects indicate a high identification with the player character for the NFS-players.

An ANOVA was computed to detect differences between the two player groups. Results support our assumptions. While the IAT-effect for the CoD-players is 21.53 ms (SD = 64.32), it is -61.4 (SD = 67.59) for participants who had played the car-racing game NFS game (F(1/56) = 22.91; p < .001; η² = .29).
Results of the SDT

Basis of analysis for the questionnaire-based assessment of identification with the player character was the self-description on given attributes on a 11-point scale (SDT, cf. Goldstein and Cialdini, 2007). We assumed that participants after playing CoD would describe themselves as more soldier-like (i.e. as more patriotic, brave, allegiant, and duteous), while participants after playing NFS would evaluate themselves higher for attributes suitable to a car racer (i.e. fast, ruthless, and impulsive). In order to present a more complete list, three neutral words were included in the list as well (i.e. helpful, shy, and tolerant). For the neutral words, no differences were expected regarding the two stimulus games.

Both attribute lists displayed low internal consistencies (Cronbach’s Alpha of .47 for the soldier items and .11 for the car racer items). Therefore, all attributes were introduced individually in a multivariate ANOVA. No significant differences in mean values were detected and the descriptive result pattern was rather chaotic. Due to these results and to the fact that the selection of the attributes was rather intuitively, an explorative factor analysis was conducted. Only the seven game-related words were introduced into an explorative factor analyses with varimax rotation. Three interpretable factors with an eigenvalue of >1 were extracted and transformed into the following new variables:

- Factor 1 “Soldier” with the attributes allegiant and brave loading positively on it and the attribute ruthless loading negatively.
- Factor 2 “Emotionality” with the attributes impulsive and patriotic loading positively on it.
- Factor 3 “Speed”: This factor contains only one attribute, namely “fast” which loads positively on the third factor.

While the first factor is explicitly a factor containing soldier-attributes and the third one refers to a car-racer, the second one contains both a trait referring to the role of a soldier (“patriotic) and a trait referring to a car-racer (“impulsive”). We would thus not assume a difference for the second factor. The results of the factor analyses thus shows that our rather intuitively derived attributes might not reflect attributes that really can separate between the two roles.

The three factors were introduced into a MANOVA with the game played as factor. No significant differences were observed. Descriptive results partially contradict our assumptions: While those participants who played CoD displayed lower values for the first factor “soldier” and higher values for the third factor “speed”, respondents who played NFS displayed higher values on the “soldier” factor and lower values on the “speed” factor. Regarding the second factor “emotionality”, the ones having played CoD display higher values.

Correlations between identification measures

Participants who played CoD are assumed to display faster reaction times for compatible trials (IAT) and faster reaction times for CoD-related words (LDT). In reverse, those who played NFS, should display faster reaction times for incompatible trials and NFS-related words. Thus, reaction times of compatible trials (IAT) should correlate positively with reaction times for CoD-words and reaction times of incompatible trials (IAT) should correlate positively with reaction times for NFS-words. Results of a correlation analyses support our assumptions – despite non-significant results for the LDT: The mean reaction time of compatible trials correlates positively with the mean reaction time of CoD words ($r = .47; p< .01, N = 50$) and the mean reaction time of incompatible trials correlates positively with the mean reaction time of NFS words ($r = .34; p< .05, N = 50$).

Due to the mentioned problems with the Self-Description-Task, no correlations were analyzed.
The implicit-explicit-differentiation was addressed through a correlation analyses between IAT measures and the explicit identification index (see above). Analyses were computed groupwise due to oppositional assumptions for the CoD and NFS players (For the CoD players, a positive correlation between explicit measures and compatible trials is assumed; for the NFS players, a positive correlation between explicit measures and incompatible trials is assumed). We know from the preceding study and from social-psychological research (Hofmann et al., 2005) that there are many moderators that affect the statistical correlation between implicit and explicit measures. Due to these findings, slight to moderate correlations are expected. Analyses do not display any significant correlations, however.

Relation of identification and enjoyment

We further analyzed correlations between the IAT, explicit identification ratings and game enjoyment. Due to the fact, that we assume for the ones playing CoD low reaction times for compatible trials and for the ones playing NFS the opposite, analyses concerning IAT-data were computed with a divided dataset for the two experimental groups. Regarding the IAT, no significant correlation between enjoyment and compatible or incompatible trials were detected for either group. Explicit identification and enjoyment ratings, though, do correlate significantly ($r = 0.39; p<.01; N=61$).

Discussion

The second construct validity experiment on implicit measures of identification has again contributed to the theoretical and methodological progress of the project. Most importantly, the Implicit Association Test turns out as a valid instrument to measure identification with a player character. It is especially valuable because its outcome variable is a more or less direct empirical translation of the core theoretical assumption of HMTH’s identification model. The IAT measures the strength of associations between the players’ self and character-related concepts and thus the (change of) player’s self perception under influence of the character they identify with. The experimental differences between shooter players and car racing players in the present experiment indicate that IAT reflect players’ adoption of role-related concepts in their own self-perception like we described in our theoretical considerations (see deliverable D3.1 and D3.2): Identification in terms of a merged identity with – e.g. – a soldier in CoD leads to stronger associations between “me” and typical character attributes of a soldier. These strengths of associations turned out to be measurable through the Implicit Association Test. The results thus qualify the IAT as valid measurement approach for the concept of identification.

Regarding the Lexical Decision Task, results are less convincing than in the preceding study. Results display the assumed directions, but do not differ significantly, though. These results might be attributed to procedural design problems: In contrast to the first study with 16 words per category, this time only 7 words per category were presented. Furthermore, not “identification vs. non-identification” was tested like in the first study, but “identification with a soldier vs. identification with a car racer”. Maybe these two roles are more alike than assumed so that identification with one of the roles also (partly) activates concepts that we had assigned to the other role, so the LDT would not produce significant group differences as it did in the preceding study where the control group played “Tetris” and simply could not identify with any reasonable role or character. Finally, words were again chosen on an intuitive and theoretical basis, so their choice may diverge from the semantical structures of our participants Qualitative piloting of role-related attributes and concepts seems to be necessary for further applications of LDT in the identification context. One empirical argument for the usefulness of the LDT that comes out of the study is the
interpretable correlation between LDT data and the experimentally validated IAT. Together with the clearer findings from the preceding study, LDT is thus overall evaluated as a second useful implicit measurement of identification.

The Self-Description Task did not provide the assumed results. Basically two reasons can account for these results. Firstly, attributes were again chosen intuitively and maybe did just not reflect attributes enough related to the character roles. The low internal consistency of the sub-lists we had generated to describe characteristics of the soldier and the driver roles indicate such problems in attribute selection that were already found in the previous experiment. Moreover, the non-satisfying SDT results may mirror the same problem discussed concerning the LDT, namely the potentially too similar role characteristics of race drivers and soldiers – both need to be fast, brave, and ruthless, for instance. For the IAT that puts the two compared roles as opposing categories, such overlaps of role characteristics would be less of a problem; LDT and SDT are obviously more susceptible to this facet of the research design. Just like for the LDT, qualitative role analysis and identification of distinct role characteristics seem to be necessary in advance before the content of an SDT attribute list can be defined in further identification studies. With these limitations and based on the more convincing results of the SDT in the previous experiment, the SDT is planned to be further investigated in the WP6 replication studies (with according modifications, however).

Finally, we did not observe correlations between implicit identification measures and game enjoyment. This zero-correlation was already observed in the previous study and can be explained through an expectable implicit-explicit divergence (see above) deriving from the structural difference of these two measurement types (Hofmann et al., 2005). Again, parallel to the first study, a moderate correlation of implicit and explicit measures can be expected for the same construct (e.g., implicit and explicit stereotypes or self-esteem). In the present case regarding the correlation between strengths of associations and enjoyment, we have again modeled a partial overlap between variance in identification and variance in game enjoyment (due to the simultaneous effect of other determinants of fun such as performance or curiosity). Implicit identification measures and explicit enjoyment ratings therefore do not address the same construct; rather, they are intended to measure different concepts that we expect to be correlated. From this perspective, a moderate correlation may be suppressed by the implicit-explicit divergence. As we already stated above, an explicit enjoyment rating should correlate more strongly with an explicit identification measure than with an implicit identification measure; in turn, the implicit identification measure should correlate more strongly with an implicit enjoyment measure than with an explicit enjoyment rating. In contrast to the first study, our present research design allows to test at least the correlation between explicit enjoyment and explicit identification and in fact detects a significant correlation of $r = .39$. Considerations following the first study could thus been confirmed with results of this follow-up study.

Our overall reading of the reported study is thus fully optimistic. The IAT turned out to be a valid and useful measure, the LDT has to be technically improved, but also seems to be useful – especially in regard to the first study. Moreover, theoretical assumptions could at least be confirmed with the explicit data: Identification with the player character actually contributes to game enjoyment. In sum, the reported study marks a further significant step towards the definition of implicit measurement paradigms in the context of identification with a role or game character and can be replicated in WP6.
2.2.3 *1st* Construct validity study of self-esteem

**Introduction**

The basic assumption for the self-esteem research of HMTH within the FUGA project is that playing video games allows to resolve tasks and challenges, with positive impact on self-esteem and accompanying enjoyable emotions (pride; see D2.1; D3.1). Implicit measures of self-esteem were thus investigated in order to see if they would be able to detect game-induced changes in self-esteem. On the basis of our experiences from the pilot studies of self-esteem processes, a new set of measurements was designed in order to assess self-esteem implicitly. In the following sections, the design and procedure of the experiment are described; we then present the key findings and discuss them in the overall context of the FUGA project.

**Design of Implicit Measurement of Self-Esteem**

The assessment of self-esteem is – as well as the assessment of identification – laid out as a multi-component procedure. In their research project of the implicit assessment of personality, Egloff, Schmukle and Back (http://www.uni-leipzig.de/~diffdiag/projekte/index.html#indsw) found out that the frequently used self-esteem IAT (Greenwald & Farnham, 2000) is not stable over time, but rather state-sensitive, which makes it useful for our purposes and thus constitutes the first step of our measurement. As a second step, the *Name Letter Evaluation Task* (NLET, e.g. Koole, Dijksterhuis & Knippenberg, 2001; Nuttin, 1985; Nuttin, 1987) is integrated in the design as questionnaire-based measurement also assessing implicit self-esteem. Both the Self-Esteem IAT and the NLET have proved to be the most reliable and valid implicit measurements of self-esteem (Karpinski, 2004). As third part of the measurement approach, explicit self-esteem is assessed questionnaire-based through items regarding the overall (trait) self-esteem (i.e., the Rosenberg Scale, which should be stable over time; cf. Rosenberg, 1965). In addition, further relevant dimensions like feeling of competence (from the Game Experience Questionnaire developed by FUGA partner TUE) are measured.

**Self-Esteem IAT**

The self-esteem IAT (Farnham, Greenwald & Banaji, 1999) is a specific version of the Implicit Association Test (Greenwald et al., 1998). Its logic is to set a two-dimensional sorting task with the dimensions *valence* (positive-negative) and *personal reference* (me-not me). Targets appearing on screen must be assigned to a left key if they belong to “positive” or “me” and to a right key if they belong to “negative” or “not me”. A reverse block then combines “negative” and “me” on the left key and “positive” and “not me” on the right key. Response times for sorting are recorded and compared for the block that combines positive valence and personal reference against the block that combines negative valence with personal reference. The interpretation of response latency patterns is that (substantially) faster sorting performance when “me” and “positive” valence are paired is an indicator of high implicit self-esteem, as the short reaction times reflects strong cognitive associations between respondents’ self-concept and positive concepts. In turn, fast sorting performance in trials when “me” and “negative” valence are paired would suggest the interpretation of low self-esteem, as they stand for stronger cognitive associations between respondents’ self-concept and negative concepts.

The IAT is probably the most widely used implicit measurement paradigm (e.g. Greenwald, Nosek & Banaji, 2003). Applications to various target constructs (e.g., anxiety: Egloff & Schmukle, 2002) and variations of the procedure itself (e.g., De Houwer, 2003; Karpinski & Steinman, 2006) have been published. Although substantial criticism has been expressed as well (e.g., Fiedler & Bluemke,
FUGA the fun of gaming: measuring the human experience of media enjoyment Report D5.1

2005; Fiedler, Messner & Bluemke, 2006), the IAT has received more credit than any other implicit approaches, which are frequently launched in the dynamic-competitive community of social cognition research. However, IAT-based measures typically address individual differences (traits), which render their applicability in the current context questionable. Fortunately, there are two lines of research indicating the potential usefulness of the self-esteem IAT for FUGA’s measurement task. One is the criticism within social psychology that the IAT is not a robust measure of trait variables, because it is susceptible to state influences (e.g. Glen & Banse, 2004). What is a problem for trait measurement could turn out as asset for state measurement, in the current case make the self-esteem IAT useful to detect situational variation in implicit self-esteem. The other line of research has already used IAT-based measures to assess state variables. For instance, Uhlmann and Swanson (2004) used an IAT to measure aggressive self-cognitions after experimental play of violent versus non-violent video games. From this research, encouraging arguments for the potential viability of the IAT as a state-sensitive measure for game-based self-esteem can be extracted. Therefore, the first component of our implicit measurement approach for self-esteem as part of game enjoyment is the application of a conventional self-esteem IAT (following Banaji et al., 1999).

**Name Letter Evaluation Task**

The Name Letter Evaluation Task (NLET; cf. Kitayama & Karasawa, 1997; Koole, Dijksterhuis & Knippenberg, 2001; Nuttin, 1985; Nuttin, 1987) bases on the observation that people favor letters belonging to their names – and especially their initial name letter - over letters that are not part of their names. This preference for the “own” letters in general is stable over time but its degree depends on the state of self-esteem, as Koole et al. (2001) could demonstrate in their experiments. We adopted the NLET for the German language and design and let the participants rate on a 9-point-scale how much they liked every letter of the alphabet, showing one after another in a random order. Following Nuttin (1985) and Koole et al. (2001), participants were told that this study is – amongst other things – concerned with people’s aesthetic judgments of simple stimuli, that is, letters of the alphabet. It was explained that these kinds of judgments contribute to a better understanding of certain aspects of emotions and that they were probably not accustomed to tasks like that. Participants were instructed to answer as spontaneously as possible.

**Explicit Self-Esteem Scales**

The third component of the self-esteem measurement approach within the FUGA project is an explicit approach that serves purposes of implicit-explicit differentiation (see Hofmann, Gschwendner, Nosek & Schmitt, 2005) and cross-validation (with moderate correlations between implicit and explicit measures indicating validity of the implicit measures).

For explicit measurement, items regarding the feeling of competence and success and the relation of success to the difficulty of the game and own performance and capacity were developed. Furthermore, the trait self-esteem scale introduced by Rosenberg (1965) was used. The reason for combining state and trait accounts is to test for the state-sensitivity of the self-developed items and control for trait self-esteem that is likely to moderate situational responses to success and failure in the game situation (Klimmt & Hartmann, 2006).
Methods

Participants and Procedure

51 voluntary male university students aged between 20 and 30 years (M = 22.4, SD = 2.14) took part in the experiment. We eliminated one participant due to technical problem during the experiment. 44 of the 50 analyzed participants play at least sometimes Computer, almost all of them (48) at least played in the past. They all had some experience with the genre of the game used in the experiment, first-person-shooters (FPS). Before we invited the students to participate, we asked them to rate themselves on a 10-point-scale, if they were rather novices or experts regarding playing FPS (1 meaning being a novice with almost no experience, 10 meaning being an absolute expert. Respondents were then assigned to one of the stimulus games:

- Participants with expert ratings between 1 and 4 (novices) were randomly assigned to either the easy or the medium difficulty version,
- Participants with expert ratings at least 5 (experts) were randomly assigned to either the medium difficulty or the extremely difficult version of the game.

This way, an experimental variation of game difficulty was realized that took player skill into account (Skalski, Bracken & Tamborini, 2005). The participants were individually invited to a quiet room with controlled lighting conditions and were asked to do a testing phase of the speeded tasks. Afterwards, they were randomly assigned to one of the two conditions “easy level” or “difficult level” of a mod of Half-life (the different levels were created by FUGA partner HGO (see screenshot below); we thankfully acknowledge HGO’s support and collaboration for this experiment,). Depending on their self-rated expertise, rather “novice” participants in the “easy” stimulus group played a very easy level. In this level, it was almost impossible to get hurt or to die. Experts in the “easy” condition played a level with medium difficulty, which was supposed to provide them with some success and the feeling of competence. The same level was the “difficult” level for the novice participants and was supposed to let them experience great failure and thus a decrease of self-esteem. The difficult level for the experts was even very much harder and hardly to accomplish. The following table shows the allocation of novice and expert participants on the conditions.

<table>
<thead>
<tr>
<th>Table 3: Participants per condition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experimental Condition: Game Difficulty</strong></td>
</tr>
<tr>
<td>Easy</td>
</tr>
<tr>
<td>Novices</td>
</tr>
<tr>
<td>Experts</td>
</tr>
<tr>
<td>Altogether</td>
</tr>
</tbody>
</table>

After the training phase, participants played the level for 10 minutes. They were then requested to complete the self-esteem IAT (procedure described above). The IAT was followed by the questionnaire starting with the NLET (cf. above).

Furthermore, the game-related self-esteem was assessed as questionnaire-based measure of the proximate and explicit self-esteem (items as "I am proud of my performance in the game" or “I am ashamed of having performed so badly in the game”). Finally, some further relevant dimensions were assessed in the questionnaire: The Rosenberg scale of self-esteem, the competence items of
the GEQ-questionnaire, enjoyment experience and items of the perfectionism scale (dimension “Striving for excellence”; Hill et al., 2004) were assessed and mean indices were computed for further analyses (cf. Descriptive results and indices, below). After the procedure, participants were debriefed and dismissed. Each person received 5 EUR as compensation.

Data cleansing

As already mentioned above, one participant was excluded from data analyses due to technical problems during the experiment. Further three participants were excluded for data analyses of the IAT due to extremely high or low values of self-esteem in the assessment prior to playing (as they scored about two standard deviations below/above the mean value of the IAT effect). For analyzing the NLET and the other questionnaire-based dimensions, the dataset with 50 participants was basis of analysis.

Descriptive results and indices

To prepare the analysis of IAT data, mean indices were build for the pretest “compatible” reaction times, that is, assorting positive words and “me” and negative words and “not me” (M = 746.24 milliseconds; SD = 115.6 ms) and for the pretest “incompatible” reaction times, that is assorting negative words and “me” and positive words and “not me” (M = 954.88 ms; 133.52 ms). Analogous indices were built for the compatible (M= 724.24 ms, SD = 110.74 ms) and incompatible (M = 905.03 ms, SD = 143.07ms) post-test reaction times. Subsequently, the Pre-IAT effect was computed by subtracting the average compatible reaction time from the average incompatible reaction time (M = 202.48 ms, SD = 116.49 m). The Post-IAT-effect was computed in the same way (M = 180.74, SD = 106.76).

Preparatory analysis of the NLET data required two steps: Initially, the average difference between the evaluation of each participant’s initial letters (M = 6.76; SD = 1.86) and of all remaining letters of the alphabet (M = 4.8; SD = .65) was computed. Furthermore, an alternative algorithm computed the difference between the letters composing the forename of each participant (M = 5.4; SD = 1.23) and all letters that are not part of the forename (M = 4.78; SD = .71). The Initial NLET-effect over
all participants has a mean value of 1.96 (SD = 1.66), while the average Forename NLET-effect is 0.62 (SD = 1.14).

Regarding further dimensions such as experienced competence and enjoyment, internal consistencies were determined and mean indices computed. Results are presented in Table 4.

Table 4: Indices

<table>
<thead>
<tr>
<th></th>
<th>Cronbach’s Alpha</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competence, GEQ-items (without item “I felt strong”)</td>
<td>.82</td>
<td>2.45</td>
<td>.77</td>
</tr>
<tr>
<td>Rosenberg scale</td>
<td>.83</td>
<td>4.19</td>
<td>.6</td>
</tr>
<tr>
<td>Proximate self-esteem</td>
<td>.8</td>
<td>3.46</td>
<td>.82</td>
</tr>
<tr>
<td>Enjoyment</td>
<td>.94</td>
<td>3.46</td>
<td>1.12</td>
</tr>
<tr>
<td>Striving for excellence</td>
<td>.78</td>
<td>3.12</td>
<td>.7</td>
</tr>
</tbody>
</table>

N = 50

Results

Results of the Self-Esteem IAT

To detect changes in implicit self-esteem induced by the stimulus game, the key information of the self-esteem IAT, the so-called IAT effect, was computed for the pre-game and post-game tests. Strong positive IAT effects indicate a high self-esteem. The difference between pre-game and post-game IAT effect was used for analysis. Players of different skill levels (novices versus experts) and players confronted with different game task difficulties (easy versus challenging) were compared in respect to this difference. Figure 1 indicates some interesting results. First, all participants produced lower self-esteem after playing than before. However, this is likely to be a measurement artifact, as participants became faster in responding to the IAT’s sorting task. Training effects seem to be stronger concerning incompatible tasks, as they are intuitively ‘harder’ (i.e. people may notice they need more time to make counter-intuitive decisions such as linking the target word “my” to a category that includes negative valence; they may then try to improve especially on those trials, as they perceive themselves to perform slower than average there). If participants improve through training on incompatible trials to a larger degree than they improve in compatible tasks, a reduced IAT effect will result, which we consequently interpret as measurement artifact instead of a general decrease in implicit self-esteem and thus produced less pronounced discrepancies in response latency for ‘compatible’ and ‘incompatible’ trials.

Two other observations in figure 1 are more relevant, however: Experts tended to display a higher implicit self-esteem before playing than novices did. Second, only the experts confronted with an easy game version were found to change their implicit self-esteem due to the game session, whereas all other groups (novices in both difficulty conditions, and experts confronted with the challenging game version) produced similar patterns. The drop in IAT effect values (between the two measurement occasions) is twice as strong for the experts who had played the easy game version than for the other respondent groups. A repeated measures ANOVA with the self-esteem IAT effect value before and after playing as within-subject factor and the dichotomized game expert/novice self-description as well as the game difficulty level (easy versus challenging) as between-subject
factors did not reveal the assumed interaction of the three factors as statistically significant, however, \( (F (1,21) = 2.12, p = .16, \eta^2 = .09) \).

**Figure 3:** Players’ implicit self-esteem before and after playing a “Half-Life 2” level (higher values reflect higher implicit self-esteem)

### Results of the Name Letter Task (NLET)

Results of the name-letter task mirror the latter finding: “Half Life 2” experts who had played the easy game version revealed a lower self-esteem in this test after playing than all other groups. An ANOVA with player expertise and game difficulty as between-subjects factors did not confirm the interaction effect as statistically significant \( (F(1,44) = 1.4, \text{ns}, \eta^2 = .03) \). However, after splitting the data file in experts and novices, a marginally significant difference for experts playing the easy \( (M = 1.1, SD = 2) \) versus difficult game version \( (M = 2.24, SD = 1.06) \) is revealed for the Initial NLET-effect \( (F(1,46)= 3.13, p<.10, \eta^2 = .12) \).

### Results of the Questionnaire-based Self-Esteem

The results of the questionnaire-based assessment show an interesting pattern as they are not parallel to the implicit measurements (see figure 2): Both novices and experts displayed a higher self-esteem after having played the easy game than after having played the difficult game. The difference is higher for novices \( (M = 3.92 \text{ compared to } 3.13) \) than for the experts \( (M = 3.65 \text{ compared to } 3.38) \). The main effect for the game stimulus is significant \( (F(1, 43)=9.64; p<.01; \eta^2 = .18) \).
Figure 4: Players’ explicit game-related self-esteem after playing a “Half-Life 2” level

Correlations of Self-Esteem and Enjoyment

The game enjoyment ratings, finally, show a somewhat different pattern: Game experts found the challenging level most enjoyable, whereas experts playing the easy level reported lower fun values that mirrored those enjoyment ratings from novices who did not produce a difference between game difficulty levels; rather, novices found both the easy and the challenging levels equally enjoyable (Figure 3). The interaction of player expertise and game difficulty did not reach significance in an ANOVA, however (F = 1). Post-hoc t-Tests did also not confirm the difference between experts who had played the easy game and the other groups as significant.

No further significant correlation of self-esteem measures and enjoyment was observed.

Figure 5: Players’ enjoyment rating after playing a “Half Life 2” level (measured by questionnaire items).

Discussion

The study revealed some interesting results on the importance of self-esteem for video game enjoyment, although some methodological limitations demand a careful interpretation of findings. First, the results lend some support to applications of Bandura’s (1997) framework of self-efficacy to game enjoyment (e.g., Klimmt & Hartmann, 2006): Participants who perceived themselves as
expert video game players (and knew from the recruitment conversation that they would be invited to play a video game) displayed a higher implicit self-esteem than game novices before they actually played the game. This indicates an optimistic self-expectation and high self-efficacy conviction in respect to the task ahead. This higher self-efficacy might then serve as evaluation baseline for subsequent game performance. Novices, in contrast, were less self-convincing and displayed a lower implicit self-esteem before playing a game – a task they expected themselves not to perform very well. Consequently, their evaluation criterion for their own game performance was probably much lower than the self-set evaluation standard of expert players.

A consequence of this difference in the pre-gaming implicit self-esteem is that game difficulty did not affect novice players’ enjoyment. Novices found both the easy and the difficult game version equally enjoyable. Thus, they either reached their (low) self-defined performance criterion in most cases or they did not consider a good performance as a key determinant of their entertainment experience (and rather focused on other fun factors such as Presence or immersion, see Tamborini & Skalski, 2006).

In contrast to the novices, game difficulty mattered for game experts. They produced the highest observed mean enjoyment ratings if they had played the ‘challenging for experts’ level (which was objectively very difficult to master), but reported lower enjoyment if they had been exposed to the ‘easy for experts’ version (which was still not trivial to master compared to the ‘easy for novices’ version). This result indicates that a strong challenge (and the opportunity to test their skills) is important for experts’ game enjoyment, and that the absence of challenges is a critical game feature that undermines their entertainment experience. Experts were obviously not willing or able to compose their enjoyment out of other fun factors (see above), but heavily depended on the possibility to put their self-expectations to a test. The findings on implicit self-esteem contribute to the interpretation of these findings: If confronted with a challenging game level, expert players held up their (high) self-esteem they had revealed prior to playing the game, as they were able to test (and at least partly meet) their self-expectations. Experts’ self-esteem dropped substantially, however, if they had to play an easy game level with no chance to demonstrate their (self-ascribed) high skill and to test their self-expectations. This kind of disappointment about the game lowered their self-esteem and thus reduced game enjoyment.

In terms of measurement evaluation, findings of this experiment are highly informative, although they do not allow the definition of a standardized implicit method for self-esteem-based game enjoyment. First, the self-esteem IAT turned out to be useful as state-sensitive measure of implicit self-esteem. Expert-novice differences prior to playing and a tendency to an expertise x difficulty interaction in IAT findings after playing (with experts displaying lowered self-esteem after exposure to the easy game) are the empirical observations that allow to conclude some merit in the self-esteem IAT for game enjoyment measurement. However, the specific patterns that occurred indicate either some conceptual problems (as the true connection between game performance, game difficulty, player self-esteem and game enjoyment could be more complex than hypothesized in the FUGA theory papers, see deliverables to WP2). An additional experiment will be required to clarify the capacity of the self-esteem IAT in the context of video game enjoyment. For the moment, construct validity of the self-esteem IAT within a game enjoyment context cannot be definitely confirmed, but the findings suggest a greater likelihood that additional work can demonstrate a satisfying construct validity of the self-esteem IAT.

Second, the NLET seems to be at least as useful as the IAT, as the measure produced the same result pattern as the IAT. As it turned out to be sensitive to the treatment and easy to apply, NLET is definitely a candidate that will be useful in future studies on game enjoyment and player self-esteem.
Overall, the findings of both applied implicit measures suggest that they can augment and enrich results obtained from conventional explicit (questionnaire-based) measures. While there is some correspondence between implicit and explicit measures, there is also some divergence, which is in line with general findings from social psychology (Hofmann et al., 2005) and justifies combining implicit and explicit measures to gain a better picture of the cognitive processes involved in video game play. While there are some practical problems (especially strong error variance in the IAT, which results in low test power) attached to the implicit measures, they can obviously enrich other measurement paradigms especially due to their capacity to tap into self-related cognitive processes such as self-esteem. Future studies will have to (re)produce findings that help to better understand the abilities and limitations of implicit measures in a video game context and to find out if the unexpected patterns observed in the current study have to be attributed to conceptual issues (i.e., assumptions on player performance, game difficulty, self-esteem processes, and game enjoyment are wrong and need to be adjusted) and/or methodological issues (i.e., the self-esteem measures produce systematic, however invalid result patterns). Because the findings obtained in the present experiment are interpretable within a game enjoyment framework (and are not incomprehensive or chaotic), we are optimistic that future project work in FUGA will help to further establish the construct validity of the implicit methods in self-esteem-based studies.

2.2.4 2\textsuperscript{nd} Construct validity study of self-esteem

Introduction

The 2\textsuperscript{nd} construct validity study on self-esteem processes was based on the results of the 1\textsuperscript{st} study of self-esteem processes. In this study, both the IAT, the NLET turned out to be useful measurements of the state self-esteem. Some changes in the design were indicated, due to the following observations: Participants in the first study responded very differently to the stimulus dependent on their degree of playing-expertise. While experts in the difficult game condition and novices in both the easy and difficult game condition suffered only from a very little loss in self-esteem, experts in the easy game condition responded with a greater loss in self-esteem (the general phenomenon of a loss of self-esteem through playing was probably caused by a training-effect: Participants became especially faster for the incompatible trials and less faster for the compatible ones). The greatest loss on self-esteem for the expert players in the easy game condition can be explained by their high standards of playing outcome: They started with a relatively high level of self-esteem – due to their optimistic self-expectation and high self-efficacy conviction in respect to the task ahead and then did not have the chance to give proof of their abilities. This also affected game enjoyment negatively.

Apart from these difficulties, the uses measures IAT and NLET (see above) seemed to be useful and were thus applied for the existing study as well. As a consequence of our observations, we modified the design of the second study in the following three concerns:

- In the first study, we have seen that novice players’ enjoyment and self-esteem did in contrast to expert players – not depend much on game difficulty. This can be explained by the concept of contingencies of self-worth, which differ from person to person (Crocker & Woolfe, 2001). Contingencies of self-worth are domains or category of outcomes on which a person has staked his or her self-esteem. In this case, expert players – in comparison to novices – should be more likely to base their current state self-esteem on their playing outcome. We therefore assume – and actually have seen in the last study – that expert players are more affected in their state self-esteem by a playing experience than novice players are, who probably do not base their self-esteem on the playing outcome. We
implemented this observation into our research design and let only *experts* participate in the second construct validity study on self-esteem.

- In the preceding study we were not able to elicit a rise of self-esteem through success in the game, even though this was our main assumption. We propose the reason for this to be the training-effect for especially the incompatible IAT trials. Another reason could be that the ‘easy’ game level was just too easy and participants had to attribute their success externally, which undermined the hypothesized self-esteem-enjoyment connection. In order to avoid this for the follow-up study, we added another level and now had an easy level, an intermediate “Flow”-level and a difficult one. This diversification should provide us with more detailed results that can verify the construct validity of our measurements.

- Besides from the participants and difficulty of levels, we chose a different game for this study. Instead of Half-life, we used Unreal Tournament 3 (UT3). We chose a level in which players directly fight against another person (‘duel mode’). Each round ends with the player’s own death or the death of the opponent. Thus, UT3 gives a much more direct feedback of success than Half-Life 2 did in the first study. We assume that this direct performance feedback should lead to more pronounced group differences concerning self-esteem and game enjoyment.

**Design of Implicit Measurement of Self-Esteem**

As in the first pilot study of self-esteem, we again used the self-esteem IAT (Farnham, Greenwald & Banaji, 1999) and the The Name Letter Evaluation Task (NLET; cf. Kitayama & Karasawa, 1997; Koole, Dijksterhuis & Knippenberg, 2001; Nuttin, 1985; Nuttin, 1987). Both measurements are described in 2.2.3 in detail.

Besides the IAT and the NLET, a third implicit task was implemented in the second study which is not relative to the evaluation of an “other”-category: The Signature Task (Rudman, Dohn & Fairchild, 2007; Zweigenhaft, 1977). For this task, participants are asked to sign a consent form in advance to and after their participation. The space they needed for their signature was measured and pre- and post-size was compared. An increased size (in square centimeter) after playing compared that in advance to the playing was interpreted as an increase in self-esteem, while an decreased size was interpreted as an decrease in self-esteem. In their study on automatic threat defense, Rudman et al. (2007) found out that a threat to self-esteem leads to an increase in implicit, but not explicit self-esteem which was actually confirmed by the data: Signatures of participants suffering from a threat to self-esteem increased significantly in size, whereas those of participants whose self-esteem was not threatened, decreased.

**Explicit Self-Esteem Scales**

The third component was like in study one an explicit approach that serves purposes of implicit-explicit differentiation (see Hofmann, Gschwendner, Nosek & Schmitt, 2005) and cross-validation (with moderate correlations between implicit and explicit measures indicating validity of the implicit measures).

For explicit measurement, items regarding the feeling of competence and success and the relation of success to the difficulty of the game and own performance and capacity were developed resp. adopted from the first study on self-esteem.
Methods

Participants and Procedure

74 voluntary male university students aged between 18 and 32 years (M = 21.84, SD = 2.73) took part in the experiment. All participants play at least sometimes digital games and they all had some experience with the genre of the game used in the experiment, first-person-shooters (FPS). Before we invited the students to participate, we asked them to rate themselves on a 10-point-scale, if they were rather novices or experts regarding playing FPS (1 meaning being a novice with almost no experience, 10 meaning being an absolute expert). Only the ones that rated themselves at least “5” were asked to participate in the study. These “experts” were then randomly assigned to one of the stimulus levels “easy”, “medium-heavy” and “difficult”. In the easy level, it was almost impossible to get hurt or to die. The level with medium difficulty was supposed to provide the players with some success and the feeling of competence. At last, the difficult level was very hard and impossible to win. Players necessarily get killed in this level – independent from their level of skills. The following table 5 shows the mean self-rating of expertise for the three conditions.

Table 5: Self-rated playing expertise for the different experimental conditions; Ø = 6,86 (SD = 1,26)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>6,6</td>
<td>1,29</td>
</tr>
<tr>
<td>Medium</td>
<td>7,17</td>
<td>1,03</td>
</tr>
<tr>
<td>Difficult</td>
<td>6,96</td>
<td>1,36</td>
</tr>
</tbody>
</table>

The participants were again individually invited to a quiet room with controlled lighting conditions and were asked to sign a letter of content to participate in the following procedure. This signature was later used to compare it with the signature after playing. Participants were further asked to do a testing phase of the speeded tasks, which was the first self-esteem-IAT phase in advance to the game. After the training phase, participants played the level for 10 minutes. They were then requested to complete the self-esteem IAT again (procedure described in chapter 2.2.3). The IAT was followed by the questionnaire starting with the NLET (cf. section 2.2.3). Both the self-esteem IAT and the NLET were applied in the same way as in the first study, its logic is thus described above in detail (see 2.2.3).

Furthermore, the game-related self-esteem was assessed as questionnaire-based measure of the proximate and explicit self-esteem (items as “I am proud of my performance in the game” or “I am ashamed of having performed so badly in the game”). Finally, some further relevant dimensions were assessed in the questionnaire: the competence items of the GEQ-questionnaire, enjoyment experience during and after the game, the evaluation of the own performance, estimation of the difficulty of the game and items of the perfectionism scale (dimension “Striving for excellence”; Hill et al., 2004) were assessed on five-point-scales and mean indices were computed for further analyses (see table 6 below for relevant indices). Finally, participants signed another form for the signature task. After the procedure, participants were debriefed and dismissed. Each person received 5 EUR as compensation.
Table 6: Indices

<table>
<thead>
<tr>
<th></th>
<th>Cronbach’s Alpha</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competence, GEQ-items</td>
<td>.87</td>
<td>2.62</td>
<td>.99</td>
</tr>
<tr>
<td>Enjoyment during the game</td>
<td>.93</td>
<td>3.89</td>
<td>.82</td>
</tr>
<tr>
<td>Enjoyment after the game</td>
<td>.74</td>
<td>3.92</td>
<td>.69</td>
</tr>
<tr>
<td>Evaluation of the difficulty</td>
<td>.91</td>
<td>2.82</td>
<td>1.18</td>
</tr>
<tr>
<td>Appraisal of own performance</td>
<td>.8</td>
<td>3.17</td>
<td>.9</td>
</tr>
</tbody>
</table>

N = 74

Data cleansing

Three participants were excluded from data analyses due to technical problems during the experiment. Further all extremely high or low IAT-values (compatible and incompatible trials, prior to and after the playing) were excluded for data analyses (two standard deviations below/above the mean values). This rather strict approach was chosen very considerately in order to level the variance of reaction times over the three experimental groups. The same procedure was carried out for the NLET and the signature task: All values greater than twice the standard deviation of the mean value were excluded from the data set. Thus, analyses of the IAT-data was run with N = 60-65 (depending on the value-category), analyses of the NLET with N = 68 and the data set for the signature task contained N = 70.

Results

Results of the IAT

To detect changes in implicit self-esteem induced by the stimulus game, the key information of the self-esteem IAT, the so-called IAT effect, was computed for the pre-game and post-game tests. Strong positive IAT effects indicate a high self-esteem. The difference between pre-game and post-game IAT effect was used for analysis. Players confronted with different game task difficulties (easy versus medium versus challenging) were compared in respect to this difference. The last column of the following table 7 shows the rise of self-esteem (as assessed through less reaction time for compatible trials (“me” and positive attributes) and longer reaction time for incompatible trials (“me” and negative attributes).
Table 7: IAT-effect for the experimental conditions

<table>
<thead>
<tr>
<th></th>
<th>Pre-Game-IAT-Effect</th>
<th>Post-Game-IAT-Effect</th>
<th>Difference of Pre- and Posttest: rise of self-esteem through playing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy game version</td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
</tr>
<tr>
<td></td>
<td>N = 19</td>
<td>113,93 (86,78)</td>
<td>25,75 (124,16)</td>
</tr>
<tr>
<td>Medium game</td>
<td>Mean (SD)</td>
<td>161,97 (92,33)</td>
<td>55,95 (117,92)</td>
</tr>
<tr>
<td>Difficult</td>
<td>Mean (SD)</td>
<td>119,37 (100,18)</td>
<td>6,09 (137,34)</td>
</tr>
</tbody>
</table>

N = 60; Main effect time: F(57/2) = 32,83, p<0,001; η=0,37; Main effect game version and interactional effect n. s.

As can be seen in table one, none of the groups benefits from the game regarding self-esteem. The players of the easy level suffer the least from the game, though. The difference is not significant. A change in the IAT-effect can be due to getting slower or faster when assorting the compatible trials, the incompatible trials or both. In order to detect the “point of change”, we thus analyzed the data for the compatible and the incompatible trials separately. Figure 6 below demonstrate that major changes took place for the incompatible trials. These changes though, appeared in all three groups, indicating this result being a methodical artefact due to a training-effect. Incompatible trials means having to assort “me”-related words and negative attributes to the same side of the screen. This task is for most people very difficult in the beginning and therefore even holds the chance for training-effects. Compatible trials are very much easier to accomplish right from the beginning and thus they are less susceptible for training-effects.

Figure 6: Reaction times for incompatible and compatible IAT-trials in Milliseconds

Whereas all groups perform faster in the second IAT on the incompatible trials (see figure 6a), the reaction times on the compatible trials stay almost at the same level (see figure y). Neither a training-effect nor big differences between the experimental groups turned up for the compatible
tasks. For this reason, only differences in compatible trials (i.e., those sorting tasks requiring participants to associate “me” and positively valenced attributes as well as “not me” and negative valenced attributes, which is intuitively easier and faster to do than associating “me” and negatively valenced attributes and “not me” and positive valenced attributes) are considered for further analyses.

From the first study on self-esteem (see 2.2.3), we learned that changes in implicit self-esteem not only depended on the difficulty of the game level, but also on self-expectations and feelings of self-efficacy: Rather than only taking into account the difficulty of the game level, participants’ self-esteem was affected by the interplay of difficulty, self-expectations and actual performance. Although we only have expert players in the existing sample, we included the actual game-performance as second factor in the analysis of variance. For this purpose, the distance of the number of “kills minus deaths” for each player to the mean of the respective experimental group was calculated. These values were transformed into comparable z-standardized values. This way, a performance measure that was comparable across experimental conditions was created, as participants who outperformed other participants in their experimental group received higher values. In order to introduce this performance-variable into an ANOVA, a median-split was conducted. We thus had – besides three different game versions – also two groups based on their factual performance: Good vs. bad players. Figure 7 below illustrate the results.

As assumed, performance on the game does indeed affect self-esteem. Good players start with a higher self-esteem than bad players did (F(57/2) = 9.3, p<0.01; η=0.14). This pattern could already be observed in the first study of self-esteem (see 2.2.3). Also interesting – though not significant – are the differences regarding the development of reaction times before and after the game, thus the interaction effect of time and performance factors: Bad performers are only affected from the easy game version, meaning that if the players play an easy game version and are then not even able to perform well, self-esteem is negatively affected. The opposite happens to good performers, they benefit from an easy game version, probably due to many positive feedbacks (which would be a lot of kills and few deaths in this case). The difficult game version does not affect their self-esteem much, whereas they suffer the most from the medium game version. The differences between the game versions are neither very high nor significant and should thus be interpreted carefully.
Results of the NLET

A repeated measure ANOVA with the letter category (initial vs. other) as within subjects’ factor and the experimental condition (game version) as between subjects’ factor was conducted to test the effect of the difficulty on the dependent measures. Only a main effect for the letter category was observed – which is the basic assumption of the NLET ($F(60/1) = 147.88; p<.001; \eta^2 = .71$). Furthermore, neither a main effect for the experimental condition nor an interaction effect the interplay of condition and letter category was observed. No remarkable results could be observed, too, after introducing the players’ performance into the ANOVA as second factor.

Results of the Signature Task

In order to detect differences between groups of the size of signatures before and after playing, a single value displaying this difference was computed. To standardize signature size data, the ratio between the post-game / pre-game difference and the absolute size of the pre-game signature was computed. This way, a percentage value of signature size change was obtained. An ANOVA was computed. Supporting our assumptions, the signature of the ones playing the medium game version became significantly larger – compared to the one prior to the game – than the signatures of the other participants ($F(69/1) = 8.51; p<.001; \eta^2 = .2$). This pattern appears both for the good and the bad performers, though more strongly for the good players (see figure 8), meaning that the difficulty of the game had a greater impact on those who performed well.

![Figure 8: difference of pre and post signature size depending on game version and performance](image)

Results of the explicit measurements

As already described above, several dimensions of playing experience were assessed through explicit, questionnaire-based measures. In order to test the influence of the stimulus games on five of these central dimensions, five analyses of variance were computed. As the playing performance turned out to be an important mediating and moderating variable, it was introduced into the ANOVAs as second factor. Table 8 shows the results. In the first place, results demonstrate that our stimuli worked very well. The ones in the easy group evaluate the game as much easier than the others, felt more competent and evaluated their own performance better. Furthermore, the link between performance and game enjoyment – both during and after the game – is confirmed: The easier the game and at the same time the better
the performance, the more players enjoy playing. Furthermore, the importance of own performance is underlined. There is a significant difference for all dimensions between the ones with relatively less own deaths and many kills (only for the evaluation of difficulty, the difficult game version seems to be so hard that it cannot produce any variance).

Table 8: Mean differences of explicit measures for different game versions and good vs. bad performers

<table>
<thead>
<tr>
<th></th>
<th>Easy game version</th>
<th>Medium game version</th>
<th>Difficult game version</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Good Performers</td>
<td>Bad Performers</td>
<td>Good Performers</td>
</tr>
<tr>
<td></td>
<td>(N=13)</td>
<td>(N=12)</td>
<td>(N=14)</td>
</tr>
<tr>
<td>Enjoyment * during the game</td>
<td>4.5</td>
<td>0.67</td>
<td>4.06</td>
</tr>
<tr>
<td>Enjoyment ** after the game</td>
<td>4.5</td>
<td>0.38</td>
<td>4.17</td>
</tr>
<tr>
<td>Feeling of competence (GEQ)**</td>
<td>3.78</td>
<td>0.57</td>
<td>3.38</td>
</tr>
<tr>
<td>Evaluation of the difficulty****</td>
<td>1.35</td>
<td>0.47</td>
<td>1.83</td>
</tr>
<tr>
<td>Appraisal of own performance*****</td>
<td>4.22</td>
<td>0.63</td>
<td>3.88</td>
</tr>
</tbody>
</table>

N = 71
*Main effect game version: F(65/2) =5.89, p<0.01; η=0.15;  
Main effect performance: F(65/1) = 5.32, p<0.05; η=0.08;  
** Main effect game version: F(65/2) =11.28, p<0.001; η=0.26;  
Main effect performance: F(65/1) = 6.52, p<0.01; η=0.09;  
*** Main effect game version: F(65/2) =53.05, p<0.001; η=0.62;  
Main effect performance: F(65/1) = 11.19, p<0.001; η=0.15;  
**** Main effect game version: F(65/2) =68.47, p<0.001; η=0.68;  
Main effect performance: F(65/1) = 3.49, p<0.07; η=0.05;  
***** Main effect game version: F(65/2) =33.58, p<0.001; η=0.51;  
Main effect performance: F(65/1) = 7.88, p<0.01; η=0.05.

Correlation of Performance and Enjoyment

In order to further test the relation between enjoyment and performance, correlations were computed. Both performance criteria – kills and own deaths – correlate highly significant with enjoyment during the game (kills: r = .49; p<.001; deaths: r = -.42; p<.001) and enjoyment after the game (kills: r = .55; p<.001; deaths: r = -.56; p<.001).

These correlations are even more pronounced for the good performers. Performance has thus a greater effect on enjoyment for those, who perform already well – which are probably the more passionate players.

Discussion

The second study on self-esteem also revealed some interesting results on the importance of performance and positive feedback for video game enjoyment. Like in the first study, we face some methodical limitations, too, that demand a careful interpretation of findings. First, the results lend support to the basic assumption of our theoretical framework (cf. deliverable 3.1): A good
performance and being able to cope with the given task leads to experience of self-efficacy and game enjoyment. The results of the questionnaire-based assessment of these dimensions strongly support the hypothesized relations: First, game enjoyment was higher in players who performed better (table 8); this pattern occurred in each of the three difficulty settings. Second, questionnaire data suggest interesting attribution processes that help players to maximize their performance-based game enjoyment. In the easy game version, the high number of successful rounds produced high levels of enjoyment and a high satisfaction with one’s own performance. In this case, a self-serving bias towards internal attribution of success seems to be operating: In spite of the obviously easy task, players attribute their positive results to their own competence (which is nicely documented by the GEQ scale “competence”, see table 8). In the hardest game version, enjoyment is quite close to the medium condition, and satisfaction with one’s own performance is even higher in this group than in the medium difficulty condition. This seems to result from a self-serving external attribution bias: Players use the argument of the extreme task difficulty to justify the negative results they achieved and protect their self-perception as a game player from severe damage. However, the evident high number of lost rounds does not allow to create a real positive overall gaming experience; it only prevents game enjoyment (and satisfaction with the own performance) to drop down massively. Bad performers in the medium condition, finally, seem to be in the problematic situation of not having achieved very many successful rounds and not being able to attribute this limited success to external factors, because (they found) task difficulty was not very high. So their satisfaction with their own performance is the lowest of all groups, although the medium number of successful rounds still enabled a medium level of game enjoyment. These findings are interesting in light of existing game enjoyment models (e.g., Klimmt, 2003; Behr, Klimmt & Vorderer, 2008; Vorderer, Hartmann & Klimmt, 2006) and suggest their refinement by self-serving attribution processes. They also allow to draw practical conclusions on game design (see below).

The implicit measurements, however, have not turned out to be a sufficiently valid and sensitive approach to performance and self-esteem dimensions in video game enjoyment. Regarding the self-esteem IAT, we observed a strong training effect, especially for the incompatible trials, which disqualifies the application of a self-esteem IAT as a pre/post-measurement. Ironically, while repeated measurement does obviously not work very well with the self-esteem IAT in a lab setting (the highly undesirable training effect occurred in both studies), the application of the self-esteem IAT before the actual gaming phase revealed an interesting pattern that is interpreted in light of self-efficacy theory (Bandura, 1997; Klimmt & Hartmann, 2006). Expert players and/or participants convinced of their skill to master the upcoming game situation – in other words, people with a high state self-efficacy expectation – display a stronger implicit self-esteem before playing the game. This finding is interesting, as it indicates automatic self-esteem processes involved in mental action preparation. Especially expert players construe a game situation as a performance setting (Behr et al., 2008; Vorderer et al., 2006), and they do so automatically. Past positive competence experiences raises implicit (and possibly explicit) state self-esteem, with potential consequences for the game experience to follow.

It is much more difficult to interpret the measured changes in implicit self-esteem that (may have) occurred after participants had played the game. The training effect suppresses much variation that might have been observed if it did not occur – especially since it is plausible to assume bad performance lead to lowered state self-esteem and affect response latencies precisely to the incompatible trials, as they are more difficult to handle than the compatible trials. However, the general training effect produces exactly this pattern and strongly reduces our capacity to distinguish participants with de facto lowered self-esteem from participants with stable or increased implicit self-esteem. Looking at the compatible trials only is not the optimal solution to this problem, although we suggest to understand at least an increase in response latencies to compatible trials as a
true loss of implicit self-esteem that cannot be created by the training effect: If players need more time for compatible tasks (i.e., sorting tasks linking “me” to positively valenced attributes) after the game than they needed before the game, this is plausibly the result of a damage in self-worth, potentially caused by the game experience and the player’s game performance.

With these restrictions, it is possible to derive some conclusions from IAT data. The loss of implicit self-esteem among underperforming players of the easy game level can be explained again through attribution processes: Just like most players in the medium difficulty condition, they seem to have found their own performance below what they believed was possible for them given the difficulty level of the game. Furthermore, the difficulty is not too high, which allows for an internal attribution of performance rather than “blaming” the results on the game. These findings thus indicate automatic attribution processes being involved in the game performance – self-esteem relationship. However, the evidence on which this interpretation is based is weak, as the underlying IAT result patterns are non-significant, and some other patterns are less well-explainable.

Overall, the self-esteem IAT is not useful to cover the self-esteem processes involved, as the training effect obstructs a valid measurement repetition (see above); furthermore, even high case numbers cannot prevent single extreme values from affecting the result pattern. Data cleansing (i.e. removing single cases with extreme IAT values) in the present case changed the result pattern of the IAT several times in fundamental ways.

In sum, the main problem observed in the two experiments on video game performance, implicit self-esteem, and game enjoyment is the theoretical complexity of the research object (the multiple and dynamic links between players’ self-efficacy, game difficulty, actual playing outcome, attribution processes, de-facto changes in implicit self-esteem, and self-rating of player performance) coming together with a suboptimal sensitive, unstable, and artefact-susceptible repeated implicit measure. In fact the theoretical relationship between player performance, self-esteem, and game enjoyment seems to be so complex that any individual scientific method would underperform in revealing it to a sufficient degree. In the case of implicit measures, the problem of automatic versus deliberative cognition and potential divergences between these two domains of thinking further impede the interpretation of validity information (Hofmann et al., 2005): If implicit responses to game performance differ from explicit reactions (i.e., implicit self-esteem may remain the same whereas explicit self-esteem goes down), the conceptual complexity involved in the player performance – game enjoyment relationship is doubled, as it might have to be considered separately for automatic and for reflected self-esteem cognitions.

Reduction of complexity through experimental design was thus the strategy applied in both experimental studies; however, since players’ actual performance and mental response to playing outcomes cannot be manipulated or held constant, research on performance/self-esteem and game enjoyment cannot be controlled as rigidly as other research issues in the social sciences. Because the IAT does not work effectively when applied as repeated measure, the methodological alternative of within-subject designs (i.e., let the same people go through different game conditions and see if the IAT finds predictable changes in self-esteem; cf. Rheinberg & Vollmeyer, 2001, for according experiments on flow experiences during game play) is also likely to fail.

Similar problems apply to NLET and the signature task. Result patterns are only partially conclusive (and more so in the self-esteem study I than in study II). Because the signature task was completed only after all other implicit and explicit measures of the post-playing phase of the experiment were done, its findings are potentially compromised by ex-post cognitions such as automatic self-esteem threat defence (for the detection of which is was originally used: see Rudman, Dohn, & Fairchild, 2007).
While both experiments on player performance and self-esteem help to advance the scientific understanding of video game enjoyment (see above), the main goal of establishing a construct valid implicit measurement of performance-dependent self-esteem in video game players was not achieved.

### 2.2.5 General conclusion / outlook

With the completion of WP5, HMTH has achieved significant progress in measuring specific elements of video game enjoyment with implicit methods. The time and efforts that had to be invested were larger than originally planned; however, the results of several smaller pilot studies with more than 100 participants and four major construct validity experiments with overall 246 participants (in which overall seven different implicit methods were tested) represent a substantial contribution to both the theoretical and methodological state of the art in game enjoyment research.

Concerning the research line on identification with game characters, the developed set of methods is essentially demonstrated to be valid and useful. Identification with game characters can be measures effectively by a mix of Lexical Decision Task, Implicit Association test, and Self-Description task. Experimental construct validation was successful for all three methods (with notable limitations concerning LDT and especially SDT); the relationship with overall or general game enjoyment is in need of further investigation, however. Within the major context of the FUGA project, HMTH can thus offer a set of working and useful implicit measures for a key component of game enjoyment that is highly relevant to various game genres, including role playing games, first person shooters, sports games, and adventure games. These measures hold the important advantage to be customizable to the role a specific game intends users to identify with. LDT, IAT, and SDT all allow to use concepts and attributes for measurement that are proximate to the given identification target, be it a soldier, racing car driver, magician or chief executive officer. From a construct validity perspective, then, identification measures have been advanced that follow FUGA’s criteria and scope, and scientific as well as applied dissemination of results is already in progress. The next logical step for this line of research is thus to proceed within the original FUGA work plan and test the stability and reliability of the three mentioned implicit measures of identification. A replication of the second experiment of identification, which applies all three measures within one research design, is thus intended for the upcoming WP6. This replication will test some additions and modifications to the LDT and SDT (essentially introduce more role attributes) and also help to clarify the remaining questions concerning the relationship between implicit identification processes and (explicitly measured) overall game enjoyment.

Concerning the research line on player performance, (implicit) self-esteem, and game enjoyment, the general conclusion of HMTH’s studies on piloting and construct validity is less positive. Because two rigid experimental approaches to test predictions on the connection between player performance, game difficulty, implicit self-esteem, and game enjoyment did at best reveal a few useful results, but did not allow to establish a solid implicit measurement paradigm, the construct validity studies on implicit self-esteem of video game players did not achieve a scientifically satisfying outcome. In contrast to the implicit measurement paradigms for the identification construct (see above), it is thus not indicated to proceed to WP6: There would not be much merit in testing the stability and reliability of those measures that did not turn out as sufficiently valid and sensitive in WP5.

On the other hand, the construct validity studies have confirmed the theoretically assumed importance of player performance (and game difficulty) for game enjoyment. Therefore, it would reduce the theoretical completeness of the FUGA approach if any further work on the self-esteem
dimension of video game enjoyment would be dropped. Instead of replicating the WP5 experiments to test measure stabilities, then, an alternative research strategy for WP6 is proposed.

One commonality of the two experiments conducted to assess implicit self-esteem measures’ usefulness in the game enjoyment context is a focus on short-term changes in implicit self-esteem. Participants played a video game for 10 minutes. While it is reasonable to assume de facto changes of implicit self-esteem due to the outcomes of a ten minute game performance situation (that took place in a research laboratory and thus was likely to be rendered as a social situation of “showing what I can do”), it is not clear if the applied implicit measures (IAT, NLET, Signature task) are capable to detect such short-term, potentially minor changes of implicit self-esteem.

HMTH thus proposes to begin WP6 work on implicit self-esteem measures by shifting the focus towards mid-term changes in implicit self-esteem. A strong sense of achievement that is truly relevant for players’ (implicit) self-esteem is more likely to result from longer periods of play in which large-scale tasks (accomplishment of a major story segment, for instance) are successfully resolved. Such a perspective would change the quality of the performance environment in which implicit self-esteem (may be or) is affected: Instead of a short-term goal (get to point X or win as many duels as possible), players would be required to win a huge battle or save the world, efforts that demand potentially hours of time and much cognitive energy to invest. Such a major progress as task (that comprises many smaller tasks the performance in which is irrelevant then, as long as players manage to get on to the next small task) would thus be more likely to affect implicit self-esteem in measurable (and more predictable) quantities. If the usefulness of implicit self-esteem measures can be established in such a mid-term focused research design, the practical implications would be that implicit self-esteem measures may help to evaluate a game prototype at a more general level (instead of small-scale level of individual fights or puzzles).

WP6 Studies on Self-Esteem should thus employ a design in which players are given a large-scale task (e.g., infiltrate an enemy base and destroy a communication facility). One group of players would be interrupted as soon as they reach a defined point in the game level (which is about half the way towards the defined major goal) and then confronted with the implicit self-esteem measures, primarily the self-esteem IAT. The other group is allowed to play as long as they need to resolve the major task and is then confronted with the implicit measures. Technically, the experimental factor varied in this design is ‘success versus time-out/interruption/non-success’). If the self-esteem IAT would produce differences between these groups, a practically relevant measurement pattern would occur (i.e., does a game prototype provide milestone success experiences relevant to player self-esteem?). In case construct validity of implicit measures for such mid-term changes in player self-esteem could be established, such a study could certainly be replicated within WP6, in order to test the measures’ stability, which would be the defined work goal of WP6.

The final conclusion of HMTHs work within FUGA so far is thus very positive. Both lines of research pursued (identification and self-esteem) are on the right track – from a theoretical perspective. Empirical results are only partly compatible with our assumptions and goals, but this is the very nature of empirical research. While the further work that needs to be done has been outlined above, we are also satisfied with the contribution to innovation in game enjoyment theory and measurement that has been achieved by our efforts to explore and define implicit measurement paradigms.
2.3 UKA

The objective was to validate different measures based on fMRI brain activity recordings during game playing and think-aloud assessment as measures of Game Experience and investigate how they reflect the different dimensions of Game Experience.

Game enjoyment is a complex and multidimensional construct and measuring its biological correlates is a challenge. In order to approach this task we needed a well grounded psychological theory. We applied the Flow theory (Csikszentmihalyi, 1975) that describes flow as a mental state of being completely absorbed by an activity, associated with positive feelings. Csikszentmihalyi describes flow as being accompanied by several factors that can be measured, namely:

1. Balance between ability and challenge
2. Concentrating and focusing
3. Direct feedback of game success
4. Control over the situation/activity
5. Clear goals
6. Activity is intrinsically rewarding
7. Loss of self-consciousness
8. Distorted sense of time
9. Action awareness merging (awareness is only focused on the activity).

The first five factors can be operationalized behaviorally and measured during the game, which allows us to directly measure neurobiological correlates of selected aspects of flow. For the latter four, we have no behavioral measures but we will examine neurobiological correlates. We predict that as the outcome of the game events leading to flow state in participants, we will find the activation of brain areas involved in reward, self-consciousness, sense of time, and action awareness.

2.3.1 fMRI of digital gaming and Think-Aloud

Introduction

The central objective of this study is the measurement of game experience and game enjoyment during video game play. Game enjoyment was assessed directly via Think Aloud. Our intention was to relate reported game enjoyment to objectively measurable content factors. As basis for the development of a coding scheme we drew on flow theory (Csikszentmihalyi, 1975). In simple terms, the theory describes flow as a mental state of being completely absorbed by an activity, associated with positive feelings. Csikszentmihalyi describes flow as being accompanied by several potential factors, none of them being obligatory:

1. Balance between ability and challenge
2. Concentrating and focusing
3. Direct feedback of game success
4. Control over the situation/activity
5. Clear goals
6. Activity is intrinsically rewarding
7. Loss of self-consciousness
8. Distorted sense of time
9. Merging action awareness (awareness is only focused on the activity).

We consider flow experiences being very closely related to the concept of game enjoyment (cf. Sherry, 2004). We therefore developed a coding system extracting five dimensions from the game content that might be closely related to game enjoyment since they are based on core elements of flow theory. To cover positive and negative valence aspects of game experience we postulate two opposite poles of the game enjoyment dimension, the positive one being called “game enjoyment” and the negative one being called “disenjoyment”. We assume that game enjoyment is characterized by the presence of Flow factors, whereas disenjoyment is marked by its lack. Since not all factors can be assessed from the game content, we developed a content analysis extracting the factors

1. Balance between ability and challenge (Do the player’s skills match the demands of the situation?)
2. Concentrating and focusing (How concentrated is the player?)
3. Direct feedback of game success (Does the player see his success immediately and clearly?)
4. Control over the activity (How good is the player’s ability to control his game play?)
5. Clear goals (Does the player know what to do next?)

The operationalization of the content factors was based on expert ratings. Precisely defined content descriptions ensured high inter-coder reliability.

The remaining four constructs (6.-9.) are rather internal states which do not readily allow for a behavioural analysis. Since they cannot be inferred from the observed data, these states should be assessed based on the neurophysiological correlates. However, they can be characterized in terms of expected networks that contribute the states.

Game play recordings were content analyzed with a frame-by-frame method which assessed the content factors during game play. Three independent raters (2 female and one male undergraduate students at the RWTH Aachen University Hospital) and one supervisor (M.K.) rated each video frame all factors on a 3-point scale (low/medium/high) based on the recorded game play of each individual. The coders received about 16 hr coding training, in which they discussed the different categories with experienced digital game players and learned to rate events and violent interactions according to the coding scheme. The trainings were based on one separately recorded video not used in the study. The coders coded changes of the category values with a precision of 100 ms. Inconsistent ratings were discussed with the supervisor and corrected according to the coding scheme.

Since the main objectives of the game are killing opponents and avoiding being killed, we postulate that two major factors determining enjoyment and disenjoyment are game success (killing an opponent) and game failure (being killed). Furthermore, we assume that some additional factors influence perceived enjoyment and disenjoyment. Game difficulty should change the attribution of success in the respective situation. Higher difficulty should lead to more game satisfaction when having absolved a successful game play. Mastering a difficult situation boosts the self-esteem and judgement of the own abilities as well as of the perceived own effort (Kun & Weiner, 1973). This success should be especially attributed to good abilities in persons with a high need for achievement.
in the respective task (Heckhausen, 1988), as it is the case in regular players. On the other hand, lack of success (“failure” in game play) should lead to greater disenjoyment when the situation is easy, especially when the game success is relevant for the self concept, as it is in regular FPS players. In this case, the failure should be attributed to an internal factor such as lack of effort (Meyer, 1973).

Dispositional logic and empirical research suggest that high risk increases arousal and excitation (Anderson & Brown, 1984). Following the postulates of excitation transfer theory (Zillman, 1971, 1984, 1988) and its experimental support (e.g., Zillman, Johnson, & Day, 1974), we assume that this excitation is transferred to both success and failure in game play and should therefore enhance game enjoyment in successful situations and cause more disenjoyment when the player does not meet the demands of a situation.

Like risky situations, potentially beneficial situations are known to cause increased arousal and excitation (Ladouceur et al., 2003). Following again excitation transfer theory (Zillman, 1971, 1984, 1988), we also expect that potentially beneficial situations should enhance enjoyment in successful and disenjoyment in non-successful situations.

The generation of goal-directed behaviours is based on the extraction of reward-related information from multimodal inputs and then integration into beneficial behaviours (Hollerman et al., 2003). Reward expectation and reward prediction errors, that is the difference between expected outcome (e.g. reward) and the received outcome, are thought to be critical for dynamic adjustment in decision-making and reward-seeking behaviour (Cohen, 2007). Reward prediction errors are encoded in structures including midbrain dopamine regions, the cingulate cortex and the anterior and ventral striatum (O’Doherty et al., 2004). In particular, phasic increases in activity are observed when reinforcers are better than expected (a positive prediction error), and phasic decreases in activity are observed when reinforcements are worse than expected or not given (a negative prediction error; Schultz et al., 1997, 2004). In contrasts, neural representation of expected rewards is thought to be housed in the orbitofrontal cortex and amygdala (Hikosaka and Watanabe, 2000).

**Methods**

**Subjects**

Seventeen male German volunteers (age 18-33; mean: 24.4) were recruited with ads posted at the local university, via email, and in online newsgroups. Inclusion criteria were: age between 18 and 35 years, playing at least 5 hours weekly of First Person Shooter digital games, right-handedness, and German as native language. One subject had to abandon the experiment because of claustrophobia. No individual had any contraindication against MR investigations, acute or anamnesis of major neurological, psychiatric, or ophthalmologic disorders.

**FPS game and default settings**

The FPSG “Counter Strike: Source” (CSS, Valve Corporation, Bellevue, Washington; 2004) was used as the experimental game. In this game, two groups fight against each other as terrorists or special warrior action team (SWAT). The game provides different location settings (maps), e.g. inside and around an industrial building, in a train station, or in a castle. The plot and missions of the game are typical for the FPSG genre: Hostages have to be held by terrorists or rescued by the SWAT; terrorists have placed bombs whereas the SWAT has to stop the bomb countdown etc. The virtual environment contains various violent interactions that players experience from a first-person
perspective, that is, a virtual weapon is clearly visible and the player has the visual impression of holding the weapon in his own hands.

The game is played in several rounds per map. A round is over if one of the teams has fulfilled its mission or if a pre-determined time limit is reached. If a player is (virtually) killed before the round is over, he can observe the further game play from a viewer perspective (so-called dead mode), but has no opportunity to interact with other players until the next round.

For the purpose of this study we selected the single player mode in which all game characters except the player’s avatar are generated by the computer. This setting allowed for the control that a players’ performance was not influenced by other human players’ skills or actions. A set of three different maps was chosen (de.cbbble, de.inferno, de.piranesi), based on maximum colour contrast and better visibility in the MR compatible video goggles. All maps were defusion scenarios, i.e., terrorist have to place bombs while the SWAT members try to prevent the explosion. Players could choose whether they wanted to be terrorists or SWAT members. Every map was played for 12 minutes and every round within a map was limited to 3 minutes.

To reduce usability problems arising from different familiarity of the participants with the used FPS game and the MR compatible control device, all subjects were given time to practice the game before the game sessions until they felt comfortable with the game mechanic. Additional 10 minutes of training were performed inside the scanner before the functional measurements.

**Think Aloud**

During the scanning session, the Think Aloud method (TA, van Someren, Barnard, & Sandberg, 1994) was applied; TA is an introspective technique to assess thoughts and feelings. The sessions took place directly after the game play inside the scanner. Participants focused on the difficulty of the game (How hard/easy was the game play?) and the aspect of subjective game enjoyment (How much did you enjoy playing the respective scene?). At the end of each gaming block, the subjects watched a recording of their recently played sequence. They were instructed to verbally report continuously their estimation of difficulty and the amount of enjoyment they had during playing the respective scenes.

In order to practice expression their thoughts verbally, subjects performed a 12 minutes test block outside the scanner. It consisted of a six minutes playing session followed by a six minutes Think Aloud phase on the recently played sequence. The experimenter ensured that the subject had understood the instructions and was able to complete the task.

Evaluation of the Think Aloud was done by two independent raters (female undergraduate students at the RWTH Aachen University Hospital) and one supervisor (M.K.). To cover positive and negative valence aspects of game experience we postulate two aspects of the game enjoyment, the positive one being called “game enjoyment” and the negative one we termed disenjoyment. The coders assigned all verbal utterances explicitly reporting game enjoyment or disenjoyment to their respective categories. Difficulty ratings were assigned to the categories “easy,” “medium,” and “hard.” For this purpose, each coder received about 5 hrs of intensive training on an additional Think Aloud session not used in the study. The codes were referred to the respective time points in the game recordings and fMRI data.

With respect to the amount of different information important for the assessment of game enjoyment, the Think Aloud had been completed with an additional rating. Another rating session took place in an adjacent room after the MR scan. It used the same game recordings as the Think Aloud, but this time the relevant dimensions were risk and benefit of the player’s actions. The subjects did not talk but used two input devices (magnetic sliders) with buttons ranging from 1 to 4
(1 = not at all; 4= very much), one for risk and one for benefit. Subjects’ responses were recorded with a harddisk video camera for further analysis.

**Experimental paradigm**

After informed consent, the participants received the bundle of trait questionnaires. These could be filled out at home and were brought to the examination. After arriving at the MR scanner room, the subjects first completed the standard MR safety questionnaires and consent forms. The experimenter ensured that the participants had no contraindications against MR investigations and did not have any metal parts with them. Then the participants were made familiar with the experimental paradigm and shown the experimental setup. The participants filled out eventually remaining trait questionnaires and all the pre-measurement state questionnaires. All subjects had time to practice the game with the MR compatible controller (trackball and five mouse buttons) until they felt comfortable (5-20 mins). Afterwards they were introduced into the Think Aloud method (TA, see above). The experimenter explained the technique and gave examples for possible TA comments. The participants then played another six minutes session which was recorded with the commercially available frame grabber software (Fraps, Beepa® P/L). The session was followed by six minutes of Think Aloud training on the recently played sequence; the subject received further instructions of the experimenter if necessary. In two cases the training session was repeated to make sure that the subjects were able to complete the task in an adequate way. Before continuing, the experimenter ensured a good TA performance of the participants. Afterwards the participants were lead into the MR scanner.

Before the functional measurements the participants had time to play the game inside the MR scanner for another 10 minutes. The visual and auditory signals of the FPS game were delivered by MR compatible headphones and video goggles (Resonance Technology, Inc., Los Angeles, CA). The subjects then played three 12 minute sessions, each followed by a 12 minute Think Aloud phase on the recently played scene. The speech was recorded with an MR compatible optical microphone built as a prototype from the company Sennheiser Electronics (Wedemark, Germany). The signal was filtered online to minimize the EPI noise; so the experimenter could listen to the speech and deliver visual prompts if necessary (“Please speak up!”, “Keep on talking please!”, “Please focus on difficulty and on how much you enjoyed what you were playing!”). The experimenter registered all commented events but also non-commented game events if they appeared relevant for the challenge and enjoyment of the game. Visual and auditory output of the videogame was captured by the Fraps software (Beepa® P/L). The MR measurement ended with an 11 minute anatomical scan. In total 42 playing sessions (12 min each), 40 Think Aloud-sessions, and risk-benefit ratings for 34 game sessions were obtained.

After the fMRI data acquisition post-measurement, subjects were debriefed and filled out state questionnaires. Finally another TA session with direct verbal interviewer intervention and risk-benefit rating was performed. The latter data will be analysed and used to test construct validity after completion of the main analysis.

**Imaging Parameters**

fMRI was conducted at 3 T (Magnetom TRIO, Siemens, Erlangen, Germany) by means of single-echo single-shot echo-planar imaging (EPI; flip angle: 77 degrees; TE = 24 ms; 64 × 64 matrix with 3.6×3.6×5 mm² resolution; 4 mm slice thickness plus 1 mm gap). Twenty-eight oblique-transverse slices obtained whole brain coverage with repetition time TR = 2.0 s (360 volumes per session). Interleaved slice acquisition allowed for spatiotemporal oversampling reconstruction of the volumes, i.e., volume 1 = odd and even slices from acquisition 1, volume 2 = even slices from...
acquisition 1 and odd slices from acquisition 2, volume 3 = odd and even slices from acquisition 2, etc.

After the functional scans we conducted anatomical scans (standard protocol; inversion time TI = 500 ms; TE = 3.6 ms; flip angle = 8; bandwidth: 15.63 kHz; FoV: 240x240 mm²; slice 1.5 mm; 3d partitions: 124) that we used for obtaining individual references and as incentive for research participants.

Results

Based on flow theoretical concepts (Csikszentmihalyi, 1975), we developed a content analytical coding scheme in an iterative, deductive-inductive process. After extensive tests of the coding system’s reliability, five dimensions could be behaviorally coded as content dimensions and validated by the neurophysiological data. For an overview see Table 1.

<table>
<thead>
<tr>
<th>ROI</th>
<th>cluster</th>
<th>peak</th>
<th>MNI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p(corr.)</td>
<td>size(k)</td>
<td>p (corr.)</td>
</tr>
<tr>
<td>activation to success</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>superior vermis</td>
<td>0</td>
<td>755</td>
<td>9,12</td>
</tr>
<tr>
<td>inferior vermis</td>
<td>0</td>
<td>136</td>
<td>7,15</td>
</tr>
<tr>
<td>inhibition to success</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rACC</td>
<td>0</td>
<td>321</td>
<td>7,44</td>
</tr>
<tr>
<td>IPS R</td>
<td>0,007</td>
<td>37</td>
<td>6,44</td>
</tr>
<tr>
<td>activation to failure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>visual dorsal pathway</td>
<td>0</td>
<td>1233</td>
<td>11,38</td>
</tr>
<tr>
<td>V1 R</td>
<td>0,001</td>
<td>108</td>
<td>8,86</td>
</tr>
<tr>
<td>neocerebellar cortex L</td>
<td>0</td>
<td>392</td>
<td>8,6</td>
</tr>
<tr>
<td>inhibition to failure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nucleus caudatus</td>
<td>0</td>
<td>8031</td>
<td>14,05</td>
</tr>
<tr>
<td>IPS R</td>
<td>0</td>
<td>405</td>
<td>8,49</td>
</tr>
<tr>
<td>premotor R</td>
<td>0</td>
<td>678</td>
<td>7,27</td>
</tr>
<tr>
<td>premotor L</td>
<td>0</td>
<td>328</td>
<td>7,01</td>
</tr>
<tr>
<td>IPS L</td>
<td>0,001</td>
<td>126</td>
<td>6,82</td>
</tr>
<tr>
<td>activation to focus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neocortical cerebellum L</td>
<td>0,002</td>
<td>25</td>
<td>11,56</td>
</tr>
<tr>
<td>Inferior vermis</td>
<td>0</td>
<td>269</td>
<td>10,51</td>
</tr>
<tr>
<td>V1</td>
<td>0</td>
<td>50</td>
<td>7,92</td>
</tr>
<tr>
<td>inhibition to focus</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
First we will report approaches to these aspects in separate subsections.

1. **Balance between ability and challenge**

   **Content definition:** The game structure of the FPS reveals one event of clear success and one for failure. These are killing or being killed, respectively. The rater can readily identify either of these events. The time points of the events were noted with 100 ms time resolution.

   **Neurophysiological modeling:** We assume that success and failure elicited a momentary neuronal response in circumspect networks. The BOLD response was modeled by a generic hemodynamic response to these events and achieved statistical maps thresholded at a voxel-wise corrected $p < 0.05$ (Figure 1).
Figure 1: Statistical maps on activation and deactivation during success and failure events (threshold according to p < 0.05 corrected). The success and failure events were associated with increased visual or cerebellar activity. Success led to inhibition of the rACC. The reward system was only involved by showing caudatus inhibition in response to failure.

<table>
<thead>
<tr>
<th>Activation during success</th>
<th>Deactivation during success</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Activation during success" /></td>
<td><img src="image2" alt="Deactivation during success" /></td>
</tr>
<tr>
<td>Positive of Failure</td>
<td>Negative of Failure</td>
</tr>
<tr>
<td><img src="image3" alt="Positive of Failure" /></td>
<td><img src="image4" alt="Negative of Failure" /></td>
</tr>
</tbody>
</table>

1 All coordinates are in the MNI system. In coronal and transversal slices, the right hemisphere is displayed on the right.
**Neuronal network:** Superior vermis (MNI = 4.0, -72.0, -10.0) activates and rACC (-8.0 38.0 12.0) deactivates during successful event. Failure was correlated positively with increased activity in higher visual areas (cuneus, 0.0 -84.0 26.0 and LOC, 69.0 -36.0 14.0). Strong inhibitions were found in sub-cortical structures peaking in caudate nucleus (24.0 24.0 -4.0).

2. **Concentration and focusing**

**Content definition:** We separated three different phases of concentration and focus: low focus was assumed during save exploration and interaction with game mechanics (e.g. buying equipment), middle focus if interaction could be expected, and high focus during these actually happening. The not-interactive phases were not taken into account. The time points of beginning and end of the respective phases were noted.

**Neurophysiological modeling:** Neuronal networks should be disentangled that correlated linearly with the three different levels of focus. The BOLD response was modeled by a generic hemodynamic response to these phases and achieved statistical maps thresholded at a voxel-wise corrected p < 0.05 (Figure 2).

Figure 2: Statistical maps of linear increase of activation with concentration and focus (threshold according to p < 0.05 corrected). Small cluster in the visual system and the cerebellum increased with more demanding game phases. Likewise the success event the ACC was less active and, remarkably, bilateral IPS. Attention and conscious awareness seems to shift from spatial to visual object oriented processing.

Activation increasing with focus

Activation decreasing with focus
**Neuronal network:** With higher focus neuronal activity increased in visual (MNI = -54.0 -66.0 10.0), premotor (-6.0 10.0 72.0) and neocerebellar cortex (-44.0 -70.0 -26.0) but decreased in orbitofrontal cortex (22.0 26.0 -16.0) and intraparietal sulcus (IPS; 52.0 -58.0 52.0).

3. **Direct feedback of game success**

*Content definition:* The game structure of the FPS reveals success events – which are killing (see definition of game success under 1.). The rater can readily identify these events. Moreover the events could be rated in how far they drew explicit attention. This parameter modeled the degree of feedback obtained from this event. The time points of the events were noted and the level (low, medium, high) of the attention to the respective event.

*Neurophysiological modeling:* We assume that the success elicited momentary neuronal responses in circumscribed networks. The BOLD response was modeled by a generic hemodynamic response to these events modulated by the direct feedback-parameter. Statistical maps modeled the linear effect of this parameter. Since the trend reflects a modulation only, we cannot expect circumscribed networks to respond consistently over the entire group. Therefore a descriptive threshold was set at voxel-wise un-corrected p < 0.05 (Figure 3). In addition we selected both main ROIs from the responses to success events: the affective or rostral subdivision of the anterior cingulated cortex (rACC) for increased and vermis of the cerebellum for decreased activity during success. These ROIs were tested bilaterally for response modulation depending on the attention parameter.

Figure 3: Statistical maps on activation and deactivation during direct feedback (threshold according to p < 0.01 un-corrected). At this threshold the descriptive overview suggests a parito-frontal network that increased activity and mainly subcortical areas that decrease the more direct the feedback could be perceived.

Positive with direct feedback
Negative with direct feedback

Table 2: ROI analysis for direct feedback.

<table>
<thead>
<tr>
<th>ROI</th>
<th>contrast</th>
<th>signal change</th>
<th>MNI</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>vermis</td>
<td>positive success</td>
<td>-125±34</td>
<td>-4</td>
<td>-78</td>
<td>-40</td>
</tr>
<tr>
<td>rACC</td>
<td>negative success</td>
<td>-135±43</td>
<td>4</td>
<td>16</td>
<td>-4</td>
</tr>
</tbody>
</table>

4. Control over the activity

Content definition: The game structure of the FPS reveals transitions that change the pass of the game. These are of course killing or being killed, but also deciding to quit a fight, entering a fight, assessing the equipment menu etc. The rater can readily identify whether these events are initiated by the player – i.e. active – or by the game mechanism or other game character – i.e. passive. The time points of the events and their type were noted.

Neurophysiological modeling: We assume that active and passive transitions elicited a momentary neuronal response in circumscribed networks. The BOLD response was modeled by a generic hemodynamic response to these events and achieved statistical maps thresholded at a voxel-wise corrected p < 0.05 (Figure 4).
Figure 4: Statistical maps of activation and deactivation during active transition (threshold according to $p < 0.05$ corrected). A cortico-thalamo-cerebellar network is activated during active transition. Medio-fronto and temporo-polar areas are inhibited together with bilateral IPS.

Signal increase during active transitions

Signal decrease during active transitions

*Neuronal network:* Active transitions correlated positively with a distributed cerebellar-motor and visual network. Relative inhibition during the active transition was found in orbito-frontal and temporo-polar cortex. The passive transition showed no significant voxel even at the descriptive threshold according to an uncorrected $p < 0.001$.

5. **Clear goal versus lack of goal**

*Content definition:* Goal oriented behavior can be assumed most of the time course. From a behavioral perspective remarkable phases are prolonged save stated, i.e. the participants have no actual task and do not change it over more than 10 sec. The phases were termed ‘lack of goal.’
Neurophysiological modeling: Neuronal networks should be disentangled that activate or deactivate during phases with lack of goal. The BOLD response was modeled by a generic hemodynamic response to these phases and achieved statistical maps thresholded at a voxel-wise corrected p < 0.05 (Figure 5).

Neuronal network: Lack of goal led to increased activity in the dorsal (cognitive) subdivision of the ACC extending into the motor network and the precuneus (Figure 5A). Decreased activity was found in bilateral IPS and FFA (Figure 5B).

Figure 5: Statistical maps of activation and deactivation to lack of goal (threshold according to p < 0.05 corrected). The cognitive subdivision of the anterior cingulum and well as precuneus were more active when a clear goal was lost. The IPS can reflect spatial attention and the FFA face and object recognition that was more active in other phases.

Construct validation based on subjective experience

Behavioral data and construct validation

We measured game enjoyment by the game experience questionnaire (GEQ) as developed for the research network by TUE. Moreover the Positive And Negative Affect Scale (PANAS; Watson, Clark & Tellegen, 1988) was applied before and after the game play. Therefore the increase of the positive affect after the gaming could serve as a measure for game enjoyment and the change of the negative affect as measure for disenjoyment according to our constructs.

Construct validation was considered as multiple correlation between systematic and replicable neurophysiological responses in an individual participant with its behavioral measures as indicators of game experience and enjoyment (see above). A linear regression model with each behavioral
measure as predictor for each response map was investigated across the group. Note that the intercept of these models is comparable to the neurophysiological correlates of the flow components; the second order as described in the previous section is the $0^\text{th}$-order approximation of the current linear regression model. Likewise we investigated peak activation spots in the group maps for the estimated linear trend. To that means the maps were thresholded, first, according to a voxelwise corrected $p<0.05$. Moreover we expected distributed network activity contributing to individual differences. Therefore, second, the maps were also thresholded at $T>2.72$ but only cluster exceeding a corrected threshold according to $p<0.05$ were taken into account. Therefore both approaches were equally conservative and corrected for multiple comparisons across the brain volume. The first method should be sensitive to circumspect and well-localized activations whereas the cluster threshold would indicate distributed activations which may vary across the population. We argue that the second approach should be more sensitive to detect networks contributing to individual variability.

As a first general finding, none of the considered analysis revealed a voxel with activity exceeding the corrected threshold. Even the less conservative correction with false-discovery rate statistics (less than 5% of the above threshold voxel are false positives; Genovese, Lazar & Nichols, 2002) revealed no such a brain area. Therefore in the following, only significant clusters were reported representing rather diffuse network responsiveness.

1. Construct validation for challenge: We defined success and failure events and calculated the corresponding contrast maps. Behavioral measures for disenjoyment predicted distributed cerebellar and left orbitofrontal networks being more responsive to success (Figure 6A) whereas posterior cingulate and thalamus were negatively correlated with the behavioral measure (Figure 6B). The responses to failure revealed negative correlation with disenjoyment in the right temporal pole and to a lesser degree in the left temporal pole, left orbitofrontal and premotor cortices as well as the brainstem encompassing the periaqueductal gray (Figure 6C). No significant association with game enjoyment was found.

In a reversed approach we can use the neurophysiological data to determine the items of the questionnaires which were correlated with the localized reactivity. From the failure event the nucleus caudatus and the temporal pole emerged as peaks in the maps for the mere contrast or those indicating dependency on the disenjoyment rating. As expected a relevant linear relation (absolute T-value > 1.79; $p < 0.05$) was observed for the items 3 excited (aufgeregt), 7 anxious (ängstlich), 8 hostile (feindselig), 11 irritable (reizbar), 15 nervous (nervös), 20 timid (furchtsam), 21 aggressive (aggressiv), and 22 calm (gelassen). All but the last item were negatively associated with temporal pole reactivity to failure events. For the left temporal pole items 7, 11, 15, and 21 failed the significance level. For left and right nucleus caudatus, 2 or 3 items past the threshold as expected from random numbers. Therefore we suggest that lack of the temporal pole activation to failure events is quite specifically related to irritable-aggressive feeling.
Figure 6: Statistical maps of behavioral prediction of (A) higher and (B,C) lower individual responsiveness to success (A,B) and failure events (C, threshold for cluster size according to $p < 0.05$ corrected). (A) Cerebellar and left orbitofrontal networks activity correlated positively with disenjoyment whereas (B) higher activity in posterior cingulated and thalamus predicted less disenjoyment. Only negative associations between failure response and disenjoyment were observed with the highest peak at the right temporal pole (C); other cluster comprised periaqueductal gray and orbitofrontal and premotor cortices.

(A) Positive association of disenjoyment and success

(B) Negative association of disenjoyment and success

(C) Negative association of disenjoyment and failure
2. **Construct validation for concentration and focusing**: We defined a focus metric and calculated the according contrast maps. Only behavioral measures for disenjoyment correlated only negatively with focus responses in right motor areas (Figure 7).

3. **Construct validation for direct feedback** could not be evaluated since no rigorous statistical maps were available but only ROI analyses were conducted.

4. **Construct validation for control over the activity**: We defined active and passive transitions. No activation patterns were found for the passive transition, therefore we investigated on response prediction for active transitions. Responses to active transition were negatively correlated with disenjoyment in lateral prefrontal and premotor areas (Figure 8). No positive correlation or association with enjoyment emerged.

Figure 7: Statistical maps of behavioral prediction of lower individual responsiveness to focus (threshold for cluster size according to p < 0.05 corrected).

Negative association of disenjoyment and focus

![Negative association of disenjoyment and focus](Image)

Figure 8: Statistical maps of behavioral prediction of lower individual responsiveness to active transition (threshold for cluster size according to p < 0.05 corrected). Inferior frontal gyrus and premotor cortex showed a negative association of responses to active transition and disenjoyment.

Negative association of disenjoyment and active transitions

![Negative association of disenjoyment and active transitions](Image)
5. **Construct validation for goal**: We defined phases with apparent lack of goal as more than 10 s in a safe situation. Enjoyment correlated negatively with responses in bilateral insular cluster extending into the amygdala (Figure 9A & B). Disenjoyment was associated with higher responses in bilateral medial orbitofrontal cortex (Figure 9C) and with lower responses in precuneus and hippocampus (Figure 9D).

Figure 9: Statistical maps of behavioral prediction of lower individual responsiveness to lack of goal (threshold for cluster size according to p < 0.05 corrected). Bilateral clusters revealed a negative association of brain reactivity to lack of goal in (A) the insula and (B) the amygdala. (C) Medial orbitofrontal as well as (D) left precuneus and left hippocampal responses were associated with disenjoyment ratings.

Negative association of enjoyment and lack of goal

(A) $z=2$  (B) $y=18$

Positive association of disenjoyment and lack of goal

(C)  $y=59$

Negative association of disenjoyment and lack of goal

(D) $x=16$  $y=-8$
Discussion

We argued that game enjoyment should be characterized by the presence of Flow factors, whereas disenjoyment would be marked by their lack. Based on the content analysis, we found the behavioral measures for five out of nine factors contributing to flow and we demonstrated that we can find specific brain activation patterns for each of them. The evaluated contrasts suggest contribution of the reward system, the cerebro-thalamo-cortical motor lip, the visual system, and dorsal parts of the prefrontal cortex. In addition to information on the immediate analysis, we can conclude that the constructs 1.-5. can be validated with neurophysiological data. Only the direct goal-dimension lacks in power indicating a need to refine the construct definition; alternatively one may speculate that the given operationalizing is not reasonable on an operant or neurophysiological level.

Interestingly we can moreover relate the observed pattern to neurophysiological correlates. The areas reflect specific emotions closely related to the observed project, e.g. nucleus caudatus inhibition for lack of reward, the IPS for spatial navigation and action awareness, or the cerebellum for temporal processing:

6. Activity is intrinsically rewarding which is contrasted with the inhibition when activity fails and the character dies. The involvement of the reward system and its cognitive and affective modulation is discussed in more details below.

7. Loss of self-consciousness may be reflected in the loss of conscious perception as reflected in the IPS inhibition. These situations are when active transitions are taken, clear goal are leading the behavior or the focus is increased. This finding is underpinned by the simultaneous activation of the fusiform gyrus which is also activated to active face and object recognition. In how far this loss of conscious perception is also bound to a loss of self-consciousness is not clearly decipherable from the present data. The target structure for this would be the temporo-parietal junction which could not be observed with the theoretically derived contrast estimates.

8. Distorted sense of time would be reflected in functional changes in cerebellum and prefrontal cortex (Mathiak et al., 2002, 2004). The cerebellum is involved in the generation of every of the construct driven constructs and in the processing of success events in relation to the perceived disenjoyment. The function of the cerebellum cannot be reduced to a pure temporal one, but to subserved optimal performance, we can assume that interferences at the level will influence temporal processing and eventually diction the time perception.

9. Merging action awareness (awareness is only focused on the activity) can be most readily associated with the visual system where signal is increased if fMRI activity is associated with the conscious perception in the specific domain.

These above consistent activation pattern, involving IPS, motor loop, reward system, visual and auditory domain etc, reflects the remaining four factors of flow construct for which no behavioral measures were available.

Reward system and its modulation

It was shown that structures belonging to the reward system, that is midbrain dopamine regions, the cingulate cortex and striatum, detect the difference between expected reward and the received outcome. In particular, phasic decreases in activity are observed when reinforcements are worse than expected or not given (a negative prediction error; Schultz et al., 1997, 2004). O’Doherty et al. (2004) demonstrated that activity in the anterior striatum, mainly the caudate nucleus, is correlated with the reward-prediction error during learning. Further studies extended this result reporting
caudate activity also in paradigms that did not include behavioural learning (Delgado et al., 2000; Tricomi et al., 2004). Haruno and Kawato (2006) demonstrated that caudate nucleus activation is mostly correlated with reward-prediction-error at the reward feedback. They argue that caudate role is, like ventral striatum, mainly engaged in the learning process controlled by comparing actual and predicted rewards.

We observed deactivation of caudate nucleus in subjects when they failed in the game. This is in accordance with the reported caudate role in reward-prediction error: the deactivation occurred whenever subject did not receive an expected reward (e.g., was killed instead of killing). We did neither observe caudate nucleus nor other elements of the reward system activated in the success contrast. Conceivably tonic activation of the reward system could be found during the game. That would facilitating goal-directed behaviour and responses to other events that were rewarding to the subject but more difficult to operationalize. Consequently this activation was not visible in the studied contrasts. To test this hypothesis we suggest introducing a control game containing also less rewarding phases and less stress on goal directed behaviour.

Temporal pole (TP) activation are frequently observed in complex emotional tasks, such as theory of mind, but less frequently observed in simpler emotional tasks, such as emotional face perception. In particular, this region is activated in paradigms involving socially important narratives (Olson et al., 2007). Interestingly, TP activations are frequently listed in tables but are not discussed (Beauregard et al., 2001; Moll et al., 2002). Moreover, due to its location, this is an area of strong artefacts in fMRI due to air-tissue transitions leading to magnetic field inhomogeneities and therefore its activation can often be missing from neuroimaging data. In our case, multiecho EPI sequence with alternating phase encoding polarity allows to improve partially for those artefacts and obtain robust signal from this region.

In our study, the right TP was less active in response to failure in the game in those subjects who reported higher disenjoyment after the experiment. We can conclude that TP activation in other subject allowed them to evaluate the failure events in broader social aspects of the game and protected them from lowered mood after the game. It is consistent with the studies showing that patients with the atrophy of right but not left anterior temporal lobe present changes of mood including depression and apathy and irritability (Mychack et al., 2001).

To further elucidate the function of the temporal pole and nucleus caudatus in the individual appraisal of failure, investigated the contribution of single items to the prediction and revealed a profile of relevant adjectives. Left temporal pole showed less predictive power for the item and both nucleoli caudati showed just a random association with the behavioural measures. The profile was in agreement with the feeling of anger. In humans, uni- or bi-lateral anterior temporal lobe damage can lead to symptoms of Kluever-Bucy syndrome, first described in monkeys, and encompassing fear and anger, leading to severe socio-emotional disorders. The data thus indicate that the otherwise hardly specified feeling of anger is connected with right temporo-polar areas.

It was shown that structures belonging to reward system, that is midbrain dopamine regions, the cingulate cortex and striatum, detect the difference between expected reward and the received outcome. In particular, phasic decreases in activity are observed when reinforcements are worse than expected or not given (a negative prediction error; Schultz et al., 1997, 2004).

**Construct validation**

A non-trivial pattern for the association of neuronal responses to game elements and subjective enjoyment rating appeared. As a general result, significant effects were only found in terms of cluster size but not for individual peaks. Therefore we argue that the associations of subjective
experience and brain activation patterns were rather diffuse or variable across subjects than focal and identical across the group. Individual differences of the processing of complex tasks, thus, may be rather attributed to changes of complex networks than circumscribed brain areas.

The only association with positive enjoyment ratings was that insular and amygdala activity during goal-oriented playing phases. This indicates that the livelier and the more emotional the game playing is experienced while trying to achieve a goal, the higher subjective enjoyment will be rated.

The remaining associations were seen in relation to disenjoyment measures. Globally this indicates that the neurophysiological method can preferentially detect frustration and negatively experienced events.

Specifically we found that in the orbitofrontal cortex, disenjoyment was positively correlated with responses to success and to lack of goal. The orbitofrontal was suggested as cognitive controller to inhibit emotional responses and in particular aggressive responses (Grafman & Litvan, 1999). We can hypothesize that subjects involving inhibition of aggression are more prone to get frustrated by the game.

Several responses to game events showed negative correlation with disenjoyment: in the temporal pole responses to lack of reward, premotor responses to increased focus, frontal responses to active transitions, as well as precuneus and hippocampus responses to lack of goal. These patterns may reflect cognitive control over emotional responses. Surprisingly they seem to be rather domain specific. For instance the right temporal pole predominantly inhibits negative feelings to failure whereas the frontal cortex controls the emotional outcome measured with reference to active transitions. Furthermore the areas clearly do not overlap with the structures showing significant responses in a specific domain. This underpins the idea of a hierarchical control: Emotions can be elicited by elementary gaming events but the feelings resulting from the allover integration of emotional events and their cognitive interpretation may be guided by distant and distributed cortical loops.

For this complex hierarchy, little neuroscientific evidence is available so far. Olson et al. (2007) pointed out in their recent review that the modulatory function of the temporal pole has been widely neglected or even re-interpreted. Other areas such as posterior cingulate and precuneus are even less clear with respect to their specific contribution of processing of feelings and emotions. Nevertheless a further understanding of such hierarchical control seems to be crucial to link the relatively well-defined correlates of emotions to feelings, the subjective experience or even the behavioral outcome in questionnaires or media selection (Damasio, 1996).

2.3.2 Conclusion

We presented a content analytical approach that captured dimensions of a flow state. The constructs could be largely underpinned by neurophysiological imaging of brain function with fMRI. The patterns reflected different neural subsystems that could be related to visual perception, motor planning, and spatial attention as well as the reward system. Remarkably the reward system emerged only in terms of inhibition to the failure events. The method reveals components of game experience and enjoyment as represented in the flow concept. However the networks do not indicate a convergence to a common concept. Conceivably, elements that are well-defined can be displayed with neurophysiological methods; for complex concepts as flow or enjoyment, theoretical model building is required. Finally links between measures of subjective experience and neural response modulation could be established. The patterns were distributed and domain specific.
Therefore a further iterative approximation of models of game experience based on subjective reports and those on objective measures is required.
2.4 HGO/BTH

The FUGA partner HGO/BTH wishes to explore game experience measurement dimensions using a multilayered, multimethod-measurement approach, which also makes this approach rather complicated in the analysis. This chapter reports on the first construct validation study carried out at the BTH Game and Media Arts Laboratory in Karlshamn, where psychophysiological measurements (Eye-Tracking, facial EMG, EEG, heart-rate monitoring, EDA/GSR) and video monitoring were employed alongside psychological questionnaires (BIS/BAS scales, short scale EPQ-R, Aggression Questionnaire) and the Game Experience Questionnaire (GEQ), which is described in more detail in FUGA deliverable D3.3.

Game experience, especially game enjoyment, is an intricate and multifaceted construct that is not easily measurable; this is why HGO/BTH opted for an approach correlating different psychophysiological measurements with subjective data gathered from questionnaires. Although there have been major studies using psychophysiological measurements for game event analysis, a major part of our experimental work is exploratory, because there simply are no established analysis and cross-correlation methods, yet. So, the presented example reflects a broad, descriptive analysis of game experience when using a multimeasurement approach.

The theory we want to confirm is how to narrow defining forms like game engagement, immersion, and flow. We would like to check how they inform and depend on one another. Hence, we present our first construct validation study using multiple psychophysiological measurement methods at the same time within a First-Person Shooter environment, specifically focusing on combat challenges.

2.4.1 Mixed measures construct validity pilot study of combat game experience

Introduction

Many studies on games have assigned and studied the experience of game play as a part of certain psychological mind-sets: emotional affection, immersion, frustration, presence, and flow. In the study presented here, we are looking at experiences during combat in an FPS (First-Person Shooter) game. Partner HGO/BTH specifically designed a set of stimuli levels in Half-Life 2, which was made available to all FUGA partners on the web. During preparation of this experiment, we redesigned a few levels and further adjusted them to our specific needs, so in the end we had three different Half-Life 2 levels for this study:

- **Secret Corridors.** A level with a rough playtime of 10 minutes (depending on the skills and FPS experience of the player), designed to let the player explore a larger area in an outside environment resembling the Half-Life 2 level “Ravenholm”. The areas in this level were designed for an ebb and flow of enemy sets with different strengths, where in the early parts of the level, the player fights rather easy enemies and in the later parts, the player has to face stronger and harder to reach enemies (see Figure 1). The design idea behind this level was to evoke a feeling of “flow” in the player because of the gradual increase of combat challenges and the scattered areas of reward, where the player could power up before the next combat encounter.

2 For more information see http://gamescience.bth.se/research/projects/fuga/
• **Church Walk.** This level was designed to evoke a feeling of boredom in reasonably skilled FPS players. The general playing time for this level is a bit shorter, amounting roughly to 6 minutes (again depending on the skill of the player). The level features an open area with a setting similar to the Secret Corridors level, but here the enemies are mostly of the same type (slow melee attackers), causing less damage to the player with each attack. There is clearly a low level of challenge throughout this level. We expected the combat encounters to elicit negative affect and boredom in the player (see Figure 2).

• **New Experiment.** In this level the player has to go through three rooms only to reach a door in the third room that will quit the level. All rooms are connected with doors that only open after the player has killed all hostile npcs in each room. The first room features slow melee enemies that pose not a low level of challenge for experienced FPS players (see Figure 3). The second room features fast and slow enemies, which constantly put the player under combat pressure. The third room has many small and fast hostile npcs that are difficult to
attack and avoid. Thus, the only way to survive this last room is for the player to evade the
attacks by climbing up a ladder to an exit door. For an extra challenge the player has only a
melee weapon (Half-Life 2 crowbar) and a slowly reloading sniper weapon (Half-Life 2
crossbow) with limited ammunition. The ammunition is scattered across the rooms. There is
just enough to shoot every npc with an arrow. We wanted to explore gradual challenge and
how the players react to it with this level.

Figure 3: New Experiment Screenshot

Automated Logging

To allow a better analysis of the psychophysiological data, we have integrated logging into Half-
Life 2, which needs some preparation for each level that is to be used as a stimulus. Logging is done
with the following:

- The Tobii 1750 eye tracker, automatically logging the object names of objects of gaze
  within the Half-Life 2 game world. All default geometry (so-called brushes) is logged
  simply as “worldspawn”, requiring the level designer to create special entities on top of
  regular brushes in order to log specific geometry in the world.
- Via the parallel port of the PC to the trigger input of the USB-receiver box of the Biosemi
  ActiveTwo System (using a stimulation cable with 25 pins male Sub-D to 37 pins male Sub-
  D), we send codes (bytes) for psychophysiological data collection. This mechanism was
  created to send event codes to the psychophysiology system, which makes the data analysis
easier (it would otherwise be blind data that would later need to be correlated to a video of
the game session for correlation of psychophysiological responses to in-game events).

To sum up, we have created an elaborate system for the automatic logging of psychophysiological
data, with special focus on simplifying the cross-correlation necessary for analysis. This experiment
was also the first test-run for our software to see how stable it performs and what data can expect
from it.
Methods

Subjects
Data was recorded from 25 healthy male higher education students, aged between 19 and 38 (Mean = 23.48 years, SD = 4.76 years). All students were participants in the local game programs at BTH, Campus Karlshamn, and other participants were working part-time for local game companies. Therefore, we could assume an avid interest in games and a large proportion of participants being from the hardcore gamer demographic. As part of the experimental setup, demographic data was collected with special regard to the suggestions made by Appelman.

Most of the participants were right-handed (88% right-handed, 12% left-handed) and a little more than half of them were not wearing glasses or contact lenses (56% no glasses, 44% glasses/contacts). All the participants owned a personal computer (PC) and 96% rated this as their preferred gaming platform (out of multiple choices), followed by Xbox 360 (56%), Playstation 3 (52%) and PS2 (48%). Although 60% of all participants owned a cell phone, only 8% rated this as their favorite gaming platform. More than half of all participants said they would buy games more than once a year (68%) and all of them played games at least twice a week (with 60% playing games almost every day). When asked how many hours they play during a day, 84% answered that they play between two and four hours. The preferred mode of play was console single player (44%) or pc multiplayer (36%), with only 8% rating pc single player as their preferred mode of play (12% for console multiplayer). Nevertheless, 36% rated First-Person Shooters (FPS) as their favorite game type.

Many of the participants started playing games in early childhood: 44% started to play digital games when they were younger than six years and 40% started between six and eight years old. This leaves only 16% that started to play between eight and twelve years. So, all the participants started playing digital games before they reached twelve years. None of the subjects received any compensation for their participation; many of the students did it because of their curiosity for game research. We did offer chocolate snacks and sparkling water as refreshments after the experiments.

Stimuli
As mentioned above, we used three different Half-Life 2 levels in this experiment as stimuli. The experiments focused on negative affect (boredom), immersion and flow (especially while experiencing gradually increasing challenges). The games were played in a fixed order as required by our lab computer setup. The order was:

1. Secret Corridors (Immersion)
2. Church Walk (Negative Affect, Boredom)
3. New Experiment (Incrementing Challenges, Flow)

Participants played the games on two separate computers due to the nature of the early setup of the psychophysiological logging script.

Procedure
We conducted all experiments on weekdays with the first time slot beginning at 10:00h and the last ending at 20:00h. General time for one experimental session was 2 hours with setup and cleanup. All participants were invited to the newly established Game and Media Arts Laboratory. There, each participant was welcomed and briefed about the procedure of the experiment (for a general
setup of the laboratory see Figure 4: Experimental setup in the Game and Media Arts Laboratory. The participants needed to fill in two forms. The first one was a compulsory “informed consent” form (with a request not to take part in the experiment when suffering from epileptic seizures or game addiction). The next one was an optional photographic release form, which most of the participants signed as well. Next, the participants were led to a notebook computer, where they filled out the initial game demographic questionnaire. After, this they filled out the BIS/BAS questionnaire, the Eysenck EPQ-R questionnaire and the aggression questionnaire. During this time, the appropriate size of the EEG headcap was also selected by the experiment leader. Then, the first sensors, the galvanic skin response electrodes, were attached to the thenar and hypothenar eminences of the subject’s left hand. The electrodes, which are part of the Biosemi ActiveTwo system, were attached first, so they have a longer assimilation time on the subject’s hand (to minimise any possible influence of the electrolyte gel).

Figure 4: Experimental setup in the Game and Media Arts Laboratory

After completion of the first questionnaires, the use of the heart-rate belt and activity watch from Polar Electro was demonstrated. The electrodes of the heart-rate belt were prepared and the participants led to a separate room, were they could attach the belt directly to their chest (skin contact). After this, the activity watch was affixed to the subjects’ left wrist tested to ensure that the heart-rate signal from the belt was received.

Subsequently, the subjects were seated in front of the eye-tracker and game pcs in a comfortable chair. The EEG headcap was fitted on their head and checked for correct alignment; during the first 10 minutes of injecting the electrolyte gel, subjects were asked to relax in the chair. Following this, during applying the electrodes to the headcap, subjects were allowed to play a level of either Half-Life 2 or Portal to familiarize themselves with the controls and environment the experiments would take place in. Next, the facial EMG electrodes were attached in the following order:

1. Orbicularis oculi
2. Corrugator supercilii
3. Zygomaticus major

For some participants, the electrodes were additionally fixated using tape. Since the electrodes were connected to the ActiveTwo base system together with the EEG electrodes, no extra ground electrodes were needed (because of the CMS/DRL electrodes from the EEG cap). Recording of the electrodes (EDA, EMG and EEG) with the Biosemi ActiView software was tested and if the signal was too noisy, the respective electrodes were cleaned and reattached until signals were clearer.

Then the participants were asked to sit comfortably and get into a position that they would keep through the experiment. Using Tobii’s ClearView software, the eye tracker was then adjusted correctly so that both eyes were tracked. Using a calibration program, the participant was asked to fixate his gaze on dots on the screen, which were used to calibrate his eye position. The calibration was sent to the Tobii Eye Tracking (TET) server and ClearView recording was started.

Then, the participants played the games in the order described under stimuli above. Each game session was set to 10 minutes, but in general participants could finish all game levels before this. After playing the first level, participants had to switch computers because of a software bug. After each level, each participant was given a paper version of the game experience questionnaire to rate their experience. Unfortunately the paper versions of the questionnaire were collected independently without markings, making a direct correlation with each participant impossible. After completion of the last questionnaire, participants were thanked for their participation and escorted out of the lab.

**Measurement instrumentation**

As already mentioned before, we have used a combination of measurement tools within our study. The measurement tools in detail are:

- **Self-reported measures.** Using the BIS/BAS scales, short scale EPQ-R and the Aggression Questionnaire, Demographic information using suggestions from Appelman and evaluative self-reports using the game experience questionnaire (GEQ) from FUGA Deliverable 3.3 after game sessions.

- **Facial EMG.** We recorded the activity from the orbicularis oculi, corrugator supercilii, and zygomaticus major muscle regions as recommended by Fridlund and Cacioppo, using BioSemi flat-type active electrodes (11mm width, 17mm length, 4.5mm height) electrodes with sintered Ag-AgCl (silver/silver chloride) electrode pellets with a contact area of 4mm in diameter (BioSemi instrumentation). The electrodes were filled with low impedance highly conductive Signa electrode gel (Parker Laboratories, Inc.). The raw EMG signal was recorded with the ActiveTwo AD-box at a sample rate of 2 kHz and recorded with ActiView acquisition software.

- **32-channel EEG.** The brain activity was recorded using 32 BioSemi pin-type active electrodes with sintered Ag-AgCl (silver/silver chloride) electrode tips electrodes on different size 32-channel BioSemi active headcaps with the addition of CMS/DRL electrodes (BioSemi instrumentation). The BioSemi Common Mode Sense (CMS) active electrode and Driven Right Leg (DRL) passive electrode replace the ground electrodes used in conventional systems. Low impedance highly conductive Signa electrode gel (Parker Laboratories, Inc.) was applied in the 34 holes in the EEG headcap using a syringe with a blunt plastic tip. The raw EEG signal was recorded with the ActiveTwo AD-box at a sample rate of 2 kHz and recorded with ActiView acquisition software.
- **Galvanic skin response (GSR).** We measured the impedance of the skin using two passive Ag-AgCl (silver/silver chloride) Nihon Kohden electrodes (1 microamp, 512 Hz). The electrode pellets were filled with low impedance highly conductive Signa electrode gel (Parker Laboratories, Inc.) and attached to the thenar and hypothenar eminences of the subject’s left hand well before the recording (~30 Minutes).

- **Eye-Tracker.** We used a Tobii 1750 eye tracker, which is integrated in a 17” TFT monitor and tracks the eyes with two infrared diodes, which generate reflection patterns on the eyes’ corneas. These patterns are then assembled together with the posture of the user making it possible to extract the pupil locations and dilations through digital image processing. The pupils are tracked at a sampling rate of 50 Hz and mapped to screen coordinates through individual calibrations (which we already described above). According to Tobii AB, the spatial resolution is 0.25 degrees and the average accuracy approximately 0.5 degrees. However, it has to be kept in mind that the calibration deteriorates over time as the pupil changes and eyes may become dry. We positioned the players approximately 60 cm away from the eye tracker, which was set to run at a resolution of 1280x1024 pixels. Gaze data was recorded using our in-house logging system (on the client running Half-Life 2) and Tobii ClearView running on a Dell XPS with a frame capture card and a Logitech QuickCam Chat attached to it, which allowed recording both of the user’s face and screens of the game levels.

- **Heart-Rate.** Heart-Rate was recorded using the Polar S810i heart-rate monitor (including the watch receiver and a belt in size M-XXL). The data was saved from the receiver on a computer through an infrared connection. All data was stored with the Polar S-series toolkit and an HP notebook computer.

- **Video observation** from over the shoulder. A Sony DCR-SR72E video camera (handycam) PAL was put on a tripod and positioned approximately 50 cm behind and slightly over the right shoulder of the player for observation of player movement and in-game activity, the game screen (on the eye tracker) being visible to the camera at all times. This data was used for coding player behavior and for correlation with the “blind” EEG and EMG data.

### Results

In this section, we will focus on presenting the results from the Game Experience Questionnaire data. The levels used in this experiment were created such that we expected the easy level to induce boredom (Church Walk), the medium level (Secret Corridors) to induce flow and immersion and our new experiment to be challenging but feasible to complete for skilled players. Since we did know that our participants come from a hardcore gaming community, some part of this was exploratory, since we could only partially anticipate player skill levels.

Table 1 shows the reliability and descriptive statistics of the GEQ items in the easy level the participants played. Our personal observations showed that players with less experience in playing Half-Life 2 liked this level a little bit better than those that reported to have played the game before. Overall, we can see the easy level is perceived as less immersive and less challenging than the other levels. The level scored highest in the areas of competence and positive affect.

This was in line with our expectations. An easy level is supposed to elicit a high competence level, and this also proved to be the most reliable of the measures (Cronbach’s alpha = 0.92). The challenge level was insufficiently challenging for most of the hardcore players to reach a flow state. This explains the lower than median scores of flow (Mean = 1.61). Even though negative affect scores low as a mean value here (0.85), it has the widest range (1.5). This could be interpreted as
different influences of the challenge level on different kinds of players. Players that played Half-Life 2 before and liked the environment and settings might not have liked the overall impression of the easy game level, which could also be related to a low score on sensory and imaginative immersion (0.87).

Table 1 Descriptive statistics and reliability of GEQ components for Church Walk (easy level)

<table>
<thead>
<tr>
<th>GEQ component</th>
<th>Cronbach’s Alpha</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Range</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensory/Imaginative Immersion</td>
<td>0.872</td>
<td>0.854</td>
<td>0.148</td>
<td>0.625</td>
<td>1.042</td>
<td>0.417</td>
<td>0.022</td>
</tr>
<tr>
<td>Tension</td>
<td>0.824</td>
<td>1.028</td>
<td>0.287</td>
<td>0.542</td>
<td>1.333</td>
<td>0.792</td>
<td>0.082</td>
</tr>
<tr>
<td>Competence</td>
<td>0.808</td>
<td>2.549</td>
<td>0.432</td>
<td>1.792</td>
<td>3.083</td>
<td>1.292</td>
<td>0.186</td>
</tr>
<tr>
<td>Flow</td>
<td>0.918</td>
<td>1.611</td>
<td>0.364</td>
<td>1.208</td>
<td>2.083</td>
<td>0.875</td>
<td>0.132</td>
</tr>
<tr>
<td>Negative affect</td>
<td>0.849</td>
<td>1.181</td>
<td>0.563</td>
<td>0.667</td>
<td>2.167</td>
<td>1.500</td>
<td>0.317</td>
</tr>
<tr>
<td>Positive affect</td>
<td>0.924</td>
<td>2.083</td>
<td>0.246</td>
<td>1.667</td>
<td>2.292</td>
<td>0.625</td>
<td>0.060</td>
</tr>
<tr>
<td>Challenge</td>
<td>0.754</td>
<td>0.819</td>
<td>0.407</td>
<td>0.333</td>
<td>1.500</td>
<td>1.167</td>
<td>0.166</td>
</tr>
</tbody>
</table>

Note: N = 24, N of Items = 6, GEQ scale has values between 0 and 4.

The internal consistencies of the items on the game experience questionnaire were found to be high, with Cronbach’s alpha ranging from 0.75 to 0.92 for this level.

Table 2 Descriptive statistics and reliability of GEQ components for Secret Corridors (medium level)

<table>
<thead>
<tr>
<th>GEQ component</th>
<th>Cronbach’s Alpha</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Range</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensory/Imaginative Immersion</td>
<td>0.788</td>
<td>1.365</td>
<td>1.100</td>
<td>0.857</td>
<td>1.714</td>
<td>0.857</td>
<td>0.089</td>
</tr>
<tr>
<td>Tension</td>
<td>0.671</td>
<td>0.833</td>
<td>1.010</td>
<td>0.429</td>
<td>1.714</td>
<td>1.286</td>
<td>0.207</td>
</tr>
<tr>
<td>Competence</td>
<td>0.856</td>
<td>2.500</td>
<td>1.137</td>
<td>1.238</td>
<td>2.905</td>
<td>1.667</td>
<td>0.403</td>
</tr>
<tr>
<td>Flow</td>
<td>0.837</td>
<td>1.937</td>
<td>1.178</td>
<td>1.143</td>
<td>2.429</td>
<td>1.286</td>
<td>0.236</td>
</tr>
<tr>
<td>Negative affect</td>
<td>0.795</td>
<td>0.714</td>
<td>1.027</td>
<td>0.238</td>
<td>1.667</td>
<td>1.429</td>
<td>0.251</td>
</tr>
<tr>
<td>Positive affect</td>
<td>0.859</td>
<td>2.603</td>
<td>1.044</td>
<td>1.905</td>
<td>3.000</td>
<td>1.095</td>
<td>0.158</td>
</tr>
<tr>
<td>Challenge</td>
<td>0.532</td>
<td>1.270</td>
<td>1.069</td>
<td>0.762</td>
<td>2.333</td>
<td>1.571</td>
<td>0.370</td>
</tr>
</tbody>
</table>

Note: N = 21, N of Items = 6, GEQ scale has values between 0 and 4.

For the next level in the experiment, Table 2 gives an outline of descriptive statistics and reliability. For this experiment the reliability of items on the GEQ was unfortunately found to be barely
acceptable for challenge (Cronbach’s alpha = 0.53) and tension (Cronbach’s alpha = 0.67) measures. The other items scored well in the range between 0.79 and 0.86. Again, competence was found to have the highest range between values (range = 1.7). In general, we could find higher ranges between the values for this level. The lowest range was luckily for the measure of game immersion, which is exactly what this level is designed for (0.86). This will allow for strong correlation with psychophysiological measures, despite the lack of exact linking of questionnaires to psychophysiological data.

This level was designed for a more immersive experience, which shows in the highest mean (1.4) for all three levels for the item sensory and imaginative immersion. However that is still lower than the general median of the GEQ (2). Thus, none of the levels managed to score highly on immersion. However, this could also be due to the experimental conditions of playing in a laboratory rather than in an ecologically valid gaming environment.

The highest means are found for items competence (mean = 2.5) and positive affect (mean = 2.6), which are also found to be the most reliable of the measures (Cronbach’s alpha = 0.86). The competence level was surprisingly similar to that of the easy level even though this level has more intense combat challenges and a higher enemy count. We could argue that this reflects higher values of flow (mean = 1.9) and challenge (mean = 1.3). It could also be argued that this is the result of the design of the level, with an ebb and flow of combat challenges (scattering relief areas inside the level) leading to higher immersion. Thus, positive affect could be connected to immersion and challenge.

Table 3 Descriptive statistics and reliability of GEQ components for New Experiment (high challenge level)

<table>
<thead>
<tr>
<th>GEQ component</th>
<th>Cronbach’s Alpha</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Range</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensory/Imaginative Imm</td>
<td>0.859</td>
<td>1.206</td>
<td>1.254</td>
<td>0.667</td>
<td>1.714</td>
<td>1.048</td>
<td>0.165</td>
</tr>
<tr>
<td>Tension</td>
<td>0.881</td>
<td>1.897</td>
<td>1.355</td>
<td>1.286</td>
<td>2.381</td>
<td>1.095</td>
<td>0.243</td>
</tr>
<tr>
<td>Competence</td>
<td>0.938</td>
<td>1.992</td>
<td>1.305</td>
<td>1.714</td>
<td>2.190</td>
<td>0.476</td>
<td>0.024</td>
</tr>
<tr>
<td>Flow</td>
<td>0.937</td>
<td>2.206</td>
<td>1.323</td>
<td>1.762</td>
<td>2.571</td>
<td>0.810</td>
<td>0.099</td>
</tr>
<tr>
<td>Negative affect</td>
<td>0.887</td>
<td>0.889</td>
<td>1.254</td>
<td>0.476</td>
<td>1.714</td>
<td>1.238</td>
<td>0.203</td>
</tr>
<tr>
<td>Positive affect</td>
<td>0.905</td>
<td>2.278</td>
<td>1.306</td>
<td>1.905</td>
<td>2.667</td>
<td>0.762</td>
<td>0.095</td>
</tr>
<tr>
<td>Challenge</td>
<td>0.679</td>
<td>2.333</td>
<td>1.321</td>
<td>1.238</td>
<td>3.190</td>
<td>1.952</td>
<td>0.547</td>
</tr>
</tbody>
</table>

Note: N = 21, N of Items = 6, GEQ scale has values between 0 and 4.

In the last level of our study, the new experiment, we provided intense combat challenges and allowed for no safe spots after combat has been entered, thus putting the player under a lot of pressure. The statistical results of the GEQ are shown in Table 3. Again, the reliability of the challenge item was found to be only just acceptable (Cronbach’s alpha = 0.68), but all other measures worked excellently with Cronbach’s alpha ranging from 0.86 to 0.94.

As expected, this level rated highest on the challenge item (mean = 2.3). The level also worked well for flow as this higher level of challenge led to players experiencing a flow state (mean = 2.2). Surprisingly, even with an increase in challenge, players still felt competent, with the competence mean being just at the median of the GEQ (mean = 2.0). Positive affect was found to be high (mean
but lower than for the medium level with higher immersion. Thus, positive affect appears to be connected to immersion. Tension scored highest in this level (mean = 1.9), suggesting that tension and challenge are interrelated. Also, this level still included an ebb and flow of challenges, but on a more direct level and with a more intensive combat, related to weapon reload time and enemy spawning time.

Lastly, we found that standard deviation on the easy level was low for all items (0.2 to 0.6), a little higher on the medium level (1.0 to 1.2) and highest on the challenging level (1.3 to 1.4). This could be explained by individual differences being more visible as the challenge of the gameplay increases.

**Discussion**

The preceding section presented the results of the Game Experience Questionnaire (GEQ), which we used to provide added construct validity of the GEQ and as input to ongoing analyses incorporating the other psychophysiological data collected.

From a design aspect, we created the three levels with different aspect of the game experience in mind, low challenge and boredom (Church Walk), medium challenge and immersion (Secret Corridors) and high challenge in combination with combat-induced flow (New Experiment).

While in the first two levels, the player could choose among various weapons, the last level was explicitly designed only around one shooting weapon (the crossbow). The amount of ammunition was counted to be just enough if the player shot each enemy with one arrow. Of course, this did not happen often because of the frequency of attacks and players trying to find a safe spot in the room. If ammunition was out, the players had to fall back to the melee weapon, the crowbar, and jump around while trying to fight off the enemies. The attack frequency was raised throughout the level, which would naturally also raise the challenge level. This was the only level where players could die more than once and explicitly because they were killed by the enemies. What is interesting about this is that one would expect a higher level of frustration and maybe negative affect than that revealed by the experiment.

The results above show that the last level is not the one with the highest negative affect mean. The low challenge and boredom level is the one that scores highest on negative affect, thus we could say the absence of enough challenge leaves the players in a negative emotional state, while more challenge provides a much better experience in general, rating high on positive affect and flow. This may also be associated with an unusual aspect of the level design, compared to regular Half-Life 2 gameplay, concentrating on fighting and managing resources (ammunition); this reduces general Half-Life 2 gameplay to the minimal aspect of mastering different forms of combat.

**Future research aspects**

As already stated above, the large amount of data collected in a newly established laboratory means that analysis and cross-correlation of the different measures is ongoing. This will include correlating especially the EEG and EMG data with findings derived from the game experience questionnaire, to further understand the connections between positive affect and challenge.

Also, from a design perspective, it interesting to look at certain events inside the game levels and find out, what response was elicited by a certain event. We would also like to integrate our current findings with the other self-report measures such as the personality and aggression questionnaires that were used.
2.4.2 Conclusion

It may be concluded that in terms of construct validity the GEQ works well for our purposes. We would like to further examine the low reliability of challenge with two of our levels, because the item results were not as expected. Further stimulus tuning may be required for ongoing experiments, since the analysis of flow may require careful design of difficulty for our levels. Overall the results of this experiment are very encouraging for ongoing work aimed at creating a more thorough understanding of the constructs of flow and immersion.
2.5 TUE

The current chapter reports on a first construct validation study in which we experimentally investigated two complementary approaches to measuring player experience. First, we report on a set of behavioral indicators, such as pressure patterns exerted on a physical control device, and postural responses, that could potentially serve as real-time indicators of player experiences. Secondly, we report on the construct validity of the Game Experience Questionnaire (GEQ), which was developed earlier in the FUGA project, and has been reported separately in Deliverable D3.3.

In our experimental work, we are taking a multi-method, multi-measure approach whereby we anchor and cross-validate the different behavioral measures using a variety of other measures, including psychophysiological measures (heart rate and galvanic skin response) and measures of self-report, including the GEQ. This multi-method approach is central to the FUGA philosophy, as it allows a fuller insight into the player experience of digital games. The combination of multiple measurement modalities can reduce uncertainty associated with measuring a single modality, resulting in increased robustness and wider applicability of the total set of measures. Limitations particular to one measure may be overcome or compensated by using corroborating evidence emerging from another measure.

For construct validating both the behavior-based set of measures, as well as the GEQ, we have designed an experiment which used customized levels of a digital game, varying in difficulty, to induce boredom, flow or a state of optimal enjoyment, and frustration. By inducing these player experiences, we do not need to infer these states, nor wait for their spontaneous occurrence during natural game play. Thus, this allows us to more reliably associate behavioral response patterns with affective states (see also Scheirer, Fernandez, Klein & Picard, 2002). It is fair to say that automatic behavior-based measures are not yet well-established or tested in literature, and that there are currently no workable ‘off-the-shelf’ solutions available, neither in terms of sensors, nor in terms of data analysis and interpretation. Hence, this work can be regarded to be highly innovative and explorative in nature.

It should be noted that in the large body of literature on media reception and reaction processes, the behavioral impact of media is usually discussed in terms of how media affect behavioral tendencies after episodes of media exposure. For example, a significant body of digital games research is looking at potential associations between exposure to violent games and the development and manifestation of antisocial (e.g., aggressive) behaviors. However, when we refer to behavioral responses in the current chapter, we are referring to naturally occurring observable or otherwise measurable behaviors as they are exhibited during an episode of gameplay, as part of a direct interaction with, or response to, unfolding game events.

In this chapter, we will first present the construct validation study of behavioral indicators of player experience, with a focus on indicators of flow, frustration, and boredom. After a brief discussion of relevant literature on real-time behavior-based measures of emotion, we will describe the measures we employed in the experiment, including video observations of posture, accelerometer data, pressure chair data, and force measurements on the interface devices (mouse, keyboard). Construct validity of these measures is based on the sensitivity of each measure to pick up on the difficulty-level manipulation. Convergent validity is based on the extent to which similar results are found using self-report measures of player experience, including the in-game GEQ (iGEQ), the Self-Assessment Mannequin (SAM), and the FlowGrid.

Subsequently, we will discuss the results of the GEQ in light of the experimental manipulation, providing a further basis for the construct validity of the GEQ, beyond the validation results that
were already reported in D3.3. A valid self-report measure is of critical importance to the FUGA project as a whole, as it is an essential metric to allow comparison of results between partner sites, and between different experiments and measures. Thus, the GEQ can be regarded to act as a kind of 'ground truth' measure of player experience within FUGA.

2.5.1 Construct validation of behavioral indicators of flow, frustration, and boredom

Introduction

Csikszentmihalyi (1975; 1990) studied what makes experiences enjoyable to people. He was interested in people’s inner states while pursuing activities that are difficult, yet appear to be intrinsically motivating, that is, contain rewards in themselves – chess, rock climbing, dance, sports. In later studies, he investigated ordinary people in their everyday lives, asking them to describe their experiences when they were living life at its fullest, and were engaged in pleasurable activities. He discovered that central to all these experiences was a psychological state he called flow, an optimal state of enjoyment where people are completely absorbed in the activity. Flow is a state where someone’s skills are well balanced with the challenges posed by a task. It is characterized by a deep concentration on the task at hand, a perceived sense of control over actions, a loss of preoccupation with self, and transformation of one’s sense of time.

Flow certainly sounds familiar to frequent players of digital games. When a game is effective, the player’s mind can enter an almost trance-like state in which the player is completely focused on playing the game, and everything else seems to fade away - a loss of awareness of one’s self, one’s surroundings, and time. Sweetser and Wyeth (2005) have adopted and extended Csikszentmihalyi’s conceptualization of flow in their ‘GameFlow’ model of player enjoyment, formulating a set of useful design criteria for achieving enjoyment in electronic games. Csikszentmihalyi’s original work on flow suggests that these peak experiences are quite rare – the exception rather than the rule. Nevertheless, the flow model of game enjoyment clearly illustrates the importance of providing an appropriate match between the challenges posed and the player’s skill level. The flow experience can easily break down when the player’s skills systematically outpace the challenges the game can offer (leading to boredom) or when game challenges become overwhelming in light of the available skills (resulting in frustration). Being able to detect frustration and boredom is of importance as indicators of when a person is not experiencing flow, but also, and perhaps more interestingly, because successful games strike a balance between positive and negative emotions (see, e.g., Ravaja et al., 2004). This is in line with the view that games are often being designed with the aim to develop a negative emotion in the face of challenge, only to be followed by a positive emotional peak when the challenge is overcome (Keeker, Pagulayan, Sykes, & Lazzaro, 2004).

In sum, behavioral indicators of involvement or interest are required, as well as indicators of both boredom and frustration. Next, we will briefly review the relevant literature concerning real-time, automated measures of behavior-based indicators of affective experience.

Behavioral expression of player experiences

Behavioral expressions of subjective states are well known to both lay-people and scientists alike. A host of observable and expressive physical behaviors are associated with emotional states. We tend to smile at something funny, move towards something or somebody we like, jump up when startled, hide our heads when scared, or make strong gestures when frustrated. There are a number of
behavioral responses where the human motor system may potentially act as a carrier for the player experiences discussed previously.

Mota and Picard (2003) demonstrated that postural patterns can be indicative of learner interest. They developed a system to recognize postural patterns and associated affective states in real time, in an unobtrusive way, from a set of pressure sensors on a chair. Their system is reportedly able to detect, with an average accuracy of 87.6%, when a child is interested, or is starting to take frequent breaks and looking bored. Thus, the dynamics of postures can distinguish with significant reliability between affective states of high interest, low interest and boredom, all of which are of relevance to a gaming situation as well.

Clynes (1973; 1977) investigated the patterns of motor output of people asked to deliberately express certain emotions through the motor channel (usually a finger pressing on a measuring surface he dubbed the ‘sentograph’). He found that there are distinguishable, stable patterns of pressure and deflection for emotions such as anger, hate, grief, love, and joy, transcending barriers of culture and language (Clynes, 1973). Support for Clynes’ original findings has been varied. Trussoni, O’Malley and Barton (1988) failed to replicate Clynes’ findings using an improved version of the sentograph. Although they did find distinguishable patterns associated with certain emotions, a significant correlation with Clynes’ original sentograms was absent, throwing doubt on the universality of sentic patterns. Hama and Tsuda (1990), on the other hand, did find support for the characteristic waveform patterns associated with ‘sadness’ (long duration of pressure) and ‘anger’ (strong intensity of pressure). Moreover, in their first experiment, Hama and Tsuda did not inform participants that they were interested in measuring emotions, which raises the interesting possibility that identifiable pressure patterns may be associated with spontaneously generated motor expression of emotions. In particular, the sentic expression of anger is of interest as a potential indicator of gamer frustration.

Research by Mentis and Gay (2002) and Park, Zhu, McLaughlin and Jin (2005) provide evidence that the force people apply to interface devices can be interpreted as an indicator of negative arousal. Mentis and Gay (2002) asked a small number of participants to complete several tasks on a word processor. Later, participants were asked to indicate whether and when they experienced a frustrating event. Their results suggest that higher pressure on the touchpad is associated with a frustrating event. Building on these findings, Park et al. (2005) manipulated frustration by asking participants to complete an impossible LEGO assembly task. The instructions for the task and optional online help were presented on a laptop computer, where the pressure exerted on the touchpad was measured. Results indicated that more pressure was exerted on the interface device when participants were encountering problems. Additionally, pressure patterns also correlated with facial expressions showing negative affect, thereby providing evidence that the pressure exerted was indeed related to frustration rather than mere arousal.

Focusing on digital games, Sykes and Brown (2003) have investigated the mean pressure exerted by players on a gamepad’s button as the difficulty level of a game (Space Invaders) was increased from easy to medium to hard. Their results show that buttons on the gamepad were pressed significantly harder in the hard condition than in either the easy or the medium condition. Although the increase in pressure on the gamepad can be assumed to be associated with higher arousal, Sykes and Brown did not determine whether this arousal was positively or negatively valenced, while both states could plausibly occur in a digital game setting. Notwithstanding this limitation, Sykes and Brown (2003) successfully demonstrated that a fairly straightforward behavioral measure such as hand or finger pressure exerted on a button can already be informative about the level of user arousal in gaming situations. In addition, given its relative simplicity, this measure has the potential to be analyzed in real-time and be used to adaptively influence the game dynamics.
In sum, evidence from literature suggests that patterns in pressure and postural movement data may be indicative for experiences such as interest, arousal, frustration and boredom. Next, we will assess the validity of such indicators for player experiences in a digital game, by exploring their response patterns to experimental manipulations of difficulty level, as well as their convergence with a set of self-report measures.

Method

Design

Employing modified levels of the first person shooter Half Life 2, game difficulty was manipulated on three levels (easy vs. moderate vs. hard), according to a within groups design. Various behavioral indicators were tested as dependent variables, including postural responses and pressure exerted on keyboard and mouse. In addition, a range of self-report measures, including the in-game Game Experience Questionnaire (iGEQ), the Self Assessment Manikin (SAM), and a specifically developed FlowGrid were used as corroborative measures.

Participants

Thirty-two participants (five females) aged between 17 and 46 ($M_{\text{age}} = 22.42$ years, $SD = 5.57$ years) took part in the experiment. All participants at least occasionally played FPS games, but a substantial part consisted of more frequent players. Two strategies were employed to recruit both casual and avid gamers. A snowball procedure was used to find FPS expert gamers i.e., people who at least play on a weekly basis. The other participants were recruited by means of an email invitation sent to the HTI group’s pool of research participants. The mail invited people to participate in a FPS game experiment and recruited both people who played FPS games on a regular (at least weekly) basis and people who had some experience with playing FPS games but only played these games occasionally (less than monthly).

Participants varied according to the frequency with which they played First Person Shooter (FPS) games: thirteen participants reported to play FPS games on an occasional basis (almost never – sometimes), nineteen participants were more frequent players of FPS games (often – always). Ten participants reported to have substantial to a lot of experience with the game Half Life 2, whereas the remaining participants had no or limited experience with Half Life 2. Participants received 10€ for their time. For one participant we encountered problems with the recording the video data so we had to exclude this participant from analyses with observational and automated measures.

Experimental setting and procedure

The experiment was conducted in the Game Experience Lab at Eindhoven University of Technology. As experimental stimuli we used three modified levels of the FPS game Half Life 2 (HL2) as created by BTH/HGO. The levels were specifically created to manipulate the difficulty of the level on an easy, moderately difficult, and hard level, while keeping other factors constant, such as the type of enemies, environment and weapons. The three levels were pilot tested with two volunteers familiar with FPS and HL2. From these preliminary tests, we concluded that the moderate and hard level created roughly the desired experiences within the time frame of 10 minutes. That is, the moderate level could almost be completed and players got close to being killed, while the hard level proved to be too difficult to complete with the players getting killed numerous times. The easy level, however, was too short for our purposes. Players could complete it in a matter of minutes. Together with the BTH/HGO group, we further altered the easy level in several ways. Firstly, the level was altered to become 'never ending'. Secondly, the level was
expanded, more than doubling the playing time. Subsequent testing of the level indicated that the altered level was suitable for use as an easy level in our experiment.

The game was played on a Dell XPS PC equipped with a quad core processor and a NVIDIA GeForce 8800GTX 768Mb graphics card, connected to a 20" TFT-screen. We recorded the game output as well as the facial expression, verbal utterances, and behaviors of the participants (see Figure 1). More specifically, three video cameras installed in the lab captured the participants from a frontal perspective, from the side and from the top (see figure below for a view of the experimental setting).

Upon entering the lab, participants were welcomed by the experiment leaders. In order to comply with the ethical requirements, the experiment leaders first inquired about participants’ gaming habits. The main goal of this conversation was to signal any problematic gaming behavior of the participants. As outlined in the ethical discussion of the FUGA project, participants who exhibit a tendency towards addictive game playing were to be excluded from experiments. In order to verify this, we opted for an informal conversation with participants instead of using a formal and distant addiction scale such as the problem video game playing (PVP) scale (Tejeiro Salguero & Bersabé Moran, 2002). Experiment leaders, however, did loosely follow these items during the conversation. They probed participants to talk about their gaming habits, the priority of gaming in their life and the time spent on playing games in comparison to other leisure activities. As such they were able to detect possible indicators of problematic gaming. None of our participants exhibited indications of problematic game behavior. In general, all participants, even the most avid gamers, report to spend ample time at other activities such as school, work, friends, sports, etc.

After this introductory conversation, experiment leaders gave a brief overview of the flow-line of the experiment. They then explained that the experiment involved the use of video and audio recordings and measurement of psychophysiological reactions. Participants signed the consent form and were seated at a desk where the game PC was installed, on a chair fitted with pressure sensors. They were then connected to psychophysiological sensors and an accelerometer (described in more detail in the measurements section).
After reading brief instructions related to the use of the controls in the game, participants played three levels of the FPS game Half Life 2. The order in which the levels were played was counterbalanced. Participants were given ten minutes to play each of the levels. However, the exact play time could be shorter if, for example, the player completed the level within ten minutes. Leaving participants fill the remaining time would have likely resulted in annoyance or irritation, and may have severely influenced participants' subjective evaluation of the game experience. In these cases, the experimenter cut short the experimental session telling the participant that the time was completed and blaming their apparent intrusion during the game on a fault of their own; e.g. not having set the time right. Because more experienced players usually finished the easy level in less than ten minutes, we fixed the playtime for this level at eight minutes.

After each level, participants rated their experiences during game-play on a range of self-report measures administered on a laptop PC by means of an Authorware file (see below). Lastly, at the end of the session, potentially relevant traits were measured and we also collected some general characteristics related to socio demographics and gaming habits. These questionnaires were administered on paper. After this, participants were paid, thanked and dismissed.

**Measurement instruments**

As already mentioned in the procedure outline we have used a range of measurement tools within this study. These measurements include:

- self report measures (iGEQ, GEQ, SAM, and FlowGrid) measuring the subjective experience during and after the game play, and two trait measures
- video observations from three positions (face, side and top views) as well as the game-screen allowing coding of people's behavior during game-play
- accelerometer data acquisition measuring tilt and acceleration on three axes as a measure for people's movement during game-play
- sitting position using a sensor-equipped chair was recorded providing information on people's movement and posture during game-play
- force measurements on the interface devices (mouse, keyboard)
- psychophysiological measures (skin resistance and heart rate)

We describe each of these measurement instruments in more detail below.

**Self report measures**

After each level participants completed a set of self report measures in which they rated their experiences during game-play. Self report measures included manipulation checks, in-game experiences measured by the in-game Game Experience Questionnaire (iGEQ), and additional experience measures aimed at checking the construct validity (convergent validity) of the iGEQ.

The *manipulation check* included one five point bi-polar statement stating "How easy or difficult did you find it to play the level?" ranging from -2 (too easy to play) over 0 (optimal to play) to 2 (too difficult to play). Additionally, after playing all three experimental levels, participants were asked to order in retrospect the ranking of each level according to how boring, challenging, frustrating, and fun they found each level. That is, they were asked to indicate which level they experienced as most boring, which level they experienced as second most boring, and which level they experienced as least boring. This rank ordering was completed for each of these four dimensions.
After each level we administered the *in-game Game Experience Questionnaire* (iGEQ) consisting of seven dimensions with two items per dimension. These dimensions were Positive affect, Negative affect, Tension, Flow, Challenge, Immersion, and Competence (see D 3.3). All GEQ items are measured by means of five point intensity scales with points anchored at *not at all* (0), *slightly* (1), *moderately* (2), *fairly* (3), *extremely* (4). For our analyses, we used the mean value of the two items per dimension. We used the iGEQ, the shorter in-game version of the GEQ, because we did not want to interrupt participants too long between the different levels of game-play. However, after the third and last level, we administered the iGEQ followed by the remaining items of the full GEQ. Since our design was counterbalanced we were able to administer the full GEQ for one third of the levels played. As such, for one third of the levels played, we can compare results of the full GEQ across participants (see Section 2.5.2).

As a measure to check convergent validity of the iGEQ, we used the *Self Assessment Manikin* (SAM), a visual self report scale developed by Lang (1980) and based on Mehrabian and Russell’s (1974) Pleasure-Arousal-Dominance-theory (PAD). The SAM-scale visualizes the three PAD-dimensions. Each dimension is depicted through a set of five graphic figures (manikins) and for every dimension respondents have to indicate which figure corresponds best with their feelings on a nine point scale (see Figure 2). The first dimension P (displeasure/pleasure) ranges from extreme sadness to extreme happiness. The second dimension A (non-arousal/arousal) ranges from very calm or bored to extremely stimulated. The third dimension D (submissiveness/dominance) ranges from a feeling of being controlled or dominated to a feeling of total control. Additionally, we included a SAM-based measure of presence developed by Schneider et al. (2004) as a fourth emotion dimension that possibly applies to digital game experience. This dimension ranges from a feeling of total presence to a feeling of total absence. For each SAM dimension we asked participants to indicate (on a 9-point scales listed below the graphical presentation) which manikin corresponded with their experiences during game-play. Scale values ranged from -4 to 4, with ascending scores corresponding higher pleasure, arousal, dominance and presence ratings. We recoded the presence measure for our analysis such that increasing values represent increasing presence experiences.

**Pleasure dimension:**

![Pleasure Manikins](image1)

**Arousal dimension:**

![Arousal Manikins](image2)

**Dominance dimension:**

![Dominance Manikins](image3)
Presence dimension:

![Figure 2: Graphic representations used at measuring the SAM dimensions](image)

The fourth self report measure is new and was specifically developed for the present study: the FlowGrid. This grid was developed to measure people's subjective experience of their skills in relation to the challenge they experienced during the game (see Figure 3). The FlowGrid provides a quick means of assessing people's experience as a balance between their skills and the experienced challenge. As suggested in flow theory (Csikszentmihalyi, 1975; 1990) this balance between skills and challenge can be related to experiences of flow, frustration and boredom. Within this theory, flow is assumed to arise when there is a match between skills and challenge. When skills exceed challenge, boredom is thought to follow the experience. On the other hand, when challenge outpaces skills, people are expected to experience frustration. The two dimensions in the FlowGrid correspond with this conceptualisation. As such, we are able to link the dimensions of game experience as identified and measured with the GEQ to people’s experiences as captured in the flow theory.

Although the FlowGrid matches closely the conceptual framework of the flow theory, it is a subjectively rated balance of skills and challenge rather than an objective measure of the two dimensions. Participants were, however, specifically asked to weigh each of these two and provide one answer. To make this weighing possible in an easy and intuitive way, we were inspired by the AffectGrid (Russel, Weis, & Mendelsohn, 1989). The AffectGrid was developed as a means of assessing affect along the pleasure-displeasure and arousal-sleepiness dimension. Like our FlowGrid the AffectGrid was designed with a similar goal in mind: i.e. measuring multiple dimensions as a single-item scale. Similar to the AffectGrid, participants were asked to tick a single cell on the FlowGrid spanning across the subjectively rated skills on the vertical axis, and challenge on the horizontal axis (see figure below for a depiction of the FlowGrid as used in the current study). For our analyses we constructed a measure from the FlowGrid ranging from maximum boredom (top left corner of the FlowGrid) through flow (the diagonal in the FlowGrid representing balance between skills and challenge) to maximum frustration (lower right corner of the FlowGrid). The scale ranged from -4 (boredom) to +4 (frustration) and was constructed by subtracting the value on the subjective skill dimension from the value on the challenge dimension.
After the last game session and corresponding self-report measures had been concluded, two trait measures were administered. The first was a Dutch version of the BIS/BAS scale (Carver & White, 1994, translated and validated in Dutch by Franken, Muris, & Rassin, 2005). The second was the Aggression Questionnaire (Buss & Perry, 1992, translated and validated in Dutch by Morren & Meesters, 2002).

**Observations**

Participants were observed using three camera positions: one recording facial expressions, another recording the participants from the side, and a third recording the participants from above (see Figure 4 below for a screenshot of each of the camera views).

![FlowGrid with instruction as used in the experiment](image-url)
These three camera positions were chosen to provide us with accurate recordings of people's movements, their interaction with the interface devices and facial expressions of the participants during the experiment. In addition to the visual recordings a microphone recorded the sounds made by the participants during the experiment. This microphone was equipped with a directional head in order to record people's vocal expressions with minimal disruption in the signal from game events such as explosions and sounds made by attacking or dying zombies. To capture the game output, we directly recorded the video signal to the screen on which the participants played the game. Rather than using another camera, the signal was directly used as a video feed.

Each of the video feeds was recorded on a specially equipped observational PC. This PC was equipped with a video observation card specifically designed to synchronously record the four video feeds. The video files were tagged with a time stamp from the system time clock.

The video files were coded using the Noldus Observer XT software. Importantly, most analyses are dependent on the interactions in the game. Therefore a first step included the coding of the beginning and ending of the initial instruction, the beginning and ending of each of the levels (easy, moderate and hard), and the beginning and ending of each block of questions following the levels. This coding of the flow-line was done for all participants of whom we had video recordings (N=31). Additionally, for 11 participants people's in-game actions were coded in detail. These were randomly chosen from our main sample. However, since we strive for an equal balance between men and women, we did include all five women from our main sample into this sub sample.

In-game events coded include:

- coding of the active weapon at any time in the game (shotgun, submachine gun, crowbar, pistol, grenade, or utilizing an object)
- actions using the left mouse button: attacking a non person character (NPC) opening an object (or attempts to do so), and attacks on the environment
The coding procedure firstly included testing for inter-rater reliability. For this purpose two raters both coded the in-game actions for two participants. For all coded events, except for the forward-backward movement, the reliability proved to be satisfactory to excellent (see Cohen's kappa values in Table 1). On the basis of these reliabilities, we decided that the remaining participants could be reliably coded by one rater rather than having two raters code the actions of all participants, drastically cutting down on the time required for coding.

Besides in-game actions which basically serve as background variables necessary for detailed analysis of other indicators, the video recordings are also used to register overt participant behaviors as indicators of experience. These analyses are still very much in progress (the time required to code the in-game actions we have to date totals up to 8 hours of coding per participant). For the current set of 11 participants, so far only random, forward and backward movement of the person have been coded. Coding of facial expressions is especially time consuming and has not yet been completed at this time and therefore will not be included in the results section of this chapter.

Random, forward and backward movements have been counted as instances where participants showed clear movements either (1) haphazard or randomly around a central position, (2) moving towards the screen, or (3) moving away from it respectively. Very minor movements of the body – people rarely sit perfectly still – were not considered for coding. Since initially we did not establish sufficient reliability for forward-backward movement, the two coders resolved inconsistencies by discussion. Logically this resulted in a substantial increase in kappa values. For the remaining participants, we will continue to code this variable with two coders. However, due to time constraints at the time of writing, we currently only have observations from one coder.

Table 1: Reliability statistics of observations

<table>
<thead>
<tr>
<th></th>
<th>active weapon</th>
<th>left mouse action</th>
<th>right mouse action</th>
<th>forward-backward movement*</th>
</tr>
</thead>
<tbody>
<tr>
<td>participant 1</td>
<td>0.96</td>
<td>0.69</td>
<td>no use of right mouse</td>
<td>0.86</td>
</tr>
<tr>
<td>participant 2</td>
<td>0.98</td>
<td>0.78</td>
<td>0.68</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Cohen's kappa values for two participants as a reliability measure between two independent coders

*this reliability was established after discussion of previous inconsistencies. The actual coding of the behaviors was done separately.

Accelerometer

For each participant, an accelerometer was attached to the back, at the base of the neck, to automatically capture movement of the upper body. The accelerometer used was a Phidgets 3 axis version measuring tilt on the x, y, and z-axes, and acceleration to a maximum of 3G's, which is more than enough for the expected movement of the participants during game play. For the analyses we used the accelerometer data converged over all axes (sum of the squared values for each of the three other axes). This metric represents the acceleration as a function of movement in any direction. These values were averaged per level providing an indication of the average movement during each level.

Sitting position

A second automatic indicator of movement was recorded via a pressure sensitive chair. Sitting position and the number of shifts in position are potential indicators of boredom and of interest. In addition to observed and coded sitting position (forward-backward movement) using the video
streams, we also designed a system to measure people's body position during game-play automatically. A sensor equipped chair was constructed allowing for object measurement of the forward-backward and sideways movements of the participant during game-play. The chair was constructed using force-sensitive sensors built into the legs of the chair. The sensors used were Flexiforce sensors (built by TekScan) designed to measure up to 25Lbs (approx. 11.3 Kg) of force applied to them (for an image of the chair and the measuring system see Figure 5).

![Figure 5: Pressure sensitive chair](image)

The sensors were linked with a measuring system designed by Martin Boschman (technical coordinator of the Game Experience Lab at TU/e). In addition to the force measurements per leg, aggregated measures indicating weight distribution for forward-backward and sideways movements were collected. Forward-backward movement was calculated as the difference between the total force applied to the front legs minus the total force applied to the back legs. All measures were provided with a time-stamp allowing retrospective synchronization with the video recordings. These measurements form a continuous stream of data allowing relating person movement (direction and number of shifts) as observed from the video with these force measurements. Two

---

3 The force applied to the flexiforce sensors used in the measuring devices is returned as a value between 0 and 1000, with 0 representing no force, and 1000 representing maximum force.
measures were constructed from forward-backward movement data: average position and maximum range of movement. For constructing the average position the mean forward-backward movement during the experiment was subtracted from the forward-backward movement chair data. This resulted in a relative forward-backward position of the person in each level relative to the average position. This was done to compensate for likely differences in people’s sitting positions. The second measure distilled from the chair was the range in movement. The range was calculated as the difference between the maximum and minimum output from the chair representing the maximum forward force and backward force applied to the chair due to movement. These measures were calculated for each of the levels separately. The maximum range possible would be from -2000 (maximum pressure to the back legs without pressure to the front legs) to +2000 (maximum pressure to the front legs without pressure to the back legs).

**Force measurements**

People use input devices to interact with games. Potentially, the force exerted on these devices is an indicator of their experience during interaction. In the current study we focused on the force people applied on the two interface devices used to control HL2: the keyboard and the mouse. Both these interface devices were equipped with Flexiforce sensors (built by TekScan) designed to measure up to 1Lbs (approx. 453.6 grams) of force applied to them.

The mouse sensors were mounted on top of two buttons. To increase the likelihood that the participants would press on the sensors when operating the mouse the paddles were reduced in size and the sensors were topped with a small rubber patch. This patch raised the surface of the sensor over the rest of the paddle and discriminated the surface texture of the paddles. The patch thus naturally inclined people to keep their fingers on top of the sensors (see Figure 6 for a view of the augmented mouse). The keyboard was equipped with four pressure sensors attached to the bottom of the keyboard, one on each of the corners of the keyboard. This setup allowed measurement of the force applied to the keyboard regardless of the keys used for the interaction.
Both the mouse and keyboard pressure data were recorded continuously allowing for synchronization of the force on the input devices with discrete in-game events. The data could also be aggregated over lengths of time, e.g., complete sessions. This provides opportunities for event-based analyses, and correlation analyses with self-report measures. For analyses of the force applied to the mouse, two measures were constructed. For both measures values between 0 and 10 were deleted from the dataset as these values are the result of noise in the measurement system. The first measure was constructed using the maximum value of force applied to the left mouse per level. The second measure constructed was the average force applied to the mouse based on the maximum force per event, thus leaving out all values between the onset and end of the mouse press other than the maximum force. As the mouse is used more frequently and consistently throughout the game, we will focus on the results of the mouse data in the current report. The keyboard pressure data will be reported at a later stage.

Psychophysiological recordings

Psychophysiological measures were taken using a TMSi Mobi 6 Bluetooth device. Skin resistance was measured using dry GSR electrodes fitted on participants’ little and ring finger. Heart rate was measured using a NONIN ear clip pulse oximeter model 9000Q. Sampling rate was set to 128 samples per second. These indicators were recorded as corroborative measures and can potentially be used to cross validate measures at a later stage. They will not be reported in the present report.

Results part one: construct validity

The present results section is divided in two parts. The first part reports construct validity tests of all the measures employed in the experiment, including both behavioral indicators and self-report measures. These tests typically comprise of analyses testing the effects of the difficulty manipulation. The second part of the results section addresses convergent validity by exploring correlations between the various self-report measures and behavioral indicators.

Self report measures

The three mods used in this study were designed to represent an easy, a challenging, and a hard level in terms of difficulty, ideally inducing boredom, flow, and frustration. In this section we will first present the results of the manipulation check and then the results for the remaining self-report measures. All effects are tested employing repeated measures ANOVAs.

The one-item manipulation check ‘How easy or difficult did you find it to play the level?’ showed significant differences between each of the three difficulty levels in the expected directions, \( F(2,30) = 120.77; p < .001 \). The easy level was rated as the "easiest" followed by the moderate level, and the hard level as the most difficult level to play. Means for each condition and results of pairwise comparisons are reported in Table 2.

Participants also performed three rank orderings of the levels, according to how boring, challenging, and frustrating they were. As can be seen in Figure 7 and in line with our expectations, the easy level was most frequently rated as the most boring level, followed by the moderate level and the hard level (\( \chi^2 > 13.93, \forall p's > .001 \)). Also as expected, the hard level was most frequently rated as the most challenging, followed by the moderate level and the easy level (\( \chi^2 > 21.12, \forall p's < .001 \)). Additionally, the hard level also proved to be most frequently rated as the most frustrating level (all participants rated this level as most frustrating). The moderate and easy level were equally
often rated as second or least frustrating ($\chi^2$s = 0, n.s.). The moderate level was rated as most fun most frequently, and least frequently rated as least fun ($\chi^2$s > 11.31, all $p$'s < .003).

The components of the iGEQ revealed several significant differences, indicating that the three experimental levels induced different experiences (again, mean scores and pairwise comparisons are reported in Table 2). The iGEQ’s component most closely tied to the difficulty manipulation is of course the challenge dimension. This scale indicated a systematic increase in perceived challenge from the easy level to the hard level, with significant differences between each of the levels ($F(2,30) = 70.02; p < .001$). Subsequently, because this study aimed to explore behavioral indicators of flow, boredom, and frustration, we were interested in whether the three experimental game levels differed on these experiences in a meaningful way. For frustration this was clearly the case. Tension, which taps into frustration, was highest in the hard level followed by the moderate level and the easy level and each subsequent level differed significantly from the previous one ($F(2,30) = 82.08; p < .001$). Negative affect, which taps into boredom, was higher in the easy level than in the moderate and the hard level ($F(2,30) = 12.26; p < .001$). The latter two did not differ and revealed relatively low experiences of negative affect. As for the experience of flow ($F(2,30) = 14.71; p < .001$), both the moderate and the hard level yielded relatively high levels of flow compared to the easy level. The moderate level and the hard level however did not differ significantly.

The iGEQ’s competence dimension ($F(2,30) = 73.89; p < .001$) also showed this reverse pattern, with the highest competence experienced in the easy level, followed by the moderate level and the hard level in which the experience of competence was low. As expected, positive affect ($F(2,30) = 29.56; p < .001$) was highest in the moderate level, slightly (although significantly) higher than in the easy level. The hard level elicited the lowest experience of positive affect. The experience of immersion ($F(2,30) = 5.65; p < .01$), on the other hand, only differed between the easy level and the moderate level, with more experience of immersion in the latter level. The hard level did not significantly differ from neither the easy nor the moderate level on self reported immersion. In sum, both the moderate and the hard level seemed to be equally engaging, but only in the moderate level this was experienced in a positive way.
Three out of four SAM scales showed significant effects of the difficulty manipulation. Pleasure was significantly lower in the hard level compared to the easy level and the moderate level, $F(2,30) = 11.26; p < .001$. The latter two did not differ significantly on pleasure experiences as measured by the SAM. The experience of arousal systematically increased from the easy level to the hard level with all levels differing significantly, $F(2,30) = 44.33; p < .001$. We observed the reverse pattern for the experience of dominance, showing a systematic decrease from the easy to the hard level, $F(2,30) = 68.5; p < .001$. Results do not reveal any significant differences for the experience of presence as measured by the SAM presence scale, $F(2,30) = 0.93; p = 0.41$.

Table 2: Comparison of the self report measures for the different experimental levels

<table>
<thead>
<tr>
<th>Measure</th>
<th>Easy (1)</th>
<th>Moderate (2)</th>
<th>Hard (3)</th>
<th>Mean</th>
<th>SD</th>
<th>Easy (1)</th>
<th>Moderate (2)</th>
<th>Hard (3)</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEQ Negative Affect***</td>
<td>1.39(2,3)</td>
<td>1.10</td>
<td>0.48(1)</td>
<td>0.77</td>
<td>0.61(1)</td>
<td>0.82</td>
<td>0.48(1,3)</td>
<td>0.85</td>
<td>0.77</td>
<td>0.61(1,2)</td>
</tr>
<tr>
<td>GEQ Tension***</td>
<td>0.44(2,3)</td>
<td>0.63</td>
<td>0.95(1,3)</td>
<td>0.85</td>
<td>2.20(1,2)</td>
<td>0.84</td>
<td>0.95(1,3)</td>
<td>0.85</td>
<td>2.20(1,2)</td>
<td>0.84</td>
</tr>
<tr>
<td>GEQ Flow***</td>
<td>1.87(2,3)</td>
<td>0.89</td>
<td>2.64(1)</td>
<td>0.91</td>
<td>2.66(1)</td>
<td>0.87</td>
<td>2.64(1,3)</td>
<td>0.91</td>
<td>2.66(1,2)</td>
<td>0.87</td>
</tr>
<tr>
<td>GEQ Challenge***</td>
<td>1.00(2,3)</td>
<td>0.78</td>
<td>1.98(1,3)</td>
<td>0.85</td>
<td>3.22(1,2)</td>
<td>0.68</td>
<td>1.98(1,3)</td>
<td>0.85</td>
<td>3.22(1,2)</td>
<td>0.68</td>
</tr>
<tr>
<td>GEQ Immersion*</td>
<td>1.28(2)</td>
<td>0.97</td>
<td>1.75(1)</td>
<td>0.81</td>
<td>1.55</td>
<td>0.93</td>
<td>1.75(1)</td>
<td>0.81</td>
<td>1.55</td>
<td>0.93</td>
</tr>
<tr>
<td>GEQ Competence***</td>
<td>2.86(2,3)</td>
<td>0.74</td>
<td>2.50(1,3)</td>
<td>0.89</td>
<td>0.97(1,2)</td>
<td>0.62</td>
<td>2.50(1,3)</td>
<td>0.89</td>
<td>0.97(1,2)</td>
<td>0.62</td>
</tr>
<tr>
<td>SAM Pleasure**</td>
<td>1.41(2)</td>
<td>1.34</td>
<td>1.72(3)</td>
<td>1.05</td>
<td>0.06(1,2)</td>
<td>1.78</td>
<td>1.72(3)</td>
<td>1.05</td>
<td>0.06(1,2)</td>
<td>1.78</td>
</tr>
<tr>
<td>SAM Arousal***</td>
<td>-</td>
<td>1.67</td>
<td>0.62(1,3)</td>
<td>1.77</td>
<td>1.75(1,2)</td>
<td>1.32</td>
<td>1.67</td>
<td>1.77</td>
<td>1.75(1,2)</td>
<td>1.32</td>
</tr>
<tr>
<td>SAM Dominance***</td>
<td>2.66(2,3)</td>
<td>1.66</td>
<td>1.47(1,3)</td>
<td>1.59</td>
<td>-1.41(1,2)</td>
<td>1.16</td>
<td>1.47(1,3)</td>
<td>1.59</td>
<td>-1.41(1,2)</td>
<td>1.16</td>
</tr>
<tr>
<td>SAM Presence</td>
<td>0.87</td>
<td>2.24</td>
<td>1.47</td>
<td>2.24</td>
<td>1.34</td>
<td>2.35</td>
<td>1.47</td>
<td>2.24</td>
<td>1.34</td>
<td>2.35</td>
</tr>
<tr>
<td>FlowGrid***</td>
<td>-</td>
<td>1.40</td>
<td>-0.50(1,3)</td>
<td>1.56</td>
<td>1.44(1,2)</td>
<td>1.04</td>
<td>-0.50(1,3)</td>
<td>1.56</td>
<td>1.44(1,2)</td>
<td>1.04</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001 (within analysis), N = 32

*1, 2, 3* differs significantly (pairwise comparisons, $p < .05$) from easy level*1*, moderate level*2*, hard level*3*.

Result show significant difficulty manipulation effects on the FlowGrid, $F(2,30) = 92.26, p < .001$. In line with expectations, the easy level scored lowest on the FlowGrid followed by the moderate level and the hard level (see Figure 8). Each of the levels differed significantly from the other. We further observed that the mean for the easy level not only was lowest on the FlowGrid, it was also negative, indicating that players felt their skills were not maximally challenged. Similarly, the mean score for the FlowGrid in the moderate level is close to the midpoint of the scale (i.e., 0), indicating the level’s challenge matched their skills. The hard level scored highest and positive on the FlowGrid, indicating that for most players the challenge outran their skills.
Additionally, the scores for the FlowGrid were explored as frequencies over the levels. In both Figure 8 and 9 a remarkably dip in the middle of the graph indicates that few participants reported that their skills exactly matched the challenge. In terms of flow theory (Csikszentmihalyi, 1975; 1990) there seems to be a markedly low number of instances in which participants experienced flow.

**Observational coding of player movement**

Participants’ movements were coded and counted using the video recordings. The first thing that should be noted is that players typically moved sparsely and that participants show substantial variation. Random movements in particular were scarce, but both forward and backward
movements appeared to become more frequent with increasing difficulty level (see Table 3). Repeated-measures ANOVAs were employed to test the differences between the three experimental levels. Results show that participants indeed exhibited significantly more forward movement ($F(2,9) = 7.06, p = 0.01$) and more backward movement ($F(2,9) = 4.91, p = .04$) in the hard level compared to the other two levels. The easy and the moderate level did not differ with respect to movement frequency. There were no significant differences in random movement between the three experimental game levels, $F(2,9) = 0.68, p = .53$.

### Table 3: Frequency of player movement per level

<table>
<thead>
<tr>
<th>Measure</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency Forwards Movement*</td>
<td>2.09(3)</td>
<td>3.27</td>
<td>2.27(3)</td>
<td>2.87</td>
<td>7.00(1,2)</td>
<td>6.91</td>
</tr>
<tr>
<td>Frequency Backward Movement*</td>
<td>4.36(3)</td>
<td>9.29</td>
<td>5.27(3)</td>
<td>9.38</td>
<td>10.91(1,2)</td>
<td>14.94</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001 (within analysis), N = 11

(1, 2, 3): differs significantly (pairwise comparisons, p < .05) from easy level(1), moderate level(2), hard level(3)

### Accelerometer

Besides observational coding, automatic movement data of the upper body was also collected with the accelerometer attached to participants’ backs. Instead of frequencies, this measure renders continuous data. We analyzed the effects of the three game levels on both the maximum and mean scores acquired. The results indicate that there is no difference in the maximum accelerometer data between the levels (see Table 4). The mean accelerometer value did however differ between the levels, $F(2,29) = 4.18, p = .025$. The mean value for the hard level proved to be highest and significantly different from both the moderate and the easy level. This implies that in the hard level participants, on average moved more strongly than in the other levels. While the mean accelerometer values were lowest for the moderate level, this difference proved not to be significantly different from the easy level.

### Sitting position

One further indicator of player movement was acquired via the sensors in the pressure sensitive chair. This indicator takes into account not only the movement of the upper part of the body, but rather the center of gravity of the body as a whole. Sitting position was analyzed using the range from the forward-backward position on the chair. Player movement followed a U-shaped curve, with highest scores for the hard level, followed by the easy level, and lowest for the moderate level (see Table 4). Differences between the levels were marginally significant, in particular between the high and moderate difficulty level, $F(2,29) = 3.26, p = .053$.

### Force measurements

Sensors on both mouse buttons picked up the force with which players made each mouse click. The force applied to the mouse was analyzed using both the maximum left mouse force and the mean left mouse peak force. Both indicators increased with the difficulty of the game level (see Table 4). The maximum left mouse force differed significantly between the levels, $F(2,29) = 9.71, p = .001$. Pairwise comparisons showed that the maximum force was significantly higher in the hard level than in the other two. It did not differ significantly between the easy and the moderate level. The mean left mouse peak force showed a similar pattern, although it was marginally significant $F(2,29) = 3.15, p = .058$. Again pairwise comparisons showed that the mean peak force was highest in the
hard level, differing significantly from the easy level. The value for the moderate level did not differ significantly from either level.

Table 4: Means of automatically captured measures per level

<table>
<thead>
<tr>
<th>Measure</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Accelerometer value</td>
<td>0.14</td>
<td>.18</td>
<td>0.13</td>
<td>.08</td>
<td>0.14</td>
<td>.07</td>
</tr>
<tr>
<td>Mean Accelerometer value*</td>
<td>0.0171(3)</td>
<td>.004</td>
<td>0.0165(3)</td>
<td>.004</td>
<td>0.0178(1,2)</td>
<td>.004</td>
</tr>
<tr>
<td>Range Sitting Position(marg*)</td>
<td>127.06</td>
<td>155.92</td>
<td>102.10</td>
<td>102.62</td>
<td>174.77</td>
<td>183.78</td>
</tr>
<tr>
<td>Maximum Left Mouse Force**</td>
<td>189.58(3)</td>
<td>145.62</td>
<td>247.00(3)</td>
<td>229.10</td>
<td>385.42(1,2)</td>
<td>284.12</td>
</tr>
<tr>
<td>Mean Left Mouse Peak Force (marg*)</td>
<td>39.92(1)</td>
<td>20.36</td>
<td>44.18</td>
<td>28.13</td>
<td>49.96(1)</td>
<td>30.27</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001 (within analysis), N = 31

(1, 2, 3): differs significantly (pairwise comparisons, p < .05) from easy level(1), moderate level(2), hard level(3)

Results part two: correlations between measures

This second part of the results section addresses convergent validity by exploring correlations between the various self-report measures and behavioral indicators. In this paragraph we first explore how the iGEQ measure relates to the other self report measures. This provides additional insight on the convergent validity of the iGEQ measures, in addition to what was reported in D3.3. We then explore correlations between the self-report measures and the observed and automatic measures.

For thoroughly exploring correlations between variables, sufficient variation is needed. However, since the experimental levels were explicitly created to induce a specific experience, variance within each level was only modest. For this reason we, restructured the data such that the different experimental levels were treated as separate cases, creating three rows of data for each participant. By exploring correlations across levels we created variation in the different measures enabling us to report reliable conclusions about how the different measures are related.

Self report measures

The GEQ and its modules (iGEQ, SPGQ, PGQ) have been developed specifically for probing player experience in digital gaming. In contrast, the SAM presents a measure for more general use. Yet obviously, they probe related, sometimes even similar experiential constructs. Various iGEQ dimensions show substantial correlations with the SAM measures in the expected directions, indicating good convergent validity (see Table 5).

iGEQ positive affect showed a strong positive correlation with SAM pleasure and was also moderately positively correlated with SAM dominance. Tension, which taps into the experience of frustration, showed moderate to strong correlations to three SAM measures: it was negatively correlated to SAM pleasure and SAM dominance and positively correlated to SAM arousal. iGEQ flow, probing mental absorption and engagement in the game, showed a strong positive correlation with SAM arousal and a modest one with SAM presence. iGEQ challenge showed a strong positive correlation with SAM arousal, and a strong negative relation with SAM dominance. iGEQ competence showed a particularly strong positive correlation with SAM dominance and moderate
ones with SAM pleasure (positive) and SAM arousal (negative). iGEQ immersion and iGEQ negative affect (boredom) showed only modest correlations.

The FlowGrid presents a new measure probing the player’s perceived balance between challenge and skills. Its numerical value is in fact an indicator of how much players felt the game overtaxed their skills. Correlations between this measure and iGEQ components are reported in Table 5. Particularly strong correlations were found between this measure and iGEQ challenge, tension, and competence (negative).

Table 5: Bivariate correlations between iGEQ and SAM measures and FlowGrid

<table>
<thead>
<tr>
<th></th>
<th>Positive affect</th>
<th>Negative</th>
<th>Tension</th>
<th>Flow</th>
<th>Challenge</th>
<th>Immersion</th>
<th>Competence</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAM Pleasure</td>
<td>0.699***</td>
<td>-0.243*</td>
<td>-0.491***</td>
<td>0.05</td>
<td>-0.231*</td>
<td>0.199*</td>
<td>0.447***</td>
</tr>
<tr>
<td>SAM Arousal</td>
<td>-0.120</td>
<td>-0.350***</td>
<td>0.572**</td>
<td>0.544***</td>
<td>0.673***</td>
<td>0.317**</td>
<td>-0.38***</td>
</tr>
<tr>
<td>SAM Dominance</td>
<td>0.432***</td>
<td>0.182</td>
<td>-0.664**</td>
<td>-1.54</td>
<td>-0.72***</td>
<td>-0.139</td>
<td>0.769***</td>
</tr>
<tr>
<td>SAM Presence</td>
<td>0.132</td>
<td>-0.258*</td>
<td>0.05</td>
<td>0.219*</td>
<td>0.099</td>
<td>0.266**</td>
<td>0.059</td>
</tr>
<tr>
<td>FlowGrid</td>
<td>-0.301**</td>
<td>-0.35***</td>
<td>0.684***</td>
<td>0.32**</td>
<td>0.818***</td>
<td>0.308**</td>
<td>-0.693***</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001, N = 96

Observations and self report measures

The observation based movement measures were compared with the iGEQ components (Table 6). The frequency of forwards movement was negatively correlated with iGEQ positive affect and showed moderate correlations with iGEQ competence (negative) and iGEQ immersion (positive). Additionally, it correlated negatively with SAM Dominance and moderately positively with the FlowGrid. Backwards showed a similar negative correlation with SAM Dominance. Random movement was moderately negatively correlated with presence (Table 7).

Table 6: Bivariate correlations between player movement and iGEQ dimensions

<table>
<thead>
<tr>
<th></th>
<th>Positive affect</th>
<th>Negative affect</th>
<th>Tension</th>
<th>Flow</th>
<th>Challenge</th>
<th>Immersion</th>
<th>Competence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Movement</td>
<td>-0.12</td>
<td>0.077</td>
<td>0.163</td>
<td>-0.217</td>
<td>0.187</td>
<td>-0.214</td>
<td>-0.096</td>
</tr>
<tr>
<td>Forwards Movement</td>
<td>-0.505**</td>
<td>-0.003</td>
<td>0.295</td>
<td>0.19</td>
<td>0.306</td>
<td>0.348*</td>
<td>-0.362*</td>
</tr>
<tr>
<td>Backwards</td>
<td>-0.238</td>
<td>-0.119</td>
<td>-0.039</td>
<td>0.083</td>
<td>0.224</td>
<td>0.31</td>
<td>-0.21</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001, N = 33

Table 7: Bivariate correlations of player movement with SAM dimensions and FlowGrid

<table>
<thead>
<tr>
<th></th>
<th>SAM Pleasure</th>
<th>SAM Arousal</th>
<th>SAM Dominance</th>
<th>SAM Presence</th>
<th>FlowGrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Movement</td>
<td>-0.148</td>
<td>0.120</td>
<td>-0.109</td>
<td>-0.346*</td>
<td>0.158</td>
</tr>
<tr>
<td>Forwards Movement</td>
<td>-0.182</td>
<td>0.298</td>
<td>-0.546**</td>
<td>0.073</td>
<td>0.407*</td>
</tr>
<tr>
<td>Backwards Movement</td>
<td>0.02</td>
<td>0.11</td>
<td>-0.476**</td>
<td>-0.181</td>
<td>0.255</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001, N = 33
Accelerometer and self report measures

In spite of the fact that the mean accelerometer score was sensitive to the difficulty level manipulation, no significant correlations emerged between the two accelerometer measures and any of the self report measures.

Sensor chair position and self report measures

We explored correlations between the data from the pressure sensitive chair and both self report measures and the observed movement data (Tables 8 & 9). Two modest but significant correlations emerged with iGEQ positive affect (negative) and iGEQ tension (positive). No significant correlations emerged with the SAM or the FlowGrid.

Force measurements: comparison with self report measures and observational coding

The force measurements for the left mouse button revealed significant correlations with the iGEQ, the SAM measures and the FlowGrid. The maximum left mouse force was modestly yet positively correlated with iGEQ tension and challenge, and negatively with iGEQ competence. In addition, it correlated negatively with the SAM dominance measure and positively with the FlowGrid score, though both statistics were modest. The mean left mouse peak force (an aggregate measure of the peak forces over the entire level) showed a small positive correlation with iGEQ immersion and the FlowGrid score.

Table 8: Bivariate correlations of Automatic measures with iGEQ dimension

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
<th>Tension</th>
<th>Flow</th>
<th>Challenge</th>
<th>Immersion</th>
<th>Competence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum</td>
<td>-0.158</td>
<td>-0.034</td>
<td>0.176</td>
<td>0.044</td>
<td>0.092</td>
<td>0.076</td>
<td>0.030</td>
</tr>
<tr>
<td>Mean Accelerometer</td>
<td>-0.042</td>
<td>0.069</td>
<td>0.159</td>
<td></td>
<td>-0.038</td>
<td>-0.132</td>
<td>-0.085</td>
</tr>
<tr>
<td>Range Sitting Position</td>
<td>-0.287**</td>
<td>0.073</td>
<td>0.208*</td>
<td></td>
<td>-0.008</td>
<td>-0.188</td>
<td>-0.159</td>
</tr>
<tr>
<td>Maximum Left Mouse Force</td>
<td>-0.068</td>
<td>-0.201</td>
<td>0.261*</td>
<td>0.066</td>
<td>0.284**</td>
<td>0.06</td>
<td>-0.343*</td>
</tr>
<tr>
<td>Mean Left Mouse Peak Force</td>
<td>-0.051</td>
<td>-0.191</td>
<td>0.035</td>
<td>0.031</td>
<td>0.150</td>
<td>0.229*</td>
<td>-0.181</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001, N = 93

Table 9: Correlations of Automatic measures with SAM dimension and FlowGrid

<table>
<thead>
<tr>
<th></th>
<th>SAM Pleasure</th>
<th>SAM Arousal</th>
<th>SAM Dominance</th>
<th>SAM Presence</th>
<th>FlowGrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Accelerometer value</td>
<td>-0.131</td>
<td>0.203</td>
<td>0.056</td>
<td>-0.026</td>
<td>0.076</td>
</tr>
<tr>
<td>Mean Accelerometer value</td>
<td>-0.062</td>
<td>0.008</td>
<td>0.055</td>
<td>0.01</td>
<td>-0.053</td>
</tr>
<tr>
<td>Range Sitting Position</td>
<td>-0.203</td>
<td>-0.009</td>
<td>-0.093</td>
<td>-0.14</td>
<td>-0.056</td>
</tr>
<tr>
<td>Maximum Left Mouse Force</td>
<td>-0.110</td>
<td>0.096</td>
<td>-0.244*</td>
<td>0.096</td>
<td>0.346**</td>
</tr>
<tr>
<td>Mean Left Mouse Peak Force</td>
<td>-0.038</td>
<td>-0.108</td>
<td>-0.198</td>
<td>0.089</td>
<td>0.213*</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001, N = 93
Discussion

The principal aim of this study was to explore behavioral indicators of boredom, flow, and frustration as in-game experiences and to provide insights into the construct validity of these behavioral indicators. Secondly, we also wanted to test convergent validity of the behavioral indicators with a set of self report measures, in particular, the iGEQ, the SAM and a newly developed measure, the FlowGrid.

The manipulation of difficulty level proved to be successful in eliciting experiences of boredom, enjoyment (flow) and frustration. Results from our manipulation check demonstrated that, indeed, the easy level was experienced as too easy, the moderate level as fairly well balanced, and the hard level as too difficult. This pattern was followed closely by the FlowGrid scores. Additionally, the self report measures (iGEQ and SAM) provide a similar but more detailed pattern.

Behavioral indicators: construct and convergent validity

Behavioral indicators of frustration turned out to be quite pronounced. Frequency of forward and backward movement, range in sitting position (maximum forward to maximum backward position), mean accelerometer value (indicating overall extremeness of movement), and maximum force exerted on the mouse were all highest during the hard level. For these measures the scores in the easy and the moderate level were statistically not distinguishable, although the patterns matched expectations.

Correlations of these behavioral indicators with the iGEQ, the SAM dimensions and the FlowGrid further provide insights in the relation between these indicators and the subjective experience of frustration. For instance, SAM Dominance was negatively correlated to maximum left mouse force, indicating that experiencing loss of dominance or control in the game is correlated to persons having a higher peak force during playing. Further, looking at correlations between the iGEQ dimension of tension and the behavioral measures across experimental levels, we found range in sitting position and maximum left mouse force to be positively correlated to tension. Additionally, maximum left mouse force was also positively correlated with FlowGrid, indicating that as people perceive the challenge to be higher than their skills, which according to the flow theory gives rise to frustration, the peak force during play of a game was higher as well.

Some caution should be taken for the interpretation of the findings, however. While the indicators for movement and force were highest in the hard level, they were not significantly correlated to all corresponding subjective indicators of frustration. Of course, the levels differed in more than only the experience of frustration which may also influence the observed behaviors. For instance, the pace of the game and speed of the zombies was higher in the hard level than the other levels. Future studies are needed to corroborate these findings and test further the relation between the behavioral indicators of frustration listed here and the subjective experience of frustration. The findings of CKIR do indeed warrant such caution. They found acceleration data (indicative of extremeness of movement) to be positively correlated with the GEQ measure of positive affect. Additionally they found playing mode (competition vs. collaboration) to influence acceleration data.

On a general level, our results show clear indications that more and extreme movement and more force (both absolute and aggregated over entire levels) can be used as indicators of frustration. In addition, the results leave open the potential that at least some of these measures may also be (inversely) indicative of flow and boredom. The findings reported in this chapter are based on analyses of measures aggregated over entire levels. On top of the current encouraging findings, we are currently analyzing these measures in relation to more detailed game events.
In sum then, our results bode well for the applicability of behavioral indicators to measuring player experiences in relation to digital games. From a methodological point of view, there are several advantages associated with employing such behavioral measures. First, they are relatively free from subjective bias, because they are generally not under player’s conscious control, nor do they require specific instructions from the experimenter (e.g., “please hit the button harder as you get more frustrated”) – they occur spontaneously. Secondly, when measured in an unobtrusive fashion, they do not disrupt the player experience. Third, they are time-continuous measures, that is, they are collected as the experience is unfolding, and are as such not reliant on memory or introspection on the part of the participant (unlike self-report measures). Finally, a number of these measures, such as a pressure-sensitive gamepad, could realistically be integrated with existing game technologies. This is a clear advantage when these measures are to be integrated in commercial games, where specialist peripheral hardware will only scarcely be adopted. As the findings from the current experiment shows, some measurements, e.g. maximum force, may indeed be candidates for automation.

Future research plan

Due to time constraints at the time of writing, we were unable to report results for all measures we included in this study in the current chapter. More specifically, three additional analyses deserve priority for our future analysis of this study. First we want to check how expertise level influences our results. It could be that experts are more in control during game play leading to less extreme expression of their frustration. As such, we would expect the relation between frustration and the behavioral indicators to be attenuated for expert players. Further, we want to check whether and how personality traits (BIS/BAS and trait aggression) influence our results. Finally, we are currently coding facial expressions participants showed during game play. In future analyses we aim to compare the instances of specific facial expressions between the different experimental game levels and correlate them to the self-report measures to check whether they are reliable and valid behavioral indicators of game experience.

2.5.2 Construct validity of the Game Experience Questionnaire

Introduction

During the first year of the FUGA project, the Game Experience Questionnaire (GEQ) was developed by TU/e. In the absence of a comprehensive and validated scale of player experience, this self-report measure was developed to serve as measure against which each of the FUGA partners could assess their new measures. Development of the items, scale construction and internal consistency tests of the GEQ and corroborative modules (the social presence in gaming questionnaire – SPGQ - and the post-game questionnaire - PGQ) based on a large sample of Dutch gamers was reported in D3.3. Since then, a second explorative large-scale survey has been performed among English and American gamers, resulting in the same factor structure. Factor structure and scale reliabilities in the other languages in the FUGA project will be explored when these data become available from partners. In the preceding paragraph we already reported the results for the iGEQ, a shorter version of the full GEQ, consisting of only two items per component.

4 The results of this study will be reported in a separate addendum to D3.3.

5 Explorations of the Finnish data demonstrated excellent convergence with the Dutch and English scales, and good to excellent scale reliabilities.
The current section tests the construct validity of the Game Experience Questionnaire (GEQ) by comparing scores on the seven components for three difficulty levels of Half-Life 2, in a between Groups design. The experiment was in fact part of the experiment reported in Section 2.5.1, which tested the construct validity of behavioral indicators. The participant sample, setting, manipulations and procedure are thus identical to that reported in the previous section and only briefly described here. For more detailed descriptions we refer to the previous section.

**Method**

**Design**

In contrast to the analyses reported in section 2.5.1, the current analysis followed a between groups design, as only the third (and last) game session of the full experiment was assessed using the full GEQ. To minimize the burden on the players, the first and second session only employed the (shorter) in-game version of the GEQ. In the present analysis, only the data of the final sessions was used.

**Participants**

Thirty-two gamers participated in the experiment. Five of them were female, and their ages ranged between 17 and 46 (\(M_{age} = 22.47\) years, \(SD = 5.49\) years). All participants at least occasionally played FPS games, nineteen participants were more frequent players. Recruitment is reported in detail in Section 2.5.1. Participants received 10\(\text{€}\) for their time. Age and gender were distributed evenly over conditions, FPS experience appeared relatively high in the easy-level condition (see Table 10).

<table>
<thead>
<tr>
<th></th>
<th>Easy level</th>
<th>Moderate level</th>
<th>Hard level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Male</td>
<td>9</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>23.4 (7.7)</td>
<td>23.3 (5.1)</td>
<td>22.5 (2.4)</td>
</tr>
<tr>
<td>FPS play frequency</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occasional</td>
<td>2</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Frequent</td>
<td>9</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

**Experimental setting and procedure**

The experiment was conducted in the Game Experience Lab at Eindhoven University of Technology. Three modified levels of the FPS game Half Life 2 (HL2), created by BTH/HGO, were used to induce three difficulty levels: easy, moderately difficult, and hard. The different levels were played on a Dell XPS PC equipped with a quad core processor and a NVIDIA GeForce 8800GTX 768Mb graphics card, connected to a 20" TFT-screen. Various recordings were made and are described in Section 2.5.1. However, the present section only deals with the data collected with full GEQ.

---

6 These are the same participants as reported in Section 2.6.1.
Upon entering the lab, participants were welcomed by the experiment leaders. In order to comply with the ethical requirements, participants were first probed about their gaming habits in order to signal any potentially problematic behaviors. None of them exhibited indications of problematic game behavior. Experiment leaders then gave a brief overview of the flow-line of the experiment, and explained that the experiment involved the measurement of psychophysiological indicators, video and audio recordings. Following this, participants signed the consent form and were fitted with the psychophysiological sensors.

Participants played three levels of the FPS game Half Life 2, each for ten minutes, or eight minutes for the easy level. The order of the levels was counterbalanced. After each play session, participants rated their experiences using the In-Game Experience Questionnaire, and additional self-report measures (see Section 2.5.1). Only after the third and last level, participants completed the full Game Experience Questionnaire (35 items).

Results

This section presents the effects of the difficulty level manipulation on the seven components of the full GEQ. As the manipulation was meant to induce increasing difficulty levels, particularly strong effects were expected on challenge. In addition, the levels were created such that the easy level would be rather boring (indicated by the GEQ negative affect scale), and the hard level would be too difficult and induce frustration (indicated by the GEQ tension scale). The moderate level would present the player with an optimally challenging, i.e. just manageable, difficulty level. GEQ flow was therefore hypothesized to show a curvilinear relation with difficulty, with a maximum for the moderate difficulty level. However, first internal consistencies of the scales are computed.

All internal consistencies were acceptable to excellent, with Cronbach’s alpha varying between .71 and .87 (see Table 11). Variables were computed as the numeric average of the scores on each scale’s items. The mean scores on all seven GEQ components are reported in Table 11 and visualized in Figure 10.

Table 11: Scale reliabilities and descriptives of GEQ components

<table>
<thead>
<tr>
<th>GEQ component</th>
<th>No of items</th>
<th>Cronbach’s alpha</th>
<th>Easy level</th>
<th>Moderate level</th>
<th>Hard level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immersion</td>
<td>6</td>
<td>.71</td>
<td>1.0 (.54)</td>
<td>1.8 (.63)</td>
<td>1.8 (.70)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(2,3)</td>
<td>(1)</td>
<td>(1)</td>
</tr>
<tr>
<td>Competence</td>
<td>5</td>
<td>.87</td>
<td>2.7 (.65)</td>
<td>2.5 (.60)</td>
<td>1.1 (.57)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(3)</td>
<td>(3)</td>
<td>(1,2)</td>
</tr>
<tr>
<td>Negative affect</td>
<td>5</td>
<td>.71</td>
<td>1.4 (.66)</td>
<td>0.6 (.45)</td>
<td>0.7 (.42)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(2,3)</td>
<td>(1)</td>
<td>(1)</td>
</tr>
<tr>
<td>Flow</td>
<td>5</td>
<td>.87</td>
<td>1.9 (.70)</td>
<td>3.0 (.70)</td>
<td>2.7 (.72)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(2,3)</td>
<td>(3)</td>
<td>(1)</td>
</tr>
<tr>
<td>Tension</td>
<td>5</td>
<td>.87</td>
<td>0.5 (.80)</td>
<td>1.0 (.59)</td>
<td>2.1 (.54)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(1)</td>
<td>(3)</td>
<td>(1,2)</td>
</tr>
<tr>
<td>Positive affect</td>
<td>5</td>
<td>.76</td>
<td>2.1 (.51)</td>
<td>2.5 (.87)</td>
<td>2.2 (.56)</td>
</tr>
<tr>
<td>Challenge</td>
<td>5</td>
<td>.83</td>
<td>1.1 (.46)</td>
<td>1.7 (.75)</td>
<td>3.2 (.42)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(2,3)</td>
<td>(1,3)</td>
<td>(1,2)</td>
</tr>
</tbody>
</table>

Note: N = 32, scale values range between 0 and 4.

(1, 2, 3): differs significantly (pairwise comparisons, p < .05) from easy level (1), moderate level (2), hard level (3).

In spite of the modest number of participants in each condition, a very clear picture emerged. Effects of the difficulty level manipulation were subsequently tested with analyses of variance (ANOVAs), with the component scores as dependent variable, and difficulty level as the between-groups variable.
As expected, the strongest effect appeared on the GEQ challenge component. Reported challenge increased significantly with each subsequent difficulty level $F(2,29) = 38.99, p<.001$, $\eta^2=.73$.

A similarly strong and positive effect also emerged for GEQ tension. Tension increased significantly with difficulty level $F(2,29) = 17.07, p<.001$, $\eta^2=.54$, but only the contrast between the easy and moderate level on the one hand and hard level on the other reached significance $p<.001$.

Competence showed opposite effects of the manipulation, i.e., decreasing perceived competence with increasing difficulty levels $F(2,29) = 23.74, p<.001$, $\eta^2=.62$.

Figure 9 does indeed show the hypothesized inverse U-shape for GEQ flow, with the highest score for the moderate difficulty level. The ANOVA showed a significant effect of the manipulation, $F(2,29)=6.79, p=.004$, $\eta^2=.32$ but the curvilinear effect was not completely supported, as the contrast between the moderate and hard level did not reach significance.

Immersion increased with difficulty level $F(2,29) = 5.61, p=.009$, $\eta^2=.28$, but only the contrast between the easy level on the hand and moderate and high levels on the other reached significance.

Negative affect (boredom) showed effects opposite to that of immersion, $F(2,29) = 8.66, p<.001$, $\eta^2=.37$. Boredom decreased significantly between the easy and moderate difficulty level, but not between moderate and hard difficulty.

Lastly, positive affect was not affected by the difficulty level manipulations, $F<1$, NS.

![GEQ components for difficulty levels](image)

**Figure 10: Mean values of GEQ components for Difficulty Level**

**Discussion**

The current section explored the results of the Game Experience Questionnaire (GEQ) in light of the experimental manipulation, providing a further basis for the construct validity of the GEQ, beyond the validation results that were already reported in D3.3.
In spite of the modest number of participants in each condition, due to the between groups design employed here, a very clear picture emerged. The strongest effect appeared on the GEQ challenge component. The pattern of incrementally increasing perceived challenge perfectly reflected the increasing difficulty level of the three experimental Half Life 2 levels.

In addition, the levels were created such that the easy level would be rather boring, and the hard level would be too difficult and induce frustration. Again, the GEQ components probing these experiences (negative affect and tension respectively) reflected these manipulations exactly. GEQ negative affect was significantly higher in the easy level than in the other two; reported tension was significantly higher for players in the hard level than in the easier ones.

Ideally, the moderate level presented players with an optimally challenging, that is, just manageable, difficulty level. GEQ flow was therefore hypothesized to show a curvilinear relation with difficulty, with a maximum for this level. Although the pattern of results did suggest an optimum score in the intermediate game level, statistical analyses indicated that reported flow was equally high for players in the moderate and hard level. The absence of a significant effect between moderate and high difficulty could imply that perhaps the scale does not probe flow, or, alternatively, that the moderate level did not induce flow better than did the hard level. Both more and less experienced gamers participated in the present study, and indeed some of the avid FPS players indicated that in fact the hard level provided the perfect balance between challenge and skill that is considered prerequisite for a flow experience.

Although not directly targeted in the current experiment, some additional components of the GEQ showed significant effects of the manipulation. GEQ competence showed strong effects of difficulty: the more difficult the level was, the less competent players felt, in particular between the moderate and hard level. Additionally, immersion showed a modest increase between the easy level and the higher ones.

### 2.5.3 Conclusion

We conclude that the construct validity of three scales of the GEQ has been established: challenge, negative affect and tension all responded well to the three difficulty levels of the game and such demonstrated that they were valid and sensitive to experimental manipulations. The fourth relevant measure, flow, showed partial sensitivity and validity: it was higher for moderate than for easy levels of the game, yet showed no differences between the moderate and high difficulty level. Additional research will need to confirm whether the scale indeed can sensitively and validly pick up on flow as distinct from other experiences such as challenge, immersion and frustration. In the current experiment however, these scales did behave differently, indicating that in fact the GEQ flow dimension does probe a unique experiential construct.

A valid self-report measure is of critical importance to the FUGA project as a whole, as it is an essential metric to allow comparison of results between partner sites, and between different experiments and measures. Thus, the GEQ can be regarded to act as a kind of 'ground truth' measure of player experience within FUGA.
3 GENERAL DISCUSSION

In WP5, the FUGA partners have carried out a number of studies to construct validate the different potential measures of game experience. This work has involved the examination of how the measures differentiate the different experimental conditions that, on the basis of theoretical grounds, they should be able to differentiate. This work has also involved the examination of the relationships of the different measures with self-reported game experience. In this connection, the Game Experience Questionnaire (GEQ) developed in the FUGA project has played an important role. The construct validity of the GEQ scales (in particular, challenge, negative affect and tension) has also received support. The studies examining the potential measures of identification with game characters demonstrated that this construct can be measured by a mix of Lexical Decision Task, Implicit Association test, and Self-Description task; the relationship of these measures with overall or general game enjoyment is in need of further investigation, however. The construct validity studies on implicit self-esteem of video game players did not achieve a scientifically satisfying outcome.

Of the psychophysiological measures examined in FUGA, facial EMG appears to be the most promising measure of game experience. It is of note, however, that Facial EMG provides information on the valence (displeasure-pleasure) or positive affect and negative affect dimensions of game experience, but is not associated with other game experience dimensions. The studies also showed that both phasic and tonic facial EMG responses were associated with pleasure experiences during game playing. An important finding was that facial EMG was found to be associated with emotional experiences also during social gaming (in the presence of another player). Frontal EEG asymmetry may also provide to be a fruitful measure of game experience, although its associations with self-reported game experience were partially unexpected. Additional work is needed to examine its usefulness.

In the fMRI studies, specific brain activation patterns were found for several Flow factors, and the reward system, the cerebro-thalamo-cortical motor lip, the visual system, and dorsal parts of the prefrontal cortex were suggested to be important in this connection. Structures belonging to the reward system, including midbrain dopamine regions, the cingulate cortex, and striatum, were found to detect the difference between expected reward and the received outcome. An association was found between positive enjoyment ratings and insular and amygdala activity during goal-oriented playing phases. However, associations emerged more frequently with disenjoyment measures, which may indicate that the neurophysiological method preferentially detects frustration and negatively experienced events. The method reveals components of game experience and enjoyment as represented in the flow concept, but the networks do not indicate a convergence to a common concept.

The investigation of behavioral indicators of boredom, flow, and frustration shows that more and extreme movement and more force (both absolute and aggregated over entire levels) can be applied to index frustration. Some of these measures may also be (inversely) indicative of flow and boredom, but this issue requires additional research.

Overall, the construct validity studies of FUGA have advanced game enjoyment theory and provided valuable information on the measures that can be used to index game experience. In WP6, the FUGA partners will examine the (test-retest) reliability of the game experience measures found to be valid in WP5. The project partners will also carry out additional analyses to examine how personality traits (BIS/BAS and trait aggression) influence our results, for example.
4 REFERENCES

4.1 References for section 2.1 (CKIR)


### 4.2 References for section 2.2 (HMTH)


FUGA the fun of gaming: measuring the human experience of media enjoyment


4.3 References for section 2.3 (UKA)


FUGA the fun of gaming: measuring the human experience of media enjoyment  Report D5.1


### 4.4 References for section 2.4 (HGO/BTH)


4.5 References for section 2.5 (TUE)


FUGA the fun of gaming: measuring the human experience of media enjoyment Report D5.1


