The Effects of Framing in Gamification: A Study of Failure

Florian Brühlmann

Master's thesis

Institute of Psychology

University of Basel

October 2015

Thesis supervisors:

Elisa Mekler, M.Sc.

Prof. Dr. Klaus Opwis

Author Note

Florian Brühlmann

Delfterstrasse 26, 5004 Aarau

florian.bruehlmann@unibas.ch

Abstract

In human-computer interaction research, interest in the unique experiences that digital games offer is increasing. One characteristic of digital games is the failure of players, which is arguably the core of what makes games interesting and fun. Gamification seems to transfer some of the motivational effects of digital games in non-entertainment contexts by reframing an activity. Hence, this thesis sheds light on the psychological effects of failing in different contexts to gain a better understanding of the effects of games and gamified applications. Two mixed-design experiments ($N_1 = 148$, $N_2 = 63$) were conducted. The first examined the effect of failing in a task, game, test, and task with meaning condition. Participants in this experiment were compensated for participation. In the second experiment, the framing conditions task, game, and test were examined, and no compensation was given for participation. Results showed that framing an activity as a game did not reduce the negative effect of failing on intrinsic motivation. However, having additional information about the task seemed to buffer the negative effect of failing on positive affect in the first experiment. Furthermore, perceived value and particularly change in autonomy was found to predict playing of additional rounds after the study.

The Effects of Framing in Gamification: A Study of Failure

Introduction

In recent years, the focus of human-computer interaction (HCI) research shifted from usability aspects of digital products to a holistic view of the user experience. Thus, HCI research pays increasing attention to hedonic aspects of interactions such as affective, emotional, motivational, and aesthetic aspects of the user experience (Hassenzahl & Tractinsky, 2006). Part of this shift is the increasing interest of practitioners and researchers in the unique experiences that digital games can offer.

The idea that video game design can inspire the design of other digital products has been considered at least since the 1980s (e.g., Malone, 1981). However, designing digital games is very different from designing other applications. For example, if a user wants to set the font size in a word processor to 11 points, the user should be able to do this intuitively, quickly and without any obstacles. But a game that allows the user to reach the goal as efficiently as possible is not an interesting game for the simple reason that it lacks challenge (Malone, 1981). Games incorporate obstacles that the player has to overcome. Hence, a game designer designs deliberately — but not exclusively — for negative experiences. Players seek out games that very likely include the experience of failure, although people generally try to avoid failure and dislike losing (Schmierbach, Chung, Wu, & Kim, 2014).

In the book *The Art of Failure* (Juul, 2013), game designer and theorist Jesper Juul argues that this *paradox of failure* is the core of what makes games interesting. The paradox is defined thus: 1) players generally avoid failure, 2) players experience failure when playing games and 3) players seek out games, even though they will experience something that they normally avoid. Games make players feel inadequate but motivate them to play more to repair this inadequacy. According to Juul (2013), this is a central aspect of the enjoyment of games. Overcoming an obstacle in the progress of a game provides clear proof that a player has improved. Juul (2013) argues that examining the nature of failure in games is important because it tells us why games are enjoyable and also what effects games might have when applied outside of entertainment.

The discussion about possible effects of games on the lives of players is still ongoing.

The vast amount of time people spend playing digital games, — 73 min per day among 8- to 18-year-olds in the US (Kaiser Family Foundation, 2010) — raised concerns about possible negative effects of playing (often violent) digital games such as increased levels of aggression, social isolation, and reduced prosocial behavior (Anderson & Bushman, 2001). Other researchers have argued that playing digital games might have no practically significant negative effects (Valadez & Ferguson, 2012) or may even have positive effects on players (Bavelier et al., 2011; Ferguson, 2007). The fact that people can invest many hours in a game attracted the interest of motivation research (Ryan, Rigby, & Przybylski, 2006). Besides foundational research on games and motivation, one central area of investigation is transferring the "motivational pull" of digital games to other domains such as education and health (Rigby & Ryan, 2011).

The idea that game design principles can be applied outside of a game context resurfaced in HCI in 2010 with a new term: *gamification*. The term originates from the marketing industry, in which different types of rewards such as flight miles or discounts for returning customers have been used for a long time (Deterding, Dixon, Khaled, & Nacke, 2011). Gamification is most commonly defined as "the use of game design elements in non-game contexts" (Deterding et al., 2011). This definition indicates that gamification is not the development of a full-fledged video game, but the use of some game design elements. Gamification aims to transfer what makes digital games motivating to other contexts to make, for instance, learning, physical exercise or tagging images more interesting and to keep users motivated for a longer period of time. Using one or more game design elements in a context in which entertainment is not the main goal can be described as gamification. Implementations of gamification design elements can come in many different forms such as badges in Codeacademy, a website that uses game-like elements to teach users how to program computers, or gamified to-do lists with role-playing game elements such as EpicWin, a smartphone application that enriches the mundanity of crossing of tasks on a to-do list with experience points, avatars and story elements. What they have in common is that they apply game design elements in contexts such as learning or organization that are not pure entertainment contexts. It is, however, still debated when and how gamification should be

applied to create meaningful experiences (Nicholson, 2012).

The effectiveness of game design patterns such as points, levels and leaderboards have been studied in various domains (refer to Hamari, Koivisto, & Sarsa, 2014 and Seaborn & Fels, 2015 for an overview). But to date, little is known about the possible psychological mechanisms driving user behavior (Mekler, Brühlmann, Tuch, & Opwis, in press) and it is still unclear under what conditions gamification might work (Hamari et al., 2014). This is partly due to lack of the proper thoroughness and low generalizability of many studies of gamification. Another shortcoming is the rare combination of behavioral and psychological measures in studies with experimental design (Hamari et al., 2014). However, Lieberoth (2015) showed that introducing a game-frame to an activity can shape users' attitudes and motivation as well as behavior and increase performance in a brainstorming task. This suggests that there is some sort of psychological effect if people perceive an activity as a *game* compared to a regular *task*. One possible explanation might be that a game environment is associated with lack of consequences for poor performance, creating a secure area in which failing is allowed or even expected. Hence, in such an environment, failing is not expected to have as negative effects on intrinsic motivation as in a non-game environment.

The aim of this thesis is to study the effects of failing on motivation and behavior in a task, game, test, and task with meaning context from a self-determination theory perspective. This thesis contributes to the research by using theory of human motivation to empirically study effects of framing in an experimental design while assessing psychological and behavioral outcomes.

Theoretical background

In this section, self-determination theory, the role of context and the perception of failing will be discussed. Self-determination theory (SDT) has a rich body of research concerning the effects of external events on intrinsic and extrinsic motivation (Cerasoli, Nicklin, & Ford, 2014). The majority of research in gamification draws from this theory (Seaborn & Fels, 2015) to predict and explain the effects of game design elements on motivation since many of the common game design elements applied in gamified services use

some form of reward. Additionally, SDT has been applied in the study of motivational effects of video games (e.g., Ryan et al., 2006).

The perspective of self-determination theory

Self-determination theory originates from humanistic psychology and was developed over the last 40 years by Ryan and Deci (2000). The theory is based on three basic human needs as a source of personal growth and psychological well-being (Ryan & Deci, 2000): 1) autonomy, which is the ability to have as much freedom about one's actions and decisions as possible; 2) competence, the perception that one's own actions are the cause of desired consequences in one's environment; and 3) relatedness, the desire to interact and be connected to others. The urge to fulfill these three basic human needs is regarded as the source of intrinsic motivation (Ryan & Deci, 2000). In organismic integration theory, a sub-theory of SDT, several types of motivation can be distinguished, the most important being amotivation (the lack of motivation); extrinsic motivation, motivation regulated by external events; and intrinsic motivation, which is regulated internally. Intrinsic motivation is associated with a wide range of positive effects on well-being, learning and personal growth (Ryan & Deci, 2000). The regulation of motivation can to some extent shift from externally to more internally regulated motivation through integration of the activity in one's self-concept if it has personal value and meaning (Deci & Ryan, 2004).

Cognitive evaluation theory, another sub-theory of SDT, postulates that external events such as rewards or negative feedback can influence intrinsic motivation depending on a person's perception of how these events influence his or her competence and self-determination (Deci & Ryan, 2004). Events supporting competence can increase intrinsic motivation if these events are perceived as self-determined. Extrinsic incentives have been found to increase performance but also undermine intrinsic motivation and therefore decrease subsequent performance when the incentive is no longer present. This *undermining effect* has been documented extensively in self-determination theory research (see Deci, Koestner, & Ryan, 1999, for a review).

The use of game design elements such as points, badges, levels and leaderboards, which

are arguably the most common elements applied in gamification (Mekler, Brühlmann, Opwis, & Tuch, 2013b), can be viewed from a SDT perspective as forms of performance or completion-contingent rewards. The undermining effect is a major concern of researchers in the field of gamification (Brühlmann, 2013; Hanus & Fox, 2015; Mekler et al., 2013b; Thom, Millen, & DiMicco, 2012), but to date, there is not enough empirical evidence for the existence of this effect in gamification. Therefore, it is still inconclusive how these gamification elements affect intrinsic motivation. They might be perceived as both informational and controlling, depending on the user and the context (Mekler et al., in press).

The role of context in the perception of feedback

Lieberoth (2015) examined the effect of different degrees of gamification on intrinsic motivation and behavior. In what he called a *shallow gamification* condition, 22 participants were given a game board without any additional game mechanics. The players just progressed along the board as they discussed various business-related questions in groups of 6. In the so-called *deep gamification* condition, 25 participants were provided a game board and introduced to some game mechanics. In particular, they progressed in spaces on the board based on the rating of their discussion. In the control condition, 23 participants discussed the questions and were not given any game-related instructions. It was found that participants in the gamification conditions reported increased intrinsic motivation. Lieberoth (2015) concluded that applying a game-frame can alter perception and thus make a task more engaging and create a lighter mind-set. However, this requires further research because the sample size was small with 3 to 4 groups per condition, framing was confounded with a game artifact (the board), and the behavioral data were limited.

Burgers, Eden, van Engelenburg, and Buningh (2015) examined the effect of feedback valence (positive vs. negative) and feedback type (descriptive, comparative, evaluative) on need satisfaction, motivation and intention to play a brain-training game again. Results showed that mere informative feedback did not have significant effects, while evaluative feedback increased willingness to immediately play again and comparative feedback intention to play again in a week. Negative feedback decreased competence but increased willingness

for immediate replay. Positive feedback fostered long-term motivation and long-term willingness to play again. However, the context of this study was a brain-training game, a very specific type of game with a possible benefit outside of the game environment such as increased cognitive abilities. In their study, negative feedback increased immediate replay while reducing self-reported intrinsic motivation, which is, according to Deci and Ryan (2004), a pattern describing introjected regulation, a more internally regulated type of extrinsic motivation. The question remains: What would have happened if the game was not introduced as a game with a purpose outside of entertainment? If people do not feel as pressured as they do when playing a brain-training game, how do they respond motivationally and behaviorally to failing?

Reeve and Deci (1996) examined the role of interpersonal context in the effects of winning and losing. They found that in highly pressuring contexts, any reward or event is likely to have a negative effect on intrinsic motivation. In non-controlling environments, the same events might have a less negative effect on intrinsic motivation. In a more recent study, Vansteenkiste and Deci (2003) showed that winners were more intrinsically motivated than losers, but losers who were given a normative standard and received positive feedback for meeting the standard were less negatively affected by losing a reward in a competition. This suggests that although failing or losing has negative effects on intrinsic motivation, the type of feedback and the interpersonal context (pressuring vs. non-pressuring) can absorb some of the negative effects.

These examples show that the context and perception of failure are important, because if the informational aspect of feedback is more salient than the controlling aspect, negative feedback should have a less severe impact on intrinsic motivation (Deci & Ryan, 2004). In terms of behavior, the effect of failing depends on what people can expect from engaging in the activity again. If they can expect to improve their self-image (as in the case of brain-training game), increased engaging in the activity immediately after seems likely. For truly intrinsic regulation of behavior, meaning that engaging in the activity is desired because the activity itself is enjoyable, it is expected that negative feedback reduces the subsequent behavior because it “takes out the joy” of doing the task. However, the strength of this

reduction of intrinsic motivation depends on the relative salience of informational and controlling aspects of the negative feedback (Deci & Ryan, 2004). If people perceive negative feedback as helpful in supporting their learning and progress, instead of reprimanding and pointing to one's weaknesses, it is expected to have less negative consequences. In the game context, feedback in general and especially negative feedback might be perceived as more informational and thus might reduce autonomy less strongly. As a consequence, negative feedback might be less threatening to motivation and long-term behavior.

Failing and uncertainty in games

In the book *The Art of Failure* (Juul, 2013), Jesper Juul describes the *paradox of failure* in games, which means that people generally avoid failure but experience failure when playing games and still seek out games. A possible explanation for this behavior might be that failing in games is not painful or very unpleasant. But while this might be true in some cases, most players try to avoid failing and are upset when they fail (Juul, 2013). Another explanation explored by Juul (2013) is that failing in games has low costs, since it bears no consequences for the player outside of the game and offers plausible deniability. Players could attribute failure to a lack of skills, but they might also attribute it to the unfair game, bad luck or insufficient time invested. These attributions are seldom accepted in non-game contexts (Juul, 2013).

Players prefer games that are challenging and include the possibility of failure, but they also do not like to fail too much (Juul, 2013; Schmierbach et al., 2014). Thus, a game should neither be too easy or too hard; it should have an optimal challenge for the skill of the player. This optimal balance is part of the requirements to allow players to reach the psychological state of flow (the autotelic experience; Csikszentmihalyi, 1990) as described by Chen (2007). An extended explanation is offered by Abuhamdeh, Csikszentmihalyi, and Jalal (2015), who found that outcome uncertainty increases suspense, which in turn increases the enjoyment of a game. Players who experienced high outcome uncertainty also preferred this game over a different game with lower outcome uncertainty. This means that the possibility of failure increases suspense, which is associated with enjoyment in games and possibly with increased

levels of intrinsic motivation. Still, little is known about the reasons people continue playing a game in which they fail frequently.

In this thesis, it is hypothesized that failing within games has different motivational consequences than outside of games. A game context might offer a secure area in which actions are expected to be relatively free from outside consequences and therefore are perceived as less controlling and reduce the negative impact of failing on intrinsic motivation and behavior.

Research questions

The aim of this thesis is to examine whether framing a task as a game can reduce the negative impact of failing on intrinsic motivation. This was investigated in two experiments: In the first experiment participants received a monetary reward upon completion of the study, resulting in a motivation to participate that was tied to extrinsic rewards. Participants in the second experiment received no compensation for participating and thus were expected to partake for intrinsic reasons only.

The objective was to replicate the framing effect found by Lieberoth (2015) and extend this study with the findings of Burgers et al. (2015), who found positive effects of negative feedback on willingness to play the game again immediately in an brain-training game context. In this thesis, instead of self-reported willingness to play again actual behavior of participants was recorded. The effects of failing were examined in four different contexts: 1) task (serves as a control condition), 2) game, 3) test of cognitive abilities, and 4) task with meaning condition (similar to many gamified applications; was earlier found to increase performance in Chandler & Kapelner, 2013). Refer to the method section of Experiment 1 for details.

It is hypothesized that in a game context participants perceive failing (negative feedback) as less controlling than in the other conditions, thus diminishing perceived autonomy and consequently intrinsic motivation less strongly. Furthermore, it is expected that framing an activity as a test of cognitive abilities increases pressure and creates a controlling situation, thereby affecting intrinsic motivation more severely compared to a control condition when failing. See Table 1 for an overview of all research questions.

Table 1

Overview of the six research questions investigated in this thesis.

Nr.	Research Question
1	What are the effects of the four different conditions on the main dependent variables? (change in autonomy, competence, intrinsic motivation as well as change in affective state)
2	What are the effects of the four different conditions on the secondary dependent variables? (value, importance, tension)
3	What are the effects of the four different conditions on the score of the participants?
4	What are the effects of the four different conditions on the behavioral measures? (number of extra rounds played, additional time spent)
5	Which variables best predict whether participants play additional rounds?
6	What differences can be observed between participants in the two experiments? (compensation vs. no compensation)

Experiment 1 (with compensation)

To investigate the effects of framing on need satisfaction, motivation, and affective response when failing in an activity, an online experiment with a mixed design was implemented. The between-subjects variable was type of framing (task vs. game vs. test vs. task with meaning) and the within-subjects variable was time (pre and post experiment). The dependent motivational and affective variables were measured before and after the experiment. Participants were recruited on a crowdsourcing platform and paid 65 US cents for completion of the study.

Method

To frame an activity believably as a test of cognitive abilities, as a task, a task with meaning, and as a game, it was decided to use a relatively abstract game that is reasonably hard to master: *Flood It*[1] is a puzzle game in which a player starts on the top left corner of board which is filled with squares of different colors. The aim is to "flood" the board with one color in a certain number of steps (see Figure 1). It was decided to stick relatively closely to the original games rules, hence a maximum of 25 steps was allowed to fill the whole board. The number of allowed steps increased by 1 after losing a round to make it possible to at least win one or two rounds of the game. The aim was to keep to player motivated but still have the majority of users lose more than half of the rounds. All participants were asked to complete at least five rounds. An adapted version of the original Flood It was developed with HTML and Javascript by the author.

Procedure. Once participants commenced with the survey, they were randomly assigned to one of the four conditions. After a brief introduction, participants were led to a trial page, on which they could complete three rounds of Flood It. Following this page, a first assessment of intrinsic motivation, need satisfaction (autonomy and competence), and affect was applied to see if these variables were distributed comparatively across all conditions and to have a pre-framing measurement point. This was necessary to see how strongly intrinsic motivation, need satisfaction, and affect decreased after failing in Flood It. In a next step, the

[1] `https://bruehlmann.io/projects/floodit/start.html`

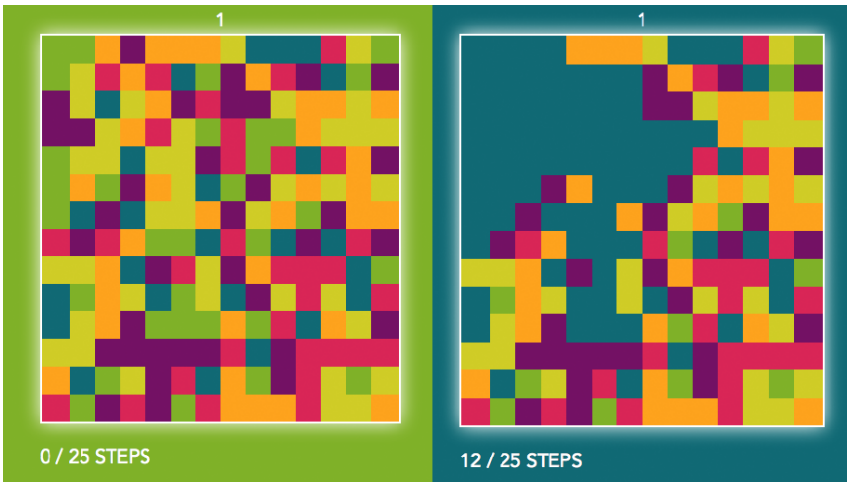

Figure 1. In Flood It, the goal is to flood the whole board with one color within a limit of steps.

framing was introduced (see Table 2). Subsequently, participants started with the activity and completed at least 5 rounds of Flood It. After each round explicit feedback was given ("You failed!" or "Well done!") and after 5 or more rounds participants had the option to return to the survey (see Figure 2). In the last part of the survey a second assessment of intrinsic motivation, need satisfaction, and affect was conducted and participants were asked to rate the activity on the value, importance, and tension scales.

Dependent variables. Intrinsic motivation was measured with the subscale interest/enjoyment of the intrinsic motivation inventory (Ryan, 1982). Perceived autonomy, competence as well as the variables perceived value ("I believe doing this activity could be beneficial to me"), importance ("It was important to me to do well at this task"), and tension ("I felt very tense while doing this activity") were also subscales of the same validated instrument. Participants rated the statements of the intrinsic motivation inventory on a scale from 1 ("not at all true") over 5 ("somewhat true") to 7 ("very true"). To measure affective state, the short form of the positive and negative affect schedule developed by Thompson (2007) was applied. The items on positive and negative affect were rated on a scale from 1 ("Not at all") to 5 ("Extremely"). All scales showed good internal consistency with Cronbach's α ranging from .82 to .93, except tension (.72; acceptable).

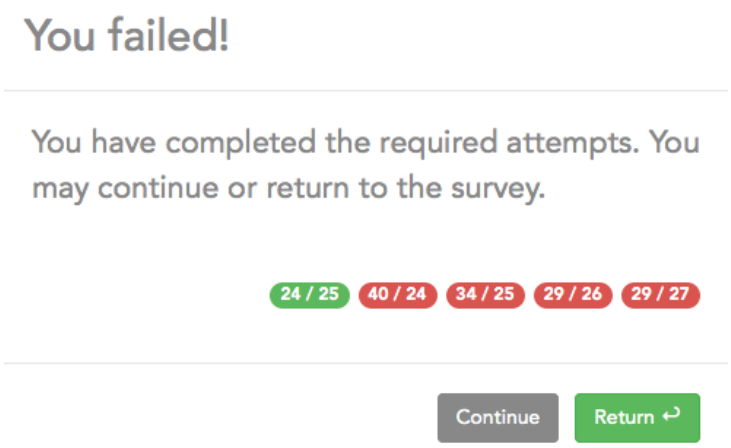

Figure 2. The modal window shown to participants who won only the first of 5 required rounds. The numbers in the bullets correspond with the performance of the participant, red bullets depict a failed attempt while successful rounds are shown in green.

Table 2

The framing that was applied to experimentally manipulate the context of the activity.

Condition	Framing
Task	Next you will complete 5 rounds of Flood It. The aim is to flood the whole area with one color in less than 25 steps. Please continue to the page to complete the task.
Game	Flood It! is a game and we are trying to get some idea of what people think about it and whether they like it. Next you will play 5 rounds of Flood It! The aim is to flood the whole board with one color in less than 25 steps. Please continue to the page to play the game.
Test	Flood It! is a test of problem solving and planning abilities; it will show how well you are able to plan ahead and solve complex problems. In fact, preliminary research with this instrument suggests that people with high intelligence tend to do better at it. Next you will complete 5 rounds of Flood It! The aim is to flood the whole area with one color in less than 25 steps. Please continue to the page to complete this test.
Task with meaning	Flood It! is a complex computational problem (NP-hard) and we would like to train our algorithm with the solution processes you generate. Next you will complete 5 rounds of Flood It! The aim is to flood the whole area with one color in less than 25 steps. Please continue to the page.

Results Experiment 1

Data cleaning. To ensure valid and useful responses, several steps were taken to guarantee data quality. Based on the recommendations by Meade and Craig (2012), the following data were removed from the full data set ($N = 236$):

1. Responses from people who dropped out before the last question.
2. Responses from people who did not select the right answer in an attention check question.
3. Respondents whose response time was identified as an outlier by the box-plot criterion (Tukey, 1977).

In some cases participants managed to win more than half of the rounds of Flood It. It is likely that those participants did not perceive their performance as insufficient and perceived the feedback differently. Thus, data of these participants were removed from the analysis.

Descriptives. Out of 157 remaining participants, one managed to win all 5 rounds (an outcome not anticipated by the author), one participant won 4 rounds, 7 participants won 3 rounds, 30 won 2 rounds, 57 won one round, and 61 participants lost all rounds. As described in the method section of Experiment 1, only participants who lost more than half (3 or more) of the required rounds were included in the analysis. Hence, a sample of $N = 148$ observations remained.

Gender. Forty participants were in the task condition (28 female, 12 male), 36 in the game (20 female, 16 male), 35 in the test (19 female, 16 male), and 37 in the test condition (16 female, 21 male). The observed distribution of gender revealed no significant differences between the conditions, $\chi^2(3, N = 148) = 5.672$, $p = .13$.

Age. The age range was diverse with values from 18 to 74, with a median of 38. No significant age differences between the conditions were observed, $F(3, 144) = 0.29$, $p = .83$.

Effects of framing on change in need satisfaction, intrinsic motivation and affect. A multivariate analysis of variance (MANOVA) was conducted to understand the effect of framing on change in the different motivational variables and affect. Subsequently, mixed-design analyses of variance (ANOVA) were conducted to examine the effects of losing in different conditions. All significance tests were carried out at an α-level of .05. The means and standard deviations of each dependent variable are presented in Table 3. Unless noted, all data were checked for normal distribution of residuals. Sphericity tests were not needed for the repeated-measures ANOVA with only two levels in the repeated factor.

MANOVA. A one-way (type of framing: task, game, test, task+meaning) MANOVA was conducted with change in perceived competence, change in perceived autonomy, change in intrinsic motivation, and the two affective state variables (change in positive and negative affect) as dependent variables. Investigation of the variance-covariance matrices suggested differences between the groups. However, given the reasonable sample size, larger variances and covariances lead to conservative probability values and reduce type-I error (Field, Miles, & Field, 2012). Nevertheless, homogeneities of the variances were investigated using the Bartlett and Flinger-Killeen test. No significant deviations were observed. In a next step, multivariate normality for each group was examined using Shapiro-Wilk tests. For all groups, significant deviations from multivariate normality were observed. Thus, multivariate Q-Q plots for each group and an Adjusted Quantile (AQ) plot[2] were examined to investigate the degree of deviation from normality and possible outliers. Based on the AQ-plot, 27 multivariate outliers were identified and excluded. The MANOVA showed no main effect for framing (Willks' $\Lambda = .89789$, $F(3, 117) = 0.829$, $p = .65$). Thus, no further univariate analysis was conducted with Δ-values.

[2]The AQ-plot of the *mvoutlier* package for *R* (Filzmoser & Gschwandtner, 2015) plots the ordered squared robust Mahalanobis distances (MD) of the observations against the empirical distribution function of the squared MD. Multivariate outliers are identified based on a 0.975 χ^2-quantile (Filzmoser, 2005).

Table 3

Means (and standard deviations) of dependent variables by condition for Experiment 1.

Variable	Framing							
	Task		Game		Test		Task + meaning	
Autonomy								
Pre	5.34	(1.57)	5.18	(1.24)	5.35	(1.24)	5.17	(1.4)
Post	5.54	(1.4)	5.27	(1.44)	5.23	(1.39)	5.18	(1.48)
Δ (change)	0.2	(0.91)	0.09	(0.83)	−0.12	(0.63)	0	(0.85)
Competence								
Pre	5.75	(0.98)	5.44	(0.92)	5.33	(0.96)	5.53	(1.02)
Post	2.91	(1.06)	3.1	(1.65)	3.33	(1.5)	3.32	(1.56)
Δ (change)	−2.84	(1.27)	−2.33	(1.55)	−2	(1.41)	−2.22	(1.79)
Intrinsic motivation								
Pre	5.53	(1.31)	5.4	(1.36)	5.58	(1.12)	5.37	(1.21)
Post	4.9	(1.49)	5.06	(1.6)	4.93	(1.55)	4.86	(1.46)
Δ (change)	−0.62	(0.83)	−0.34	(0.79)	−0.65	(1.02)	−0.51	(0.97)
Positive affect								
Pre	3.79	(0.98)	3.71	(0.87)	3.69	(0.84)	3.66	(0.93)
Post	3.26	(1.13)	3.51	(0.97)	3.44	(0.75)	3.52	(0.99)
Δ (change)	−0.53	(0.73)	−0.19	(0.56)	−0.25	(0.7)	−0.15	(0.49)
Negative affect								
Pre	1.22	(0.49)	1.24	(0.6)	1.28	(0.61)	1.23	(0.53)
Post	1.57	(0.6)	1.55	(0.74)	1.56	(0.7)	1.49	(0.65)
Δ (change)	0.36	(0.59)	0.31	(0.54)	0.28	(0.4)	0.26	(0.57)
Importance	5.49	(1.09)	5.54	(1.09)	5.55	(1.1)	5.28	(1.36)
Value	3.31	(1.47)	4.1	(1.62)	4.06	(1.45)	3.77	(1.54)
Tension	2.82	(1.02)	3.34	(1.25)	2.94	(1.42)	3.02	(1.14)
Played again	0.28	(0.45)	0.25	(0.44)	0.31	(0.47)	0.3	(0.46)

Autonomy. A two-way ANOVA with repeated measures in one factor was conducted to compare the effect of framing type on autonomy need satisfaction in task, game, test, and meaningful task conditions before and after playing the main game (factor time; see Figure 3). The analysis showed no significant main effect for framing, $F(3, 144) = 0.27$, $p = .85$, $\eta^2 = 0.01$[3]; no significant main effect for time, $F(1, 144) = 0.41$, $p = .52$, $\eta^2 < 0.01$ and the interaction between condition and time was also not significant, $F(3, 144) = 1.04$, $p = .38$, $\eta^2 < 0.01$.

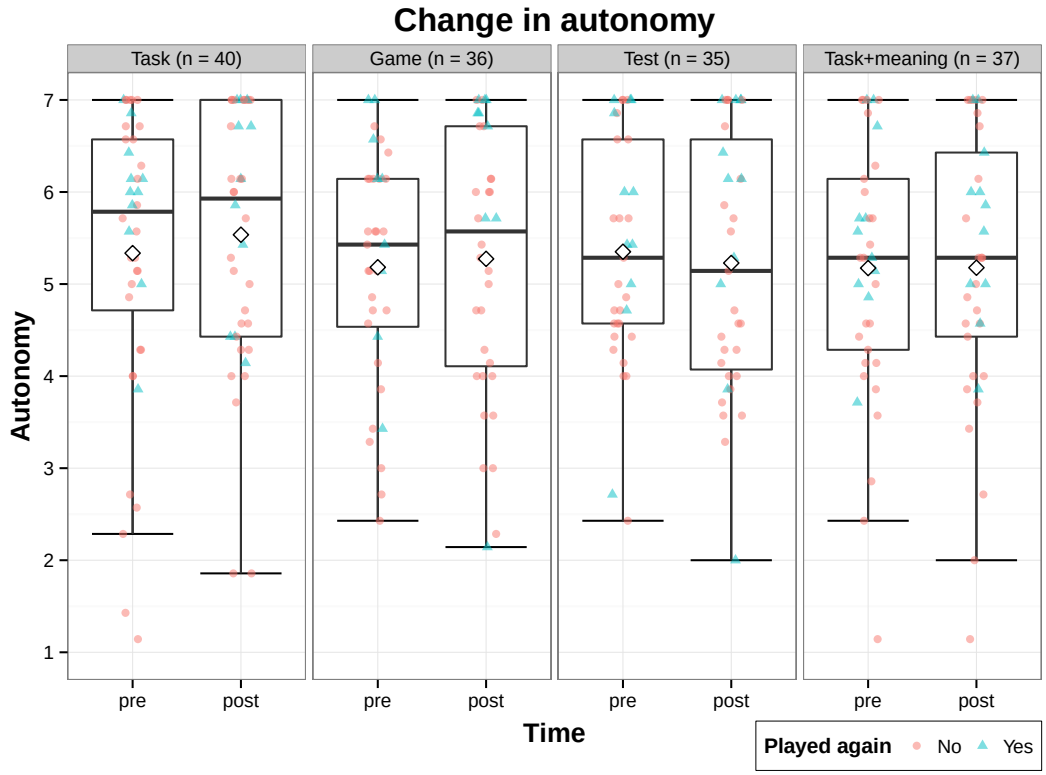

Figure 3. Change in autonomy and type of framing. The rhombus indicates the mean of the variable. *Played again* refers to whether a participant played an additional round of the game after the experiment or not.

[3]All effect sizes for ANOVA are *generalized* η^2, calculated with the formula by Olejnik and Algina (2003). Generalized η^2 can be compared to Cohen (1988) guidelines for small (0.2 - 0.3), medium (0.5) and large (0.8) effects.

Competence. A two-way repeated measures ANOVA was conducted to compare the effect of framing type on competence need satisfaction in task, game, test, and meaningful task conditions before and after playing the main game (see Figure 4). The analysis showed no significant main effect for framing, $F(3, 144) = 0.16$, $p = .92$, $\eta^2 < 0.01$; a significant main effect for time, $F(1, 144) = 354.97$, $p < .001$, $\eta^2 = 0.48$; the interaction between condition and time was not significant, $F(3, 144) = 2.10$, $p = .10$, $\eta^2 = 0.02$.

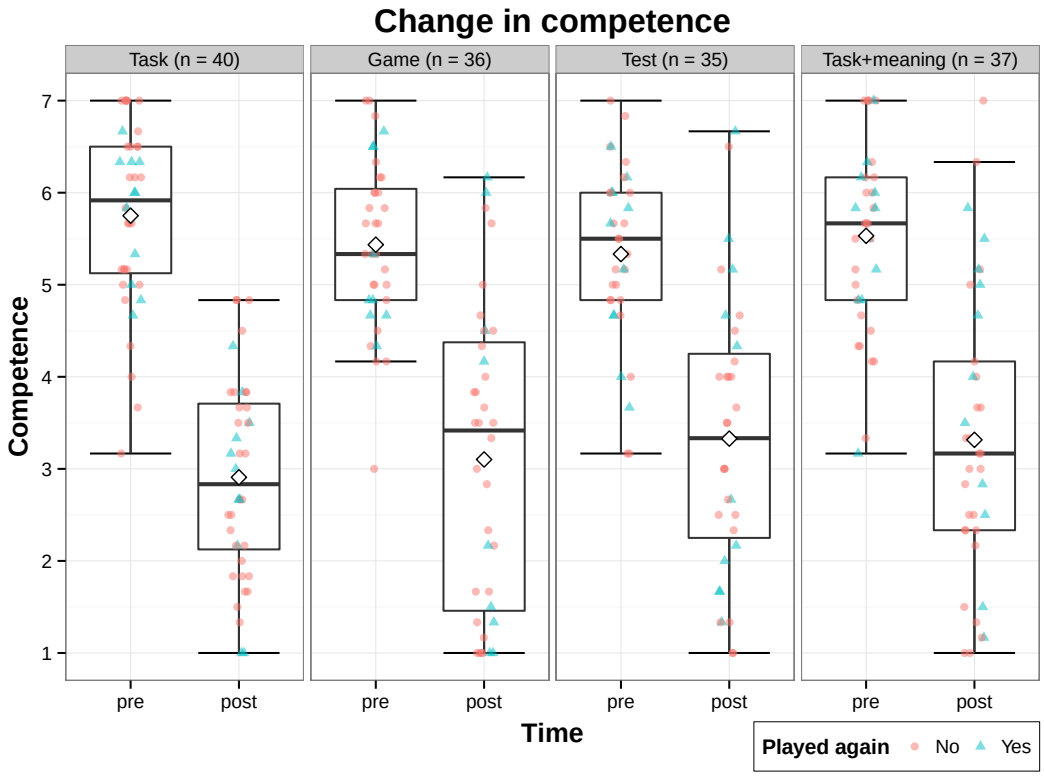

Figure 4. Change in competence and type of framing.

Intrinsic motivation. A two-way repeated measures ANOVA was conducted to compare the effect of framing type on intrinsic motivation in task, game, test, and meaningful task conditions before and after playing the main game (see Figure 5). The analysis showed no significant main effect for framing, $F(3, 144) = 0.08$, $p = .97$, $\eta^2 < 0.01$; a significant main effect for time, $F(1, 144) = 50.58$, $p < .001$, $\eta^2 = 0.04$; the interaction between condition and time was not significant, $F(3, 144) = 0.87$, $p = .46$, $\eta^2 < 0.01$.

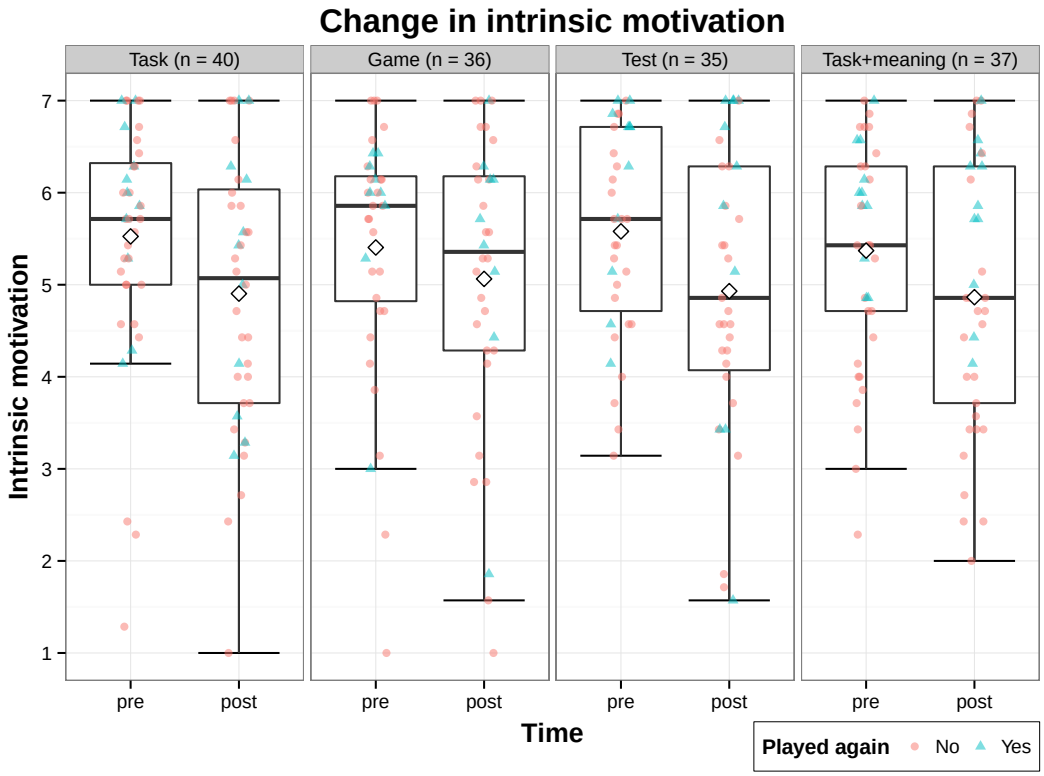

Figure 5. Change in intrinsic motivation and type of framing.

Positive and negative affect. A two-way repeated measures ANOVA was conducted to compare the effect of framing type on positive affect in task, game, test, and meaningful task conditions before and after playing the main game (see Figure 6). The analysis showed no significant main effect for framing, $F(3,144) = 0.06$, $p = .98$, $\eta^2 < 0.01$; a significant main effect for time, $F(1,144) = 29.19$, $p < .001$, $\eta^2 = 0.02$; the interaction between condition and time was significant, $F(3,144) = 2.93$, $p < .05$, $\eta^2 < 0.01$. Post hoc comparison with paired t-tests (using Holm-Bonferroni correction; Holm, 1979) revealed a significant reduction of positive affect in the Task condition $t(39) = 4.62$, $p < .0125$, $d = 0.73$, but not in the Game $t(39) = 2.07$, $p = .05$, $d = 0.19$, Test $t(34) = 2.09$, $p = .04$, $d = 0.34$ and Task + meaning $t(37) = 1.82$, $p = .08$, $d = 0.30$, conditions (see Figure 7). The analysis for negative affect showed no significant main effect for framing, $F(3,144) = 0.08$, $p = .97$, $\eta^2 < 0.01$; a significant main effect for time, $F(1,144) = 47.56$, $p < .001$, $\eta^2 = 0.06$; the interaction between condition and time was not significant, $F(3,144) = 0.26$, $p = .85$, $\eta^2 < 0.01$.

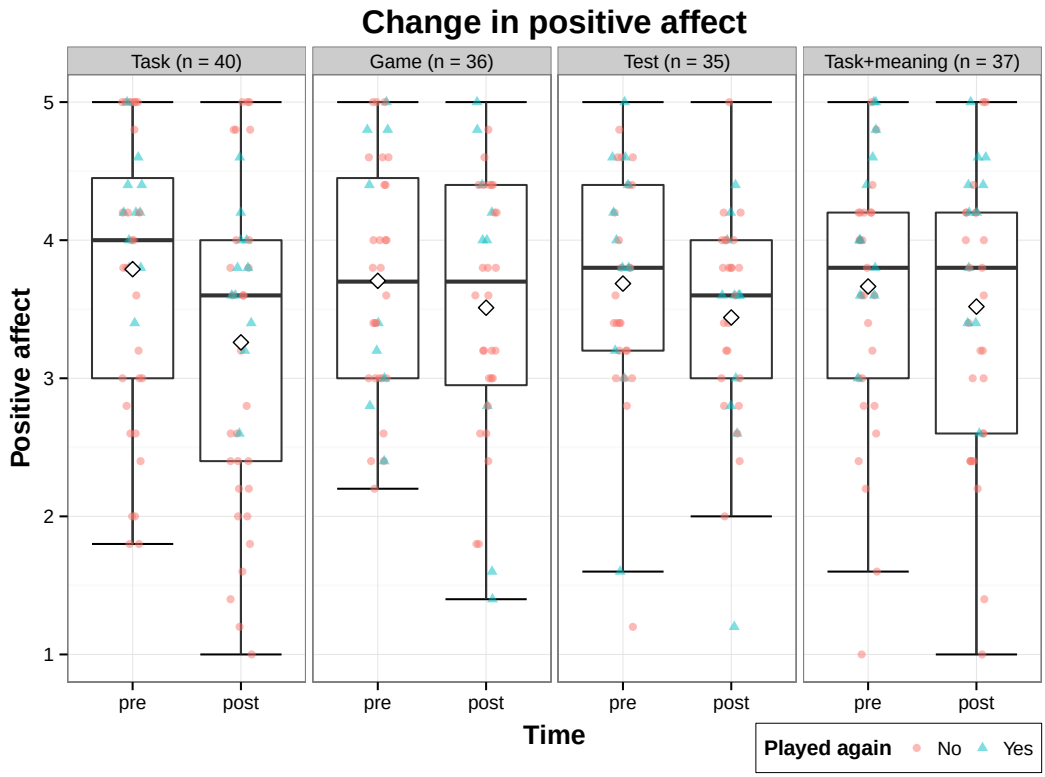

Figure 6. Change in positive affect and type of framing.

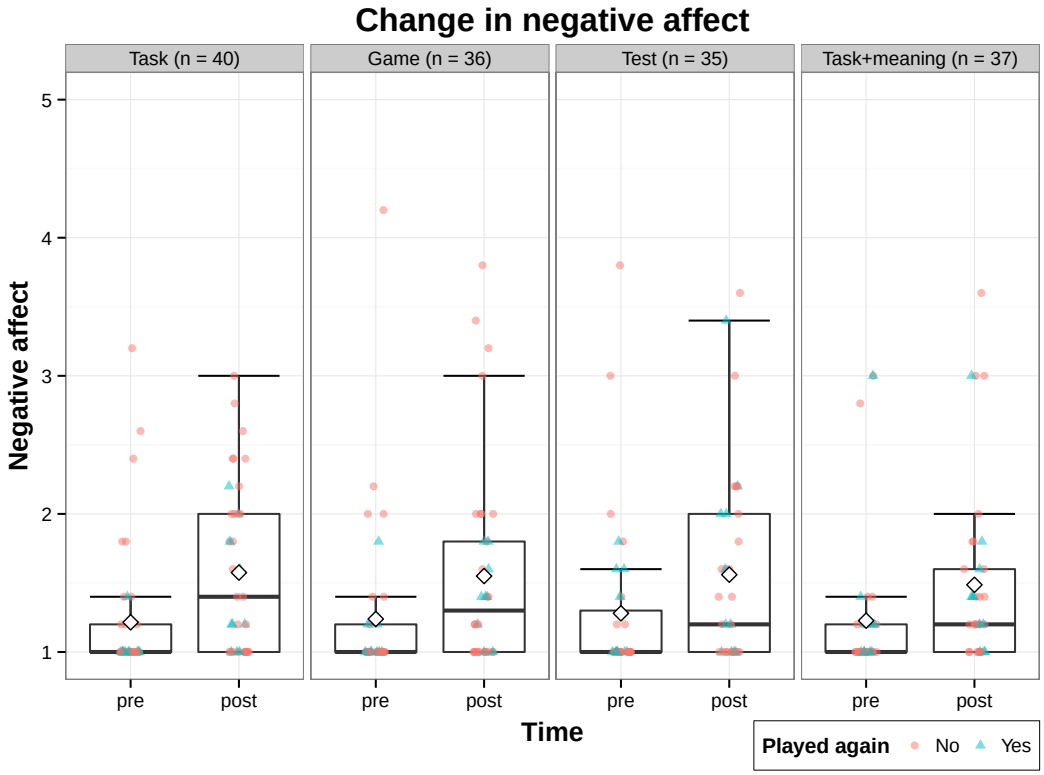

Figure 7. Change in negative affect and type of framing.

Importance, value, and tension. Three one-way ANOVA were conducted to examine the effect of framing type on personal importance, value, and perceived tension (see Figure 8). There was no significant difference in personal importance observed, $F(3, 144) = 0.42$, $p = .74$, $\eta^2 < 0.01$. Also for value and tension, no significant difference was found, $F(3, 144) = 2.21$, $p = .09$, $\eta^2 = 0.04$, $F(3, 144) = 1.24$, $p = .30$, $\eta^2 = 0.03$, respectively. Levene's Test revealed no significantly different variances between the conditions for importance, $F(3, 144) = 0.45$, $p = .72$; value, $F(3, 144) = 0.40$, $p = .75$; and tension $F(3, 144) = 1.68$, $p = .17$. The residuals of all three variables were approximately normally distributed, except for importance. However, because no significant differences were found, there was no risk of type-I error inflation (Schmider, Ziegler, Danay, Beyer, & Bühner, 2010).

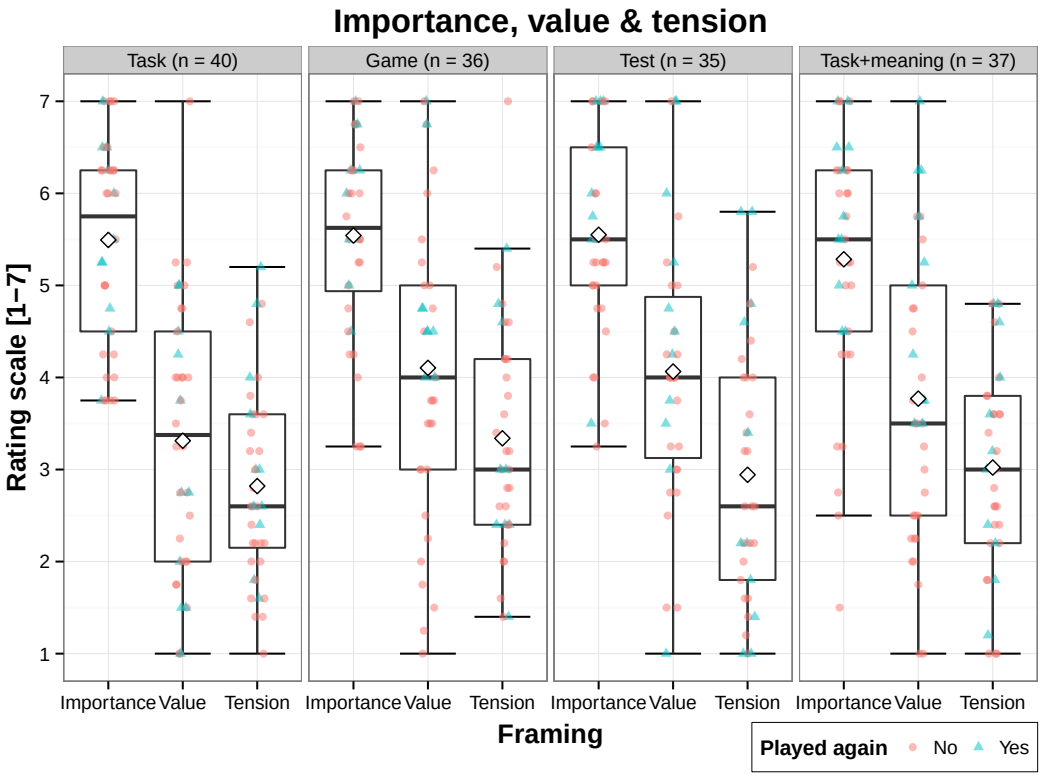

Figure 8. Ratings of covariates and type of framing.

Score and time. Score was defined as how many steps people were below the maximum allowed number of steps. There were no significant differences between the conditions observed, $F(1,144) = 0.546$, $p = .65$, $\eta^2 = 0.01$. Levene's test revealed no significantly different variances between the conditions, $F(3,144) = 0.660$, $p = .58$. In terms of the time invested (reciprocal transformed for analysis) in the five required rounds, again no significant differences were observed, $F(1,144) = 0.869$, $p = .45$, $\eta^2 = 0.02$. Levene's test was also not significant, $F(3,144) = 0.857$, $p = .47$. Values of the score and time variables are depicted in Figure 9 and Figure 10.

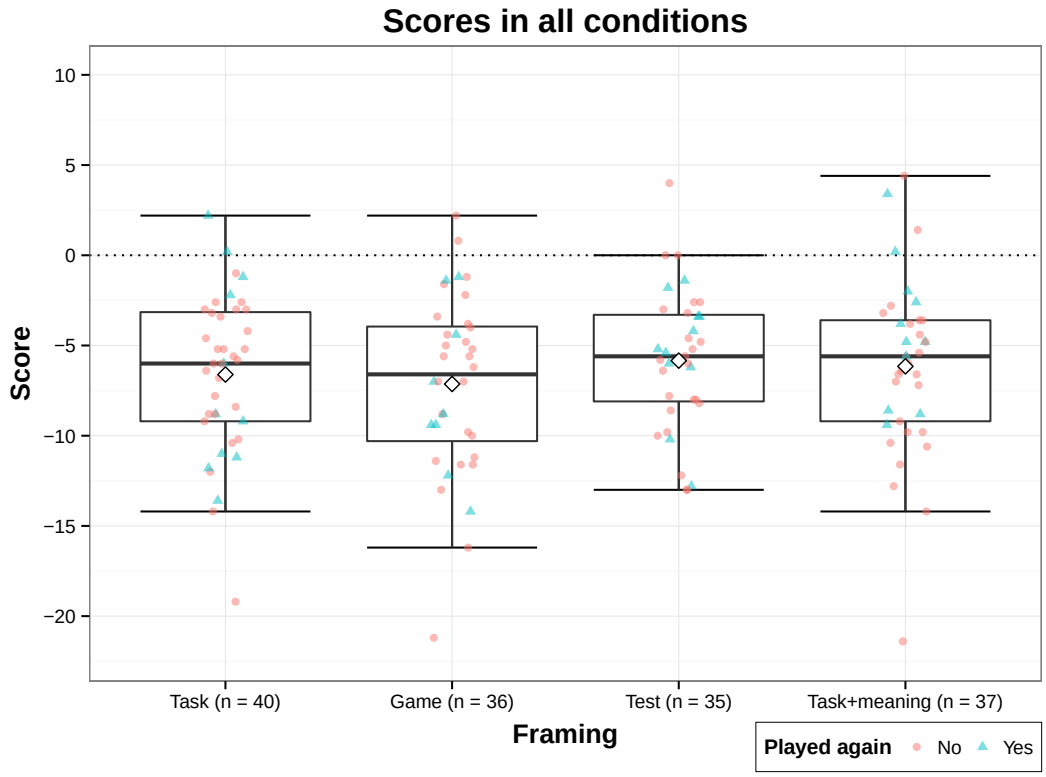

Figure 9. Score and type of framing.

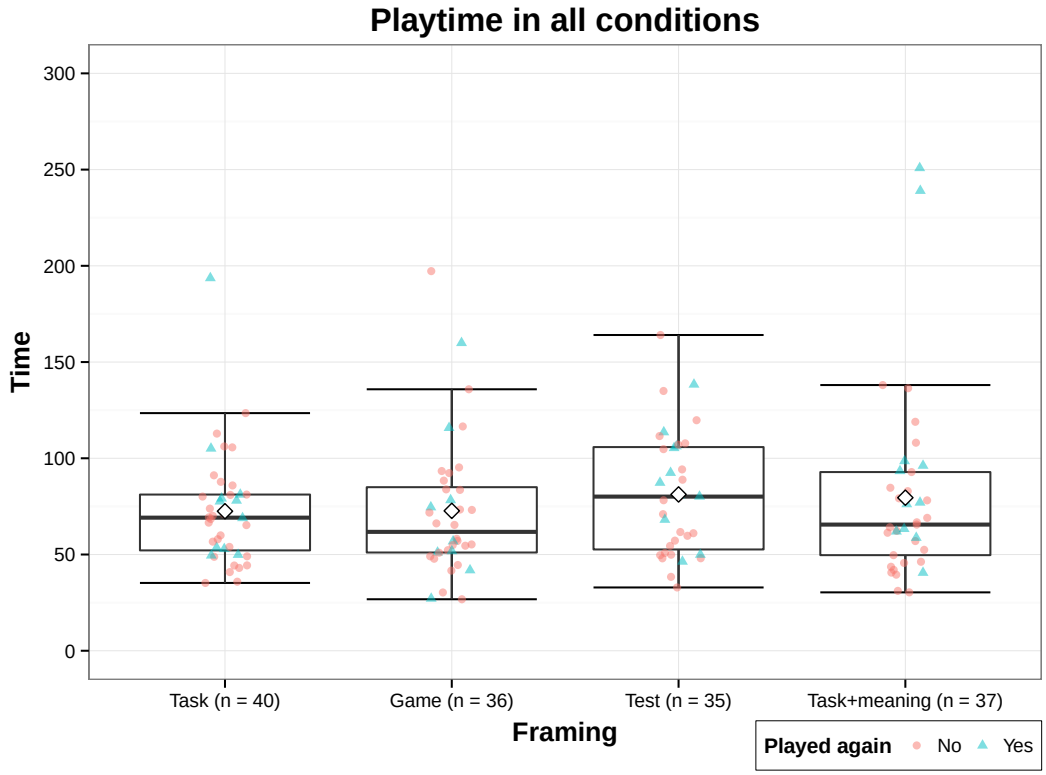

Figure 10. Playtime and type of framing.

Number of extra rounds and extra playtime. Participants had the option to play extra rounds after completing the minimum required amount of rounds. The distribution of extra rounds was right-skewed (see Figure 11) and the distribution of residuals significantly differed from normality. Log-transformation of data did not reduce the deviation, thus it was decided to run a Kruskal-Wallis rank sum test to compare the different conditions. Results showed no significant difference between the conditions $\chi^2(3) = 0.48$, $p = .92$. As with the extra rounds, the distribution of extra playtime deviated from normal distribution. Log-transformation of data reduced the deviation and it was decided to analyze data using ANOVA and Kruskal-Wallis rank sum test. Only the nonparametric results are reported: No significant difference was observed $\chi^2(3) = 3.26$, $p = .35$ (see Figure 11).

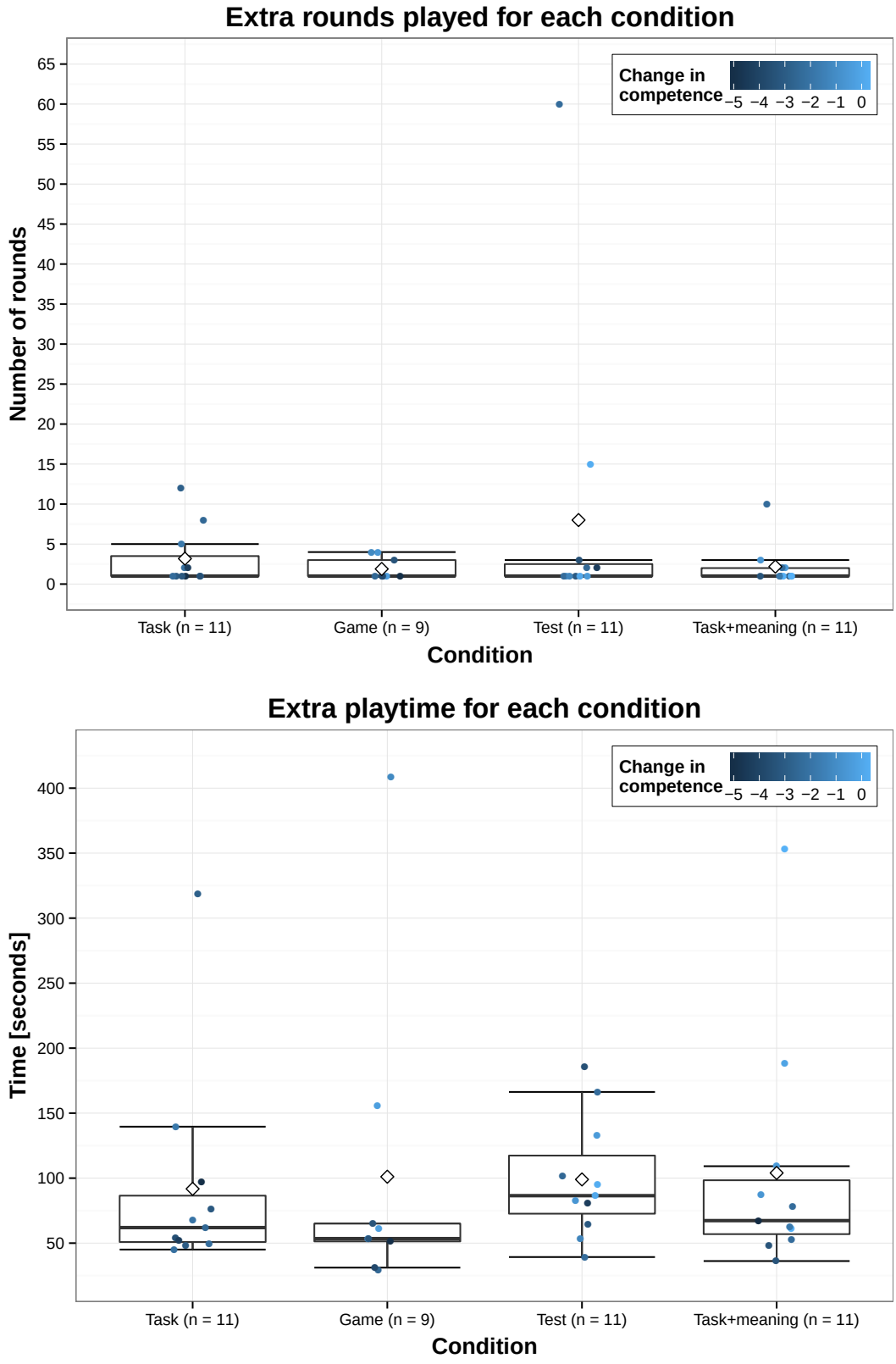

Figure 11. Number of extra rounds played by the participants and extra playtime in all four conditions.

Correlation of change in competence and intrinsic motivation. To see if losing and the subsequent loss in competence was also associated with a loss of intrinsic motivation, scatter-plots were investigated (see Figure 12). As explained before, multivariate outliers were identified using AQ-plots (shaded in blue in Figure 12). Based on the relatively large number of outliers and deviation from normal distribution it was decided to calculate Kendall rank correlation. A significant correlation of change in self-reported intrinsic motivation and change in competence was observed in the test condition ($r_\tau(N = 35) = .22$, $p < .05$, 95% CI_{bs}(.05, .39), 95%CI(.04, .4)). However, this relationship was not significant in the task ($r_\tau(N = 40) = -.09$, $p = .58$, 95% CI_{bs}(-.37, .18), 95% CI(-.36, .21)) and game ($r_\tau(N = 36) = .0$, $p = .93$, 95% CI_{bs}(-.22, .21), 95% CI(-.23, .21)) and task + meaning ($r_\tau(N = 37) = .1$, $p = .47$, 95% CI_{bs}(-.14, .37), 95% CI(-.16, .34)) conditions. For the test, game and meaningful task conditions, a loss in competence was not associated with a loss in intrinsic motivation, but in the test condition it was significantly associated.

Figure 12. Scatterplot of change in competence and change in autonomy with regression lines, 95% confidence area shaded in grey, and multivariate outliers shaded in blue.

Prediction of replay. To predict whether participants played at least one additional round after completion of the experiment, a forward and backward selection algorithm selected the best predictors of the full model with change in autonomy, change in competence, change in intrinsic motivation, change in positive and negative affect as well as value, importance and tension in a binomial logistic regression. The resulting model consisted only of perceived value as a predictor of replay. This model was then tested to assess its predictive power (see Figure 13 and Table 4). Results showed that perceived value was able to predict whether people play additional rounds in the game. However, sensitivity (the fraction of returning players correctly predicted) was low with .12, meaning that it was very difficult to correctly predict which players stay for an additional round (or more).

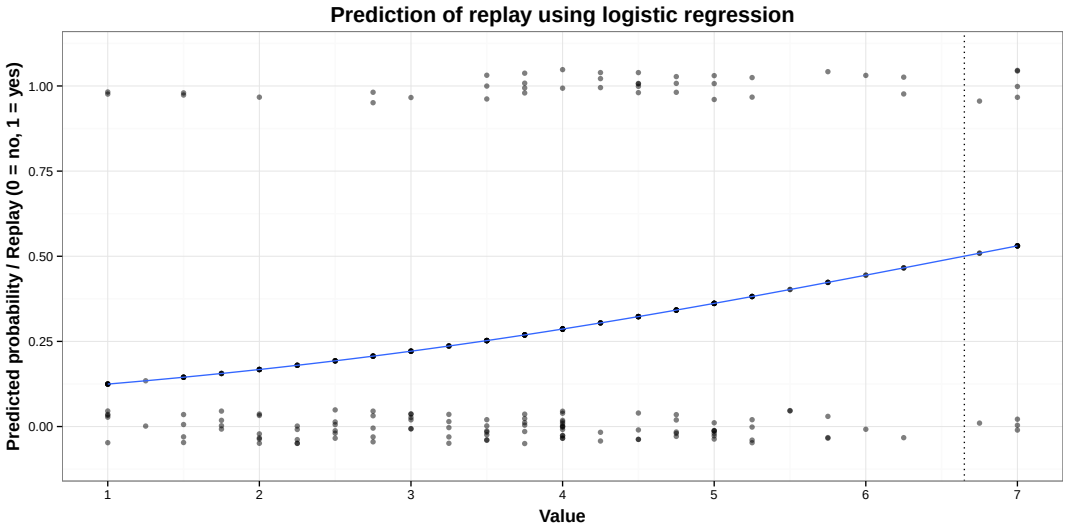

Figure 13. Predicted replay probabilities for perceived value and the observed replay behavior. The dots connected with the blue line indicate probabilities of replay. The behavior was coded 0 for no replay and 1 for at least one additional round played. The reduce over-plotting, a vertical jitter was applied. The dotted line indicates the cutoff after that the probability of replay is above .5.

Table 4

Binomial logistic regression statistics. Nagelkerke R^2 (Max rescaled R^2) = .075. Kendall's Tau − α = .120. Goodman-Kruskal Gamma = .307. Somers's D_{xy} = .293. c-statistic = 64.7%. Accuracy = .72, Precision = .56, Negative predictive value = .73, Sensitivity = .12, Specificity = .96.

Predictor	β	SE β	Wald's χ^2	df	p	e^{β} (odds ratio)	2.5%	97.5%
Constant	-2.2938	0.5537	17.2	1	<0.001	0.1009	0.0320	0.2842
Value	0.3451	0.1266	7.4	1	0.006	1.4122	1.1089	1.8271
Test			**χ^2**	**df**	**p**			
Overall model evaluation								
Likelihood ratio test			7.9488	1	0.005			
Score test			7.8298	1	0.005			
Goodness-of-fit test								
Hosmer & Lemeshow			10.526	8	0.2301			

Correlation analysis. Table 5 depicts intercorrelations among all dependent variables including all participants (148 + 9). Based on Kendall rank correlations several significant associations were found. Intrinsic motivation correlated with mean score and mean time spent per round. Time was significantly positively associated with autonomy, intrinsic motivation, value and score. Score was significantly positively associated with competence, intrinsic motivation, value and score. Value, importance, positive affect, autonomy and competence were positively associated with intrinsic motivation. Negative affect was positively associated with tension.

Table 5

Kendall rank correlations among autonomy, competence, intrinsic motivation, and affective state after playing (post) as well as score, and time (mean per round) including all participants. Values above the diagonal indicate bootstrapped 95% confidence intervals (1000 replications).

Experiment 1 ($N = 157$)	1	2	3	4	5	6	7	8	9	10
1. Autonomy (post)		(-.07, .18)	(.25, .45)	(.02, .24)	(-.31,-.07)	(.09, .32)	(.08, .29)	(-.25,-.03)	(-.02, .21)	(.03, .27)
2. Competence (post)	.05		(.10, .32)	(-.03, .22)	(-.28,-.03)	(-.14, .09)	(.13, .33)	(-.30,-.09)	(.16, .38)	(-.02, .17)
3. IM (post)	.35***	.21***		(.27, .46)	(-.27,-.03)	(.27, .49)	(.35, .55)	(-.22, .00)	(.02, .24)	(.13, .32)
4. PA (post)	.13*	.10	.37***		(-.08, .17)	(.32, .51)	(.19, .40)	(-.10, .14)	(-.16, .07)	(-.03, .19)
5. NA (post)	-.19**	-.16*	-.15*	.05		(-.07, .19)	(-.09, .16)	(.29, .48)	(-.14, .09)	(-.18, .04)
6. Importance	.21***	-.02	.39***	.41***	.06		(.23, .44)	(-.13, .13)	(-.22, .02)	(-.04, .18)
7. Value	.18***	.23***	.46***	.29***	.04	.34***		(-.11, .09)	(.01, .24)	(.00, .21)
8. Tension	-.14*	-.19***	-.11	.02	.39***	0	-.01		(-.07, .15)	(-.15, .08)
9. Score	.10	.27***	.13*	-.04	-.02	-.10	.13*	.04		(.02, .24)
10. Time	.15*	.07	.23***	.08	-.07	.07	.11*	-.04	.13*	

Note. IM = Intrinsic Motivation, PA = Positive Affect, NA = Negative Affect. $p < .001$***; $p < .01$**; $p < .05$*

Discussion Experiment 1

Need satisfaction, intrinsic motivation, and affect. No significant differences were observed between the conditions concerning autonomy, competence, and intrinsic motivation. This shows that the framing was not able to change the effect of failing on these variables. Interestingly, participants autonomy was not reduced by failure, but a medium effect for competence and a small effect for intrinsic motivation was observed. Unsurprisingly, competence suffered the most by failure. In terms of affective state, it was expected that participants' positive affect was reduced equally in all conditions. However, results showed that participants in the task condition lost more positive affect than participants in the other conditions. This can be explained by the lack of additional information they received, which might have given Flood It less meaning and strengthened the effect of failing on affect. Meaningfulness has been linked to attitudes, engagement, and behavioral outcomes (May, Gilson, & Harter, 2004). Negative effect increased equally in all conditions.

Importance, value, tension, and additional rounds. For the secondary dependent variables no significant differences were observed between the conditions. Importance of the activity was relatively high, while perceived value and tension was rather low. As with the motivational variables, participants performed equally in all four conditions and invested the same amount of time in the five required rounds. Only a relatively low number of participants decided to play an additional round, and they were equally distributed across all four conditions. This partially contradicts the finding of Burgers et al. (2015), who found that participants of their brain-training game indicated higher levels of immediate replay after receiving negative feedback. A similar effect was expected in the test condition of this experiment.

Correlation of change in competence and intrinsic motivation. Analyses of correlations between the variables showed that the loss of competence was only in the test condition associated with a loss of intrinsic motivation. In SDT, competence and autonomy (with relatedness) are antecedents of intrinsic motivation (Deci & Ryan, 2004), thus it is very likely that in an environment with a higher risk for the self-image, the resulting intrinsic motivation is more dependent on perceived competence which was also associated with score.

Prediction of replay. One of the research questions was whether it was possible to predict replay with the motivational and affective variables studied in this experiment. Perceived value was significantly associated with replay, but its sensitivity was rather low due to the large number of people who did not play an additional round. It seems that also other, not observed variables, influenced replay. One possible explanation for the moderate levels of replay would be that participants tried to finish the study as quickly as possible because they were paid for the completion of the study.

Intercorrelation pattern. The pattern of the intercorrelation between the variables measured after engaging in Flood It showed that the need satisfaction variables autonomy and competence were strongly associated with intrinsic motivation as well as the received score and the time people invested. Thus, it was shown that score was associated with time, competence, intrinsic motivation, and perceived value. Time was more strongly associated with autonomy, but also very strongly with intrinsic motivation and with value. Value was correlated with all variables except negative affect and tension. Value and importance were very strongly associated with intrinsic motivation, as well as autonomy, positive affect, and, to a lesser degree, with competence. This shows that meaning or how valuable and important this task is perceived, seem to be as relevant as need satisfaction. Especially since value was the strongest predictor of replay, although not without limitations.

Compensation for participation. However, the strong effects of value and importance could have came from the fact that people were paid to participate in the experiment. Additionally, a possible reason for the lack of difference between the conditions could be that the framing was not strong enough and was overshadowed by the effects of the monetary reward for participants. The reward given for participation could have undermined intrinsic motivation in all conditions. The undermining effect describes the negative effects of rewards on long-term motivation (Deci, Koestner, & Ryan, 2001). It was first reported by Lepper, Sagotsky, Dafoe, and Greene (1982), who observed that if students were given rewards for an activity they intrinsically enjoyed, intrinsic motivation declined below the initial level once the rewards were not given anymore.

To reduce possible effects of the reward on intrinsic motivation, it was decided to run an

additional experiment. In the next experiment participants were not paid for partaking in the survey in order to examine the effects of framing on purely intrinsically motivated participants.

Experiment 2 (no compensation)

To better understand the effects of monetary compensation on the effects of failure and framing, a second experiment was conducted without an incentive. The meaningful task condition was dropped because the main interest was the comparison of a game and test condition to the control condition (task). The procedure and dependent measures were identical to Experiment 1. All scales showed good internal consistency with Cronbach's α ranging from .81 to .93. Participants were recruited on a large internet community and through the University database. As in the first experiment, rigorous data cleaning was conducted with the full sample ($N = 201$).

Results Experiment 2

Descriptives. Out of 67 remaining participants, none managed to win all 5 rounds, four participants won 3 rounds, 26 participants won 2 rounds, 24 won one round, and 13 participants lost all rounds. As described in the method section, only participants losing more than half (3 or more) of the required rounds were included in the analysis. The same data as in the first experiment cleaning was applied. A sample of $N = 63$ remained.

Gender. Twenty-two participants were in the task condition (13 female, 9 male), 20 in the game (11 female, 8 male, 1 not indicated), and 21 in the test condition (10 female, 11 male). The observed distribution of gender revealed no significant differences, $\chi^2(2, N = 62) = 0.674$, $p = .71$.

Age. The age range was diverse with values from 17 to 65, with a median of 26. There was no significant age difference between the conditions, $F(2, 60) = 1.96$, $p = .15$.

Table 6

Means (and standard deviations) of dependent variables by condition for Experiment 2.

Variable	Framing					
	Task		Game		Test	
Autonomy						
Pre	5.65	(1.33)	5.84	(1.06)	5.55	(1.32)
Post	5.66	(1.45)	5.53	(1.01)	5.44	(1.21)
Δ (change)	0.01	(0.73)	−0.31	(0.85)	−0.12	(0.95)
Competence						
Pre	4.67	(1.3)	4.69	(0.88)	4.84	(1.03)
Post	3.28	(1.09)	3.46	(1.02)	3.2	(0.83)
Δ (change)	−1.39	(0.79)	−1.23	(1.09)	−1.64	(1.06)
Intrinsic motivation						
Pre	4.85	(1.17)	5.25	(0.91)	4.78	(1.25)
Post	4.56	(1.44)	4.64	(1.32)	4.45	(1.32)
Δ (change)	−0.29	(0.87)	−0.61	(1.01)	−0.33	(0.82)
Positive affect						
Pre	2.75	(1.14)	2.67	(0.98)	2.48	(0.84)
Post	2.6	(1.14)	2.56	(0.81)	2.57	(0.93)
Δ (change)	−0.15	(0.49)	−0.11	(1.06)	0.1	(0.37)
Negative affect						
Pre	1.18	(0.33)	1.23	(0.31)	1.33	(0.41)
Post	1.41	(0.58)	1.54	(0.52)	1.88	(0.87)
Δ (change)	0.23	(0.36)	0.31	(0.35)	0.54	(0.8)
Importance	4.06	(1.6)	3.81	(1.63)	4.25	(0.98)
Value	2.99	(1.28)	3.17	(1.43)	2.99	(1.3)
Tension	2.55	(1.22)	2.66	(1.08)	3.37	(1.21)
Played again	0.5	(0.51)	0.35	(0.49)	0.38	(0.5)

Effects of framing on need satisfaction, intrinsic motivation and affect. In contrast to Experiment 1, in this experiment participants received no compensation. The sample size was considerably smaller and thus it was decided to run only the most important statistical analysis. Thus, only ANOVAs were conducted and no MANOVA with Δ-values. The means and standard deviations of each dependent variable are presented in Table 6. Unless noted, all data were checked normal distribution of residuals.

Autonomy. A two-way ANOVA with repeated measures in one factor was conducted to compare the effect of framing type on autonomy need satisfaction in task, game, and test conditions before and after playing the main game (factor time; see Figure 14). The analysis showed no significant main effect for framing, $F(2,60) = 0.160$, $p = .85$, $\eta^2 < 0.01$; no significant main effect for time, $F(1,60) = 1.755$, $p = .19$, $\eta^2 < 0.01$ and the interaction between condition and time was also not significant, $F(2,60) = 0.764$, $p = .47$, $\eta^2 < 0.01$.

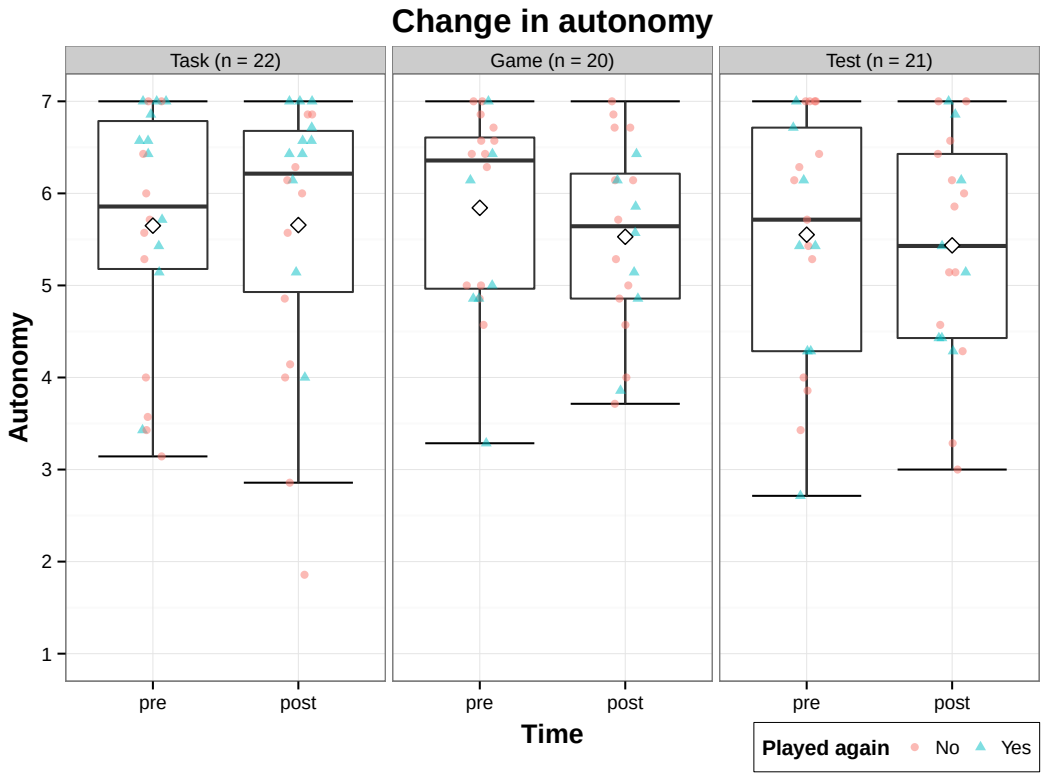

Figure 14. Change in autonomy and type of framing.

Competence. A two-way ANOVA with repeated measures in one factor was conducted to compare the effect of framing type on competence need satisfaction in task, game, and test conditions before and after playing the main game (factor time; see Figure 15). The analysis showed no significant main effect for framing, $F(2,60) = 0.05$, $p = .94$, $\eta^2 < 0.01$; a significant main effect for time, $F(1,60) = 131.182$, $p < .001$, $\eta^2 = 0.33$ but the interaction between condition and time was not significant, $F(2,60) = 0.902$, $p = .41$, $\eta^2 < 0.01$.

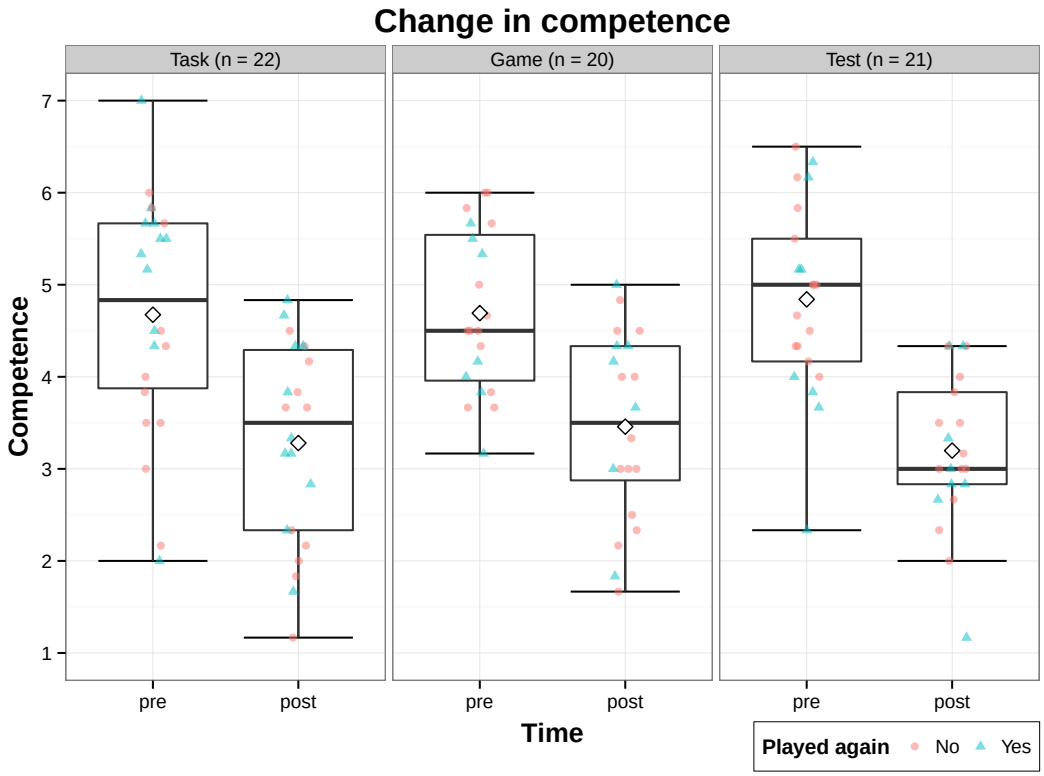

Figure 15. Change in competence and type of framing.

Intrinsic Motivation. A two-way ANOVA with repeated measures in one factor was conducted to compare the effect of framing type on intrinsic motivation in task, game, and test conditions before and after playing the main game (factor time; see Figure 16). The analysis showed no significant main effect for framing, $F(2,60) = 0.428$, $p = .65$, $\eta^2 = 0.01$; a significant main effect for time, $F(1,60) = 13.117$, $p < .001$, $\eta^2 = 0.03$ but the interaction between condition and time was not significant, $F(2,60) = 0.800$, $p = .45$, $\eta^2 < 0.01$.

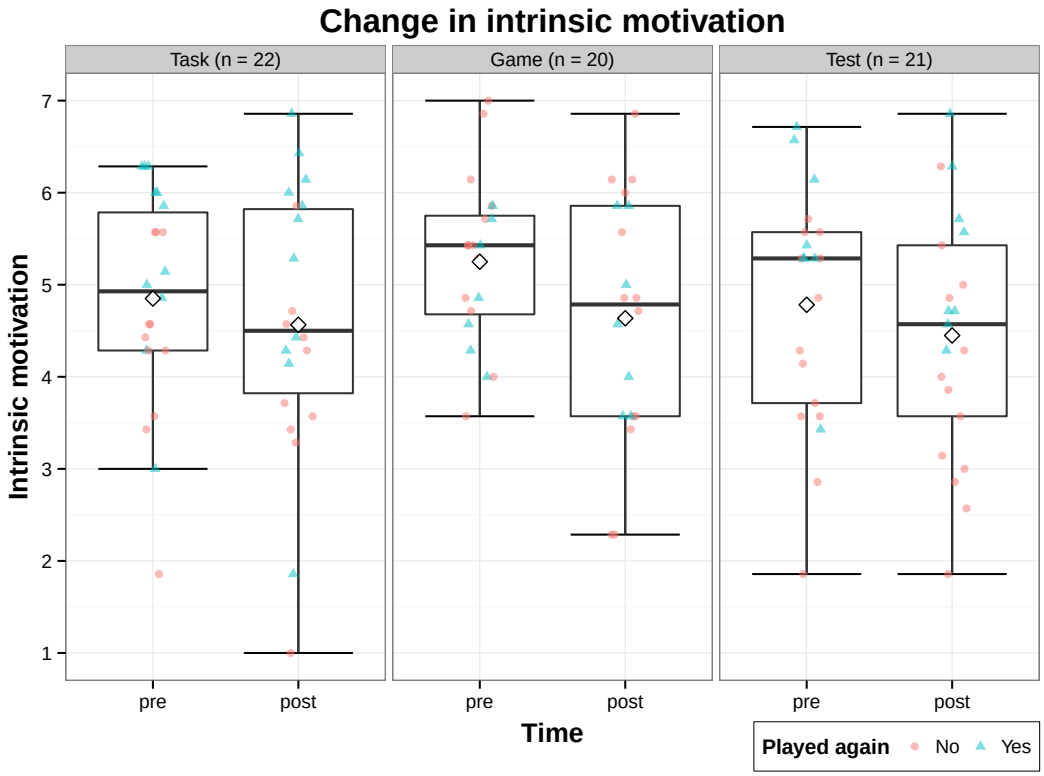

Figure 16. Change in intrinsic motivation and type of framing.

Positive and negative affect. A two-way repeated measures ANOVA was conducted to compare the effect of framing type on positive affect in task, game, and test conditions before and after playing the main game (see Figure 17). The analysis showed no significant main effect for framing, $F(2,60) = 0.141$, $p = .87$, $\eta^2 < 0.01$; no significant main effect for time, $F(1,60) = 0.372$, $p = .54$, $\eta^2 < 0.01$; the interaction between condition and time was also not significant, $F(2,60) = 0.742$, $p = .48$, $\eta^2 < 0.01$.

The analysis for negative affect (see Figure 18) showed no significant main effect for framing, $F(2,60) = 2.521$, $p = .09$, $\eta^2 = 0.06$; a significant main effect for time, $F(1,60) = 27.436$, $p < .001$, $\eta^2 = 0.11$; the interaction between condition and time was not significant, $F(2,60) = 1.917$, $p = .165$, $\eta^2 = 0.02$.

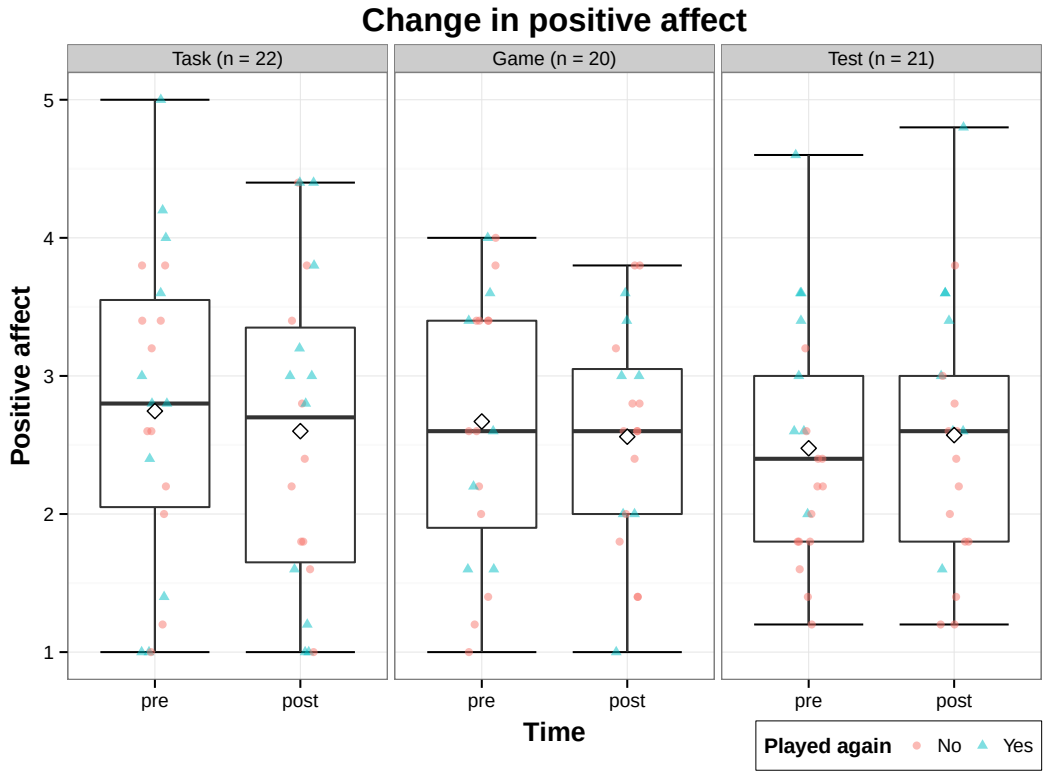

Figure 17. Change in positive affect and type of framing.

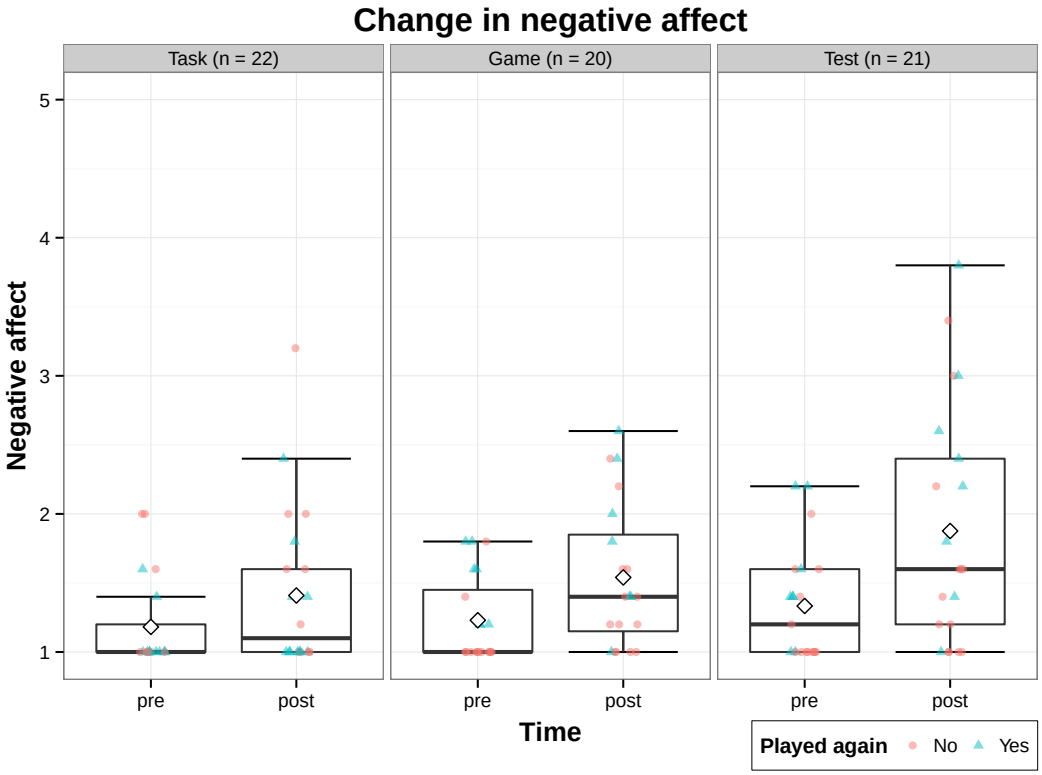

Figure 18. Change in negative affect and type of framing.

Importance, value, and tension. Three one-way ANOVAs were conducted to examine the effect of framing type on personal importance, value, and perceived tension (see Figure 19). There was no significant difference in personal importance observed, $F(2,60) = 0.478$, $p = .62$, $\eta^2 = 0.02$. Also for value and tension, no significant difference was found, $F(2,60) = 0.133$, $p = .86$, $\eta^2 < 0.01$, $F(2,60) = 3.046$, $p = .055$, $\eta^2 = 0.09$, respectively. Levene's Test revealed no significantly different variances between the conditions for importance, $F(2,60) = 3.132$, $p = .051$; value, $F(2,60) = 0.417$, $p = .66$; and tension $F(2,60) = 0.192$, $p = .83$. The residuals of all three variables were approximately normally distributed, except for value. However, because no significant differences were found, there was no risk of type-I error inflation (Schmider et al., 2010).

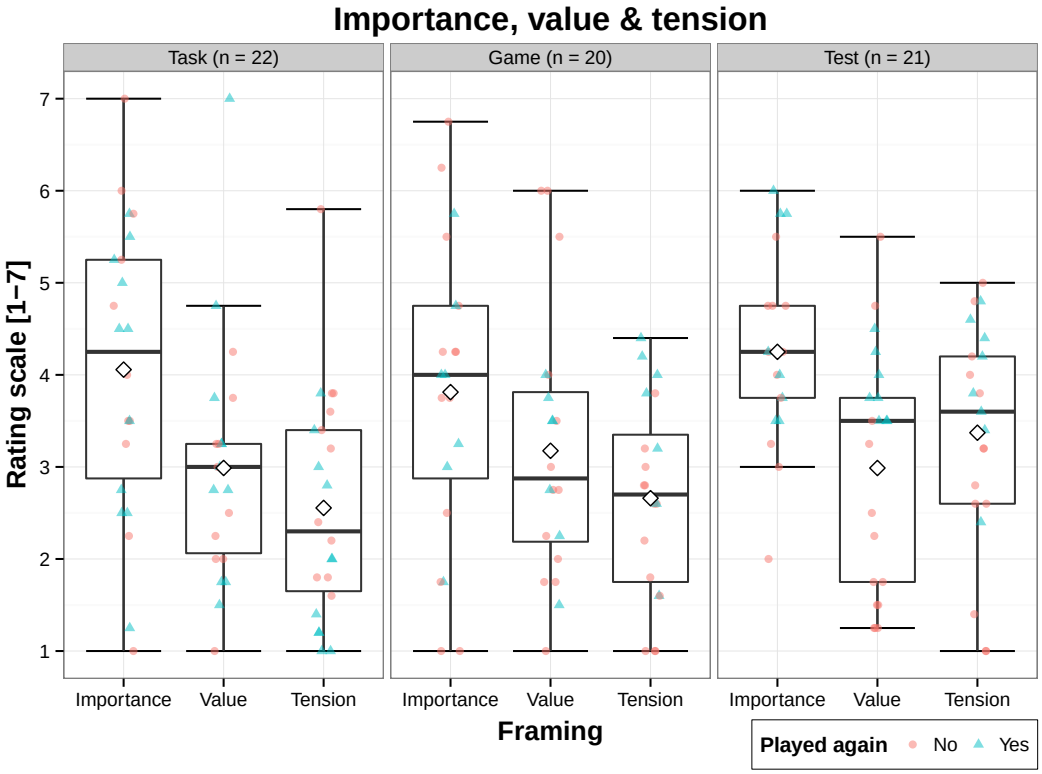

Figure 19. Ratings of covariates and type of framing.

Score and time. Score was defined as how many steps people were below the maximum allowed number of steps. There were no significant differences between the conditions observed, $F(1,60) = 0.043$, $p = .96$, $\eta^2 < 0.01$. Levene's test revealed no significantly different variances between the conditions, $F(2,60) = 0.127$, $p = .88$. In terms of the time (reciprocal transformed for analysis) invested in the five required rounds, again no significant differences were observed, $F(1,60) = 0.330$, $p = .72$, $\eta^2 = 0.01$. Levene's test was also not significant, $F(2,60) = 0.688$, $p = .51$. Values of the score and time variable are depicted in Figure 20.

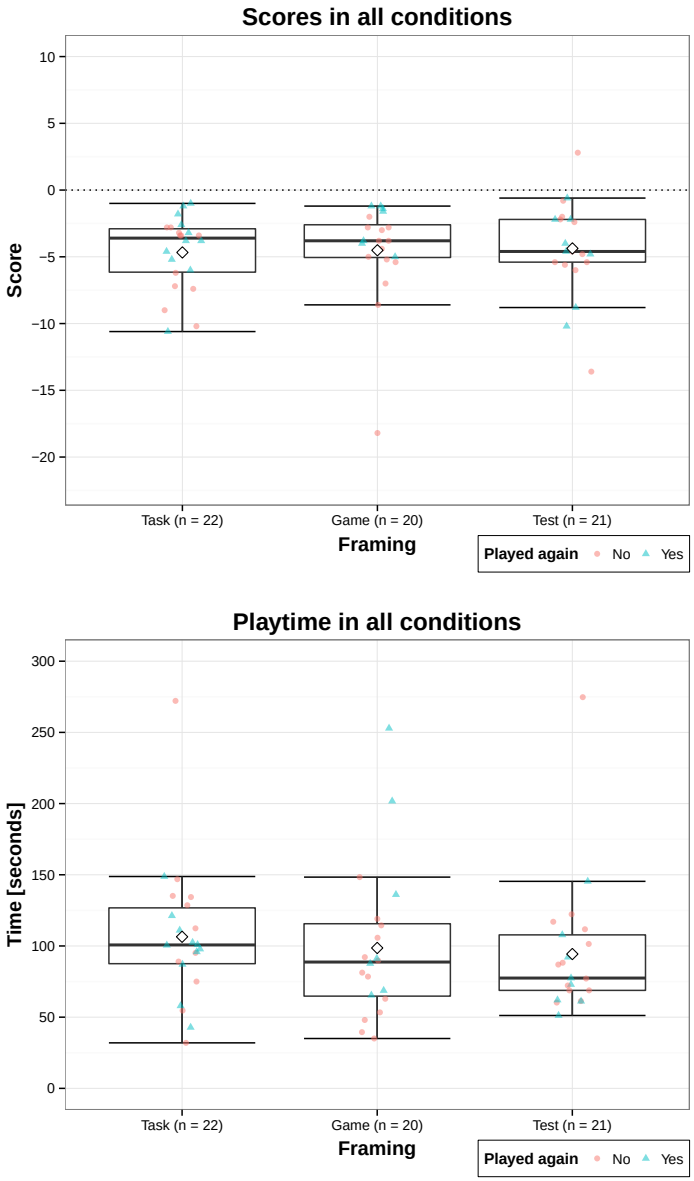

Figure 20. Score and playtime in all three conditions.

Correlation of change in competence and intrinsic motivation. To see if losing and the subsequent loss in competence was also associated with a loss of intrinsic motivation, scatter-plots were investigated (see Figure 12). As in the first experiment, multivariate outliers were identified using AQ-plots (shaded in blue in Figure 21). Based on the relatively large number of outliers and deviation from normal distribution it was decided to calculate Kendall rank correlation. A significant correlation of change in self-reported intrinsic motivation and change in competence was observed in the test condition ($r_\tau(N = 21) = .59$, $p \leq .05$, $95\%CI_{bs}(.32, .84)$, $95\%CI(.26, .81)$). However, this relationship was not significant in the task ($r_\tau(N = 22) = -.14$, $p = .37$, $95\%CI_{bs}(-.46, .18)$, $95\%CI(-.45, .18)$) and game conditions ($r_\tau(N = 20) = .03$, $p = .92$, $95\%CI_{bs}(-.41, .45)$, $95\%CI(-.39, .43)$). For the test and game conditions, a loss in competence was not associated with a loss in intrinsic motivation, but in the test condition it was significantly associated.

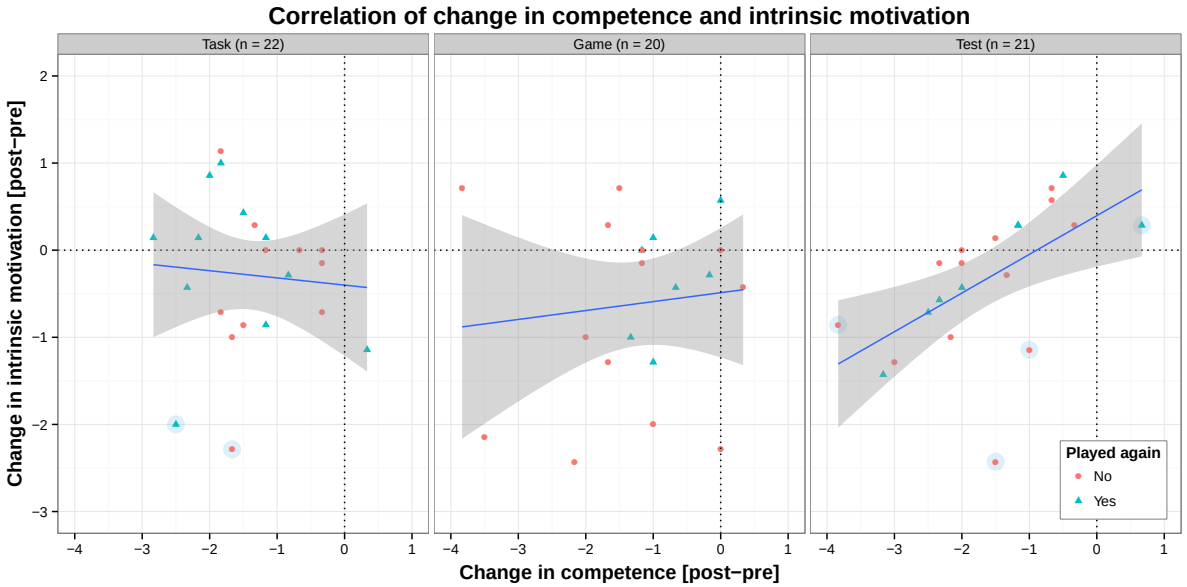

Figure 21. Scatterplot of change in competence and change in autonomy with regression lines including 95% confidence areas shaded in grey, and multivariate outliers shaded in blue.

Number of extra rounds and extra playtime. Participants had the option to play extra rounds after completing the minimum required amount of rounds. The distribution of extra rounds was right-skewed (see Figure 22), the distribution of residuals significantly differed from normality, and the sample size was very low. Thus, it was decided to run a Kruskal-Wallis rank sum test to compare the different conditions. Results showed no significant difference between the conditions $\chi^2(2) = 0.56$, $p = .76$. As with the extra rounds, the residual distribution of extra playtime deviated from normal distribution and sample size was low. Hence, data were again analyzed non-parametrically. No significant difference was observed $\chi^2(2) = 5.83$, $p = .054$ (see Figure 22). However, a trend towards longer playtime in the game condition could be observed.

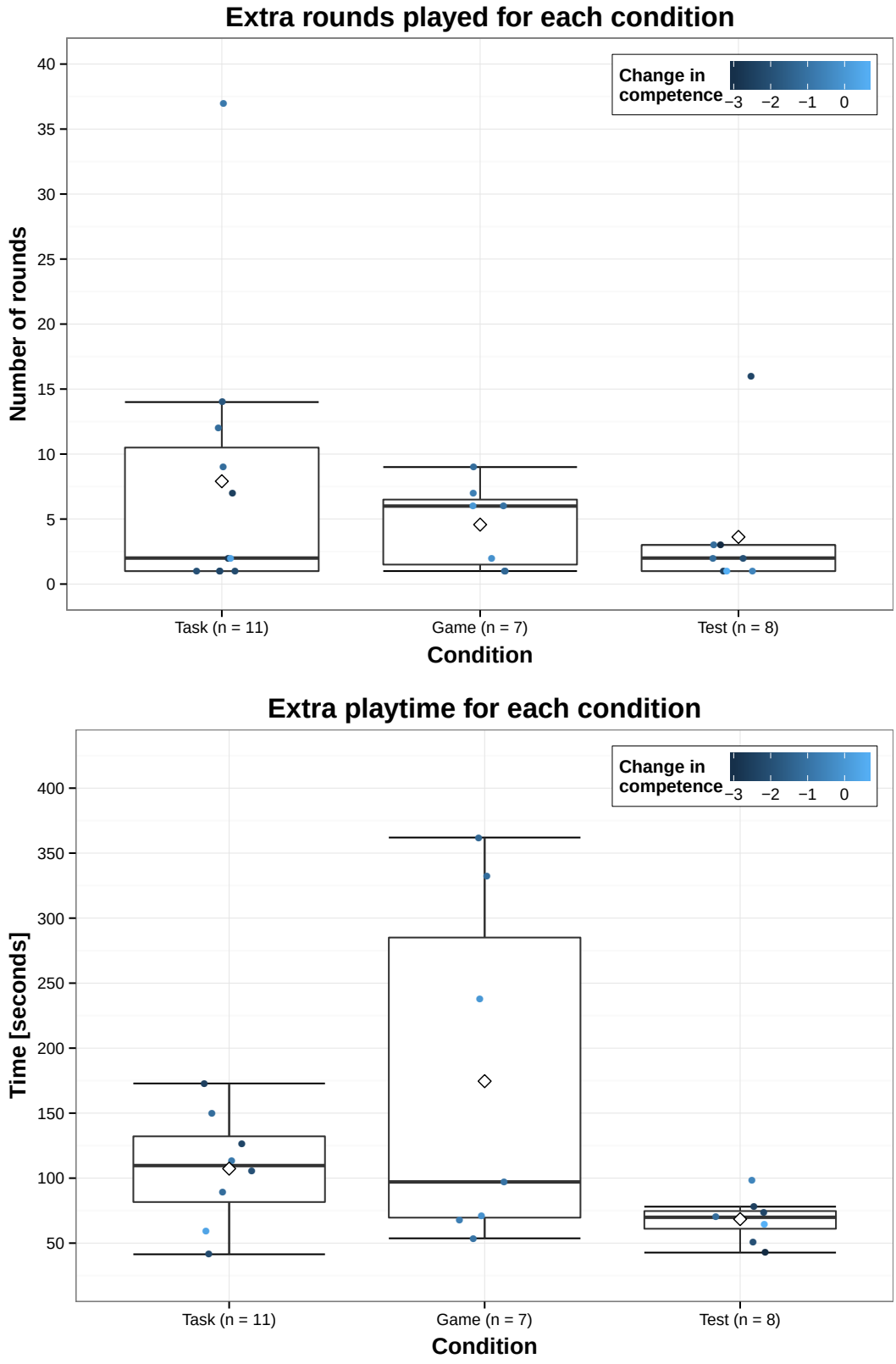

Figure 22. Number of extra rounds played and extra playtime by the participants in all three conditions.

Prediction of replay. To predict replay after completion of the experiment, a forward and backward selection algorithm selected the best predictors of the full model with change in autonomy, change in competence, change in intrinsic motivation, change in positive and negative affect as well as importance, value, and tension in a binomial logistic regression. The resulting model consisted only of change in autonomy as a significant predictor of replay. This model was then tested to assess its predictive power (see Figure 23 and Table 7). Results showed that, keeping the low sample size in mind, change in autonomy or rather lack of loss of perceived autonomy was able to predict whether people played additional rounds in the game. However, sensitivity (the fraction of returning players correctly predicted) was relatively low with .31, meaning that it was still very difficult to correctly predict which players stay for an additional round (or more).

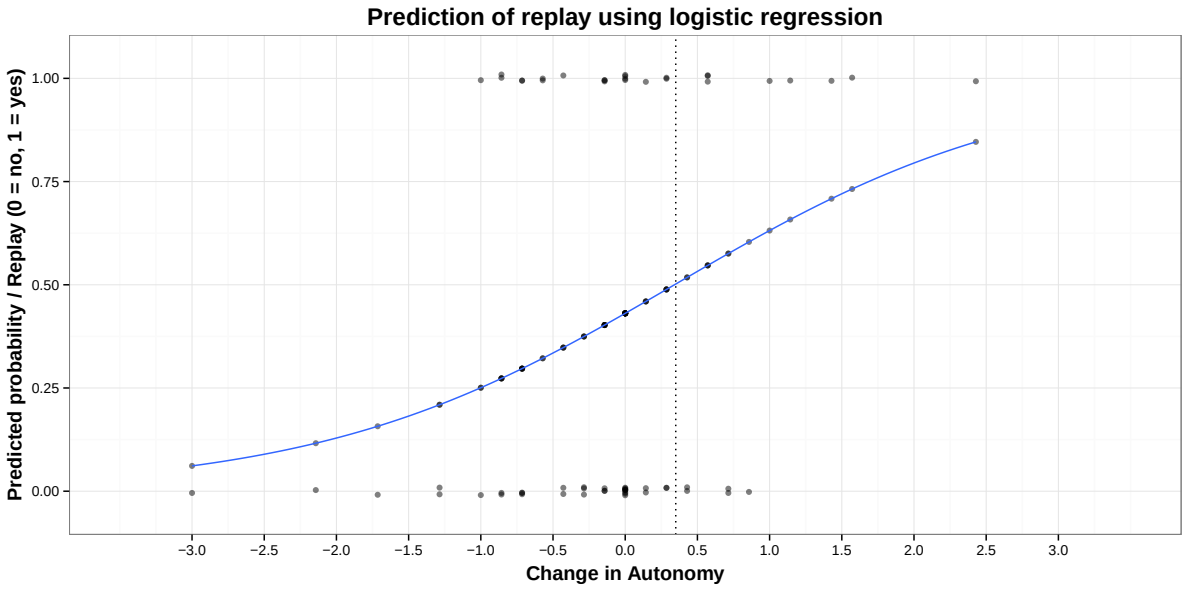

Figure 23. Predicted replay probabilities for perceived value and the observed replay behavior. The dots connected with the blue line indicate probabilities of replay. The observed behavior was coded 0 for no replay and 1 for at least one additional round played. The reduce over-plotting vertical jitter was applied. The dotted line indicates the cutoff after that replay probability is above .5.

Table 7

Binomial logistic regression statistics. Nagelkerke R^2 (Max rescaled R^2) = .114. Kendall's Tau – α = .119. Goodman-Kruskal Gamma = .258. Somers's D_{xy} = .241. c-statistic = 62.10%. Accuracy = .63, Precision = .62, Negative predictive value = .64, Sensitivity = .31, Specificity = .86.

Predictor	β	SE β	Wald's χ^2	df	p	e^{β} (odds ratio)	2.5%	97.5%
Constant	-0.2786	0.2682	1.1	1	0.300	0.7567	0.4439	1.2796
Change in autonomy	0.8167	0.3839	4.5	1	0.033	2.2630	1.1381	5.2413
Test			χ^2	df	p			
Overall model evaluation								
Likelihood ratio test			5.5761	1	0.018			
Score test			5.1347	1	0.024			
Goodness-of-fit test								
Hosmer & Lemeshow			4.7638	8	0.783			

Correlation analysis. Table 8 depicts intercorrelations among all dependent variables including all participants (63 + 4). Based on Kendall rank correlations several significant associations were found. Intrinsic motivation correlated with positive affect, value, autonomy, competence and importance, but not with any of the behavioral measures. Positive affect was positively associated with negative affect.

Table 8

Kendall rank correlations among autonomy, competence, intrinsic motivation, and affective state after playing (post) as well as score, and time (mean per round) including all participants. Values above the diagonal indicate bootstrapped 95% confidence intervals (1000 replications).

Experiment 2 ($N = 67$)	1	2	3	4	5	6	7	8	9	10
1. Autonomy (post)		(-.08, .24)	(.14, .46)	(-.11, .23)	(-.44,-.12)	(-.12, .20)	(-.12, .23)	(-.49,-.21)	(-.12, .21)	(-.05, .27)
2. Competence (post)	.08		(.11, .43)	(-.01, .29)	(-.34, .01)	(-.08, .27)	(.02, .30)	(-.25, .07)	(-.03, .28)	(-.15, .20)
3. IM (post)	.30***	.27**		(.30, .55)	(-.19, .15)	(.05, .37)	(.19, .48)	(-.25, .04)	(-.13, .19)	(-.14, .21)
4. PA (post)	.05	.14	.43***		(.04, .38)	(.22, .50)	(.23, .52)	(-.08, .24)	(-.16, .17)	(-.05, .26)
5. NA (post)	-.29**	-.17	-.02	.22*		(-.06, .26)	(-.02, .32)	(.44, .67)	(-.26, .05)	(-.24, .09)
6. Importance	.04	.09	.21*	.36***	.10		(.17, .46)	(.04, .32)	(-.05, .23)	(-.08, .26)
7. Value	.05	.16*	.34***	.38***	.15	.32***		(.03, .35)	(-.09, .22)	(-.15, .19)
8. Tension	-.35***	-.09	-.10	.09	.56***	.19*	.20*		(-.18, .13)	(-.24, .04)
9. Score	.05	.13	.03	0	-.11	.09	.06	-.03		(-.01, .30)
10. Time	.11	.03	.04	.11	-.08	.08	.02	-.10	.14	

Note. IM = Intrinsic Motivation, PA = Positive Affect, NA = Negative Affect. $p < .001$***; $p < .01$**; $p < .05$*

Comparison of Experiment 1 and Experiment 2

Autonomy and competence. For autonomy, no significant differences between the experiments were observed. For competence, a mixed-design ANOVA revealed a significant interaction between compensation and time, $F(1,168) = 23.211$, p$< .001$, $\eta^2 = 0.04$. Competence of participants in Experiment 2 reduced less strongly than those of participants in Experiment 1.

Intrinsic motivation. In Experiment 2 participants without compensations showed significantly less intrinsic motivation, $F(1,168) = 5.589$, p$< .05$, $\eta^2 = 0.03$ than participants of Experiment 1. There was no significant interaction between time and experiments observed, meaning that in both experiments participants' intrinsic motivation suffered equally.

Positive and negative affect. A significant main effect for experiment was found for positive affect, $F(1,168) = 46.580$, p$< .001$, $\eta^2 = 0.20$, showing that participants in Experiment 1 experienced more positive affect. But the significant interaction between time and experiments, $F(1,168) = 6.372$, p$< .05$, $\eta^2 < 0.01$, revealed that participants in Experiment 1 lost more positive affect. No significant differences between the experiments were observed for negative affect.

Importance, value, and tension. In Experiment 2 participants perceived the activity significantly less important than in Experiment 1, $F(1,168) = 48.671$, p$< .001$, $\eta^2 = 0.26$ than participants of Experiment 1. Levene's Test was significant $F(5,168) = 2.437$, $p < .05$, thus heteroscedasticity-corrected covariance matrices were used (HC3, based on Long & Ervin, 2000). As with importance, participants in Experiment 1 perceived the activity as more valuable than participants in Experiment 2, $F(1,168) = 11.479$, p$< .001$, $\eta^2 = 0.06$. However, no significant difference in tension was observed.

Score. Participants in Experiment 1 needed more steps to complete the required 5 rounds ($\bar{x} = 31.54$, $SD = 4.44$, $\widetilde{x} = 31$) than participants in Experiment 2 ($\bar{x} = 29.52$, $SD = 3.3$, $\widetilde{x} = 28.8$). The residuals deviated significantly from normal distribution, thus the variable score was reciprocal transformed. The Shapiro-Wilk normality tests and Q-Q plots showed only one significant deviation for the transformed variable (game condition in Experiment 2). A between-subjects ANOVA returned the following results: A significant main

effect for experiment, $F(1,178) = 8.261$, $p < .01$, $\eta_2 = 0.05$, and no significant experiment x framing interaction, $F(1,178) = 0.170$, $p = .84$, $\eta_2 < 0.01$. Levene's test was not significant, $F(5,168) = 1.547$, $p = .18$.

Time. During the 5 rounds of Flood It, participants who did not receive compensation invested more time ($\bar{x} = 99.91$, $SD = 49.71$, $\widetilde{x} = 91.19$) than those who did not ($\bar{x} = 75.3$, $SD = 32.44$, $\widetilde{x} = 69.32$). The residuals were not normally distributed, hence a reciprocal transformation was applied. Residuals still deviated significantly from normal distribution in the game and test condition of Experiment 1 and the task condition of Experiment 2 when tested with the Shapiro-Wilk normality test. Consequently, it was decided to analyze the transformed time variable with a between-subjects ANOVA and the untransformed time variable with a non-parametric permutation test (with 1000 permutations). Results of the ANOVA showed a significant main effect for experiment, $F(1,168) = 12.511$, $p < .001$, $\eta_2 = 0.07$, and no significant interaction between experiment and framing conditions, $F(1,168) = 0.440$, $p = .64$, $\eta_2 < 0.01$. The permutation test confirmed the results of the ANOVA (main effect experiment: $p < .01$, interaction experiment x framing: $p = .38$). Levene's test was not significant, $F(5,168) = 1.193$, $p = .31$.

Extra rounds and playtime. In Experiment 1, 31 (27.9%) of 111 participants played at least one additional round, in Experiment 2 this value increased to 26 (41.3%) of 63 participants. However, this difference was not significant, $\chi^2(1, N = 174) = 3.248$, $p = .071$. With the subsample of 57 participants who played at least one additional round, the time invested in those additional rounds was analyzed. Over all conditions, participants who did not receive compensation invested more time in the additional rounds ($\bar{x} = 187.8$, $SD = 230.3$, $\widetilde{x} = 93.02$) than participants who did not receive compensation ($\bar{x} = 97.04$, $SD = 82.27$, $\widetilde{x} = 65$). The number of observations were low when split up into the three conditions (range $7 - 11$). Thus it was decided to compare only the two studies. Residuals were not normally distributed, thus a reciprocal transformation was applied. Q-Q plots and Shapiro-Wilk normality tests showed no significant deviation. Results of a between-subjects ANOVA showed that the difference between the two experiments was significant, $F(1,55) = 4.309$, $p < .05$, $\eta_2 = 0.07$. Levene's test was not significant, $F(1,55) = 0.217$, $p = .64$.

Discussion Experiment 2

Need satisfaction, intrinsic motivation, and affect. As with Experiment 1, no significant differences were observed between the conditions in terms of autonomy, competence and intrinsic motivation. This shows that even in the voluntary context of an online study, framing Flood It as a game did not reduce the negative impact of failing. Similar to Experiment 1, participants' autonomy was not reduced by failure but a medium effect for competence ($\eta^2 = 0.33$) and a small effect ($\eta^2 = 0.03$) for a reduction in intrinsic motivation was observed. Participants in Experiment 2 reported less loss of competence than those in Experiment 1. Another interesting difference is that levels of intrinsic motivation were higher in Experiment 1. Additionally, more positive affect but also more loss of it after failing was reported in Experiment 1, possibly indicating that participants were more vulnerable.

Importance, value, tension, and additional rounds. In Experiment 2 participants perceived Flood It as less important and less valuable than in Experiment 1. This could also be explained by the compensation received by participants of the first experiment. As with Experiment 1, no differences between the conditions in perceived importance, value, and tension were observed. While participants in the task and game condition played more additional rounds and especially participants in the game condition invested more time in these additional rounds, the differences were not significant. However, the sample size was very low (range: $8 - 11$), hence, no final conclusion is possible.

Correlation of change in competence and intrinsic motivation. Analyses of the correlations between the variables showed that, consistent with Experiment 1, change in competence was only in the test condition associated with change in intrinsic motivation. This shows that even in the seemingly non-pressuring environment of a voluntary online experiment, participants' intrinsic motivation can be tied to competence with a test framing.

Prediction of replay. As with the first experiment it was tried to predict whether participants would engage in another round of Flood It after the study. Results of the binomial logistic regression showed that only change in autonomy was associated with replay. This means that participants who lost less autonomy were more likely to play at least one additional round. Given the low incidents of additional rounds the sensitivity of the model was

still relatively low (.31).

Intercorrelation pattern. The pattern of intercorrelations between the variables measured after failing in Flood It (see Table 8) showed a different picture than in Experiment 1. Score and time invested in the five required rounds were not correlated with any of the motivational nor affective variables. Intrinsic motivation was again associated with autonomy, competence, positive affect, importance, and value. Surprisingly, in Experiment 2 negative affect was correlated with positive affect. This seems to be paradox, but there is reasonable doubt about the bipolarity of the two affective states (Lindquist, Satpute, Wager, Weber, & Barrett, 2015). This might indicate that participants had conflicting feelings. However, this warrants further research.

General Discussion

No significant differences were observed between the different framing conditions in terms of intrinsic motivation. This contradicts the findings of Lieberoth (2015), who found that simply framing an activity as a game could increase intrinsic motivation. However, Lieberoth combined framing of an activity with game artifacts. Thus, it remains unclear from whence the effect originated. The results of this study suggest that the effects of game artifacts might have been the more influential. Participants in the study by Lieberoth reported low levels of intrinsic motivation compared to the experiments in this study. Thus, it is possible that in the present study, ceiling effects masked the influence of framing. Another explanation might be that framing in general works better with relatively uninteresting tasks. In all other motivational variables and negative affect, no differences were observed between the conditions. The behavioral data also showed no difference between the framing conditions. Framing Flood It as a game did not reduce performance compared to all other conditions, including the test condition. However, it has to be noted that in contrast to the study by Lieberoth, in which a group activity was gamified, the present study did not take place in a social context. Based on the theory of social facilitation, one could expect that a social context enhances performance in simple tasks and worsens performance in complex tasks (Geen, 1989).

Participants' intrinsic motivation and perceived competence suffered significantly from failing. Consequently, negative affect was increased, and positive affect decreased. However, in Experiment 1, positive affect of participants was only in the task condition significantly reduced. This indicates that providing a meaningful rationale for the task by giving more background information can buffer a loss of positive affect. In general, one could also expect positive effects of meaning on performance (Chandler & Kapelner, 2013), but in this study, participants' scores were almost equal in all conditions.

Failing in Flood It decreased reported intrinsic motivation, but it is unclear whether this happened because of loss in competence. Results showed that only in the test condition were change in competence and change in intrinsic motivation associated. In other words, failing affected intrinsic motivation independently from competence in all conditions except the test condition. It is conceivable that participants' scores and subsequently their perceived competence were not tied to intrinsic motivation when their performance was not linked to their self-image. Participants in the test condition, however, may have attributed their performance to their cognitive abilities, thus linking perceived competence with intrinsic motivation. It would therefore be interesting to investigate whether differences in the attribution of failure could be found between the conditions. The test condition was in this thesis framed as an assessment of cognitive abilities, resembling the brain-training game of Burgers et al. (2015). Burgers et al. found that negative feedback increased willingness for immediate replay, but this was not observed in the test condition of this study. A possible explanation could be that Flood It was perceived as unfair or too great a challenge to be able to repair competence with additional rounds. Vansteenkiste and Deci (2003) found that people can still stay motivated after failing to receive a reward if they are given positive normative feedback. The feedback given to participants in this study was not positive, but it incorporated information about the performance. Since autonomy was not negatively affected by the feedback, it is likely that it was perceived as informational in all conditions. It is conceivable that controlling feedback would have different effects in varying contexts, potentially even larger negative effects on intrinsic motivation.

More than a quarter of the participants decided to play at least one additional round.

Results of the binomial logistic regression showed that perceived value of the activity predicted replay in Experiment 1. However, sensitivity was low, and one could expect that it might be difficult to predict replay on a different sample with value. Indeed, in the second experiment, value was no longer a significant predictor. Autonomy was found to predict replay. Generally, this means that it is important to consider how valuable the activity is for the users when they are paid, such as in a corporate environment, and/or how autonomous their participation is. This supports the notion of Nicholson (2012), who argued that gamification should be meaningful to reduce possible negative effects on intrinsic motivation. If users identify with goals that are meaningful, they are more likely to produce autonomous, internalized behaviors because they can connect these goals with other values that they already hold (Nicholson, 2012). The participants who decided to not play additional rounds could have been driven by the following processes: 1) the activity was too challenging, and thus the skill-challenge balance was never achieved, but relatively high levels of intrinsic motivation contradict this notion; 2) the increased negative affect led to the desire to repair their mood or their depleted competence, which was not deemed possible with Flood It; or 3) participants wanted to complete the study as quickly as possible to do something else. However, it would have been interesting to allow the participants to chose between different difficulties of Flood It after the study to investigate whether decreased competence need satisfaction would lead to mood repair, as found in a study by Reinecke et al. (2012).

The comparison between the experiments revealed some interesting differences: Although paid participants in Experiment 1 reported higher levels of intrinsic motivation, they performed worse than people in Experiment 2. Participants in Experiment 1 did not refrain from completing additional rounds when compared to the second experiment, despite losing more competence and more positive affect than participants in Experiment 2. They also perceived Flood It as more important and valuable than participants of Experiment 2. Yet, participants in the second experiment performed better and invested more time in the activity. Similar to the first experiment, participants' autonomy did not decrease significantly in Experiment 2. Results showed that intrinsic motivation suffered equally in both experiments. This means that compensation did not affect the change in intrinsic motivation. However,

payment was not experimentally manipulated, so any differences between the experiments are not causal and need to be interpreted with caution.

Implications

This study is the first to show that different types of framing do not affect intrinsic motivation or behavior in a difficult puzzle-solving task. Giving users more information about the nature of the task or framing the task as a game or test seem to lower the reduction of positive affect when participants failed at said task. Additionally, value was positively and change in autonomy (i.e., how much perceived autonomy was lost during the task) was negatively associated with replay. This provides support for the importance of the voluntariness and meaningfulness of the task to keep people engaged, especially since this might lead to more internally regulated motivation in the long term (Deci & Ryan, 2004). The design of gamified applications should therefore be user-centered, autonomy-supporting and conveying value to the user. Designers of gamified applications should identify the types of failure that are possible in their applications and reduce their negative effects on value and autonomy. Even though one could argue that many gamified applications do not include a state of failure, most applications do quantify user input and actions. This quantification makes it possible to track progress — or lack of progress. Negative feedback in the form of not meeting one's own standard or a standard set by competing users is a form of failing that needs to be considered.

Keeping the quasi-experimental nature of the comparison of the two experiments in mind, one could argue that in Experiment 1, participants performed worse, invested less time, and were more severely affected by failure. Thus, from a motivational and behavioral perspective, it seems advisable to give users of a gamified application no monetary compensation for participating.

Limitations

The experimental paradigm reported in this thesis is not without limitations. It could be possible that the framing was not strong enough. However, special care was taken to word the instructions for the game and test conditions very similarly to the study by Ryan, Koestner,

and Deci (1991), and the instructions of the task with meaning condition were almost identical to the framing used by Mekler, Brühlmann, Opwis, and Tuch (2013a). Another possibility would be that Flood It was too interesting or motivating on its own (ceiling effect) to show differences because of the framing. It is also conceivable that Flood It incorporated too many game-like features (e.g., different rounds and the display of a score). Still, Flood It was chosen because it was abstract and did not rely on a specific content such as images of paintings (e.g., as in Mekler et al., 2013a). It still might be possible that the playful nature of Flood It and its game-like features made game-related associations more salient in all conditions, and thus the feedback was deemed equally informational. However, the different types of relationships among the key variables — intrinsic motivation and competence — as well as differences between the conditions in terms of loss in positive affect show that there was an effect of framing. This study only allows conclusions about the interaction of framing and failing. It would be interesting to decrease the difficulty to study possible effects of framing when succeeding in Flood It. The type of feedback was similar in all conditions, and one could expect that different types of feedback have distinct effects when participants are failing or winning, especially if they are more controlling.

Future Research

Future research should include also positive experiences and manipulate the outcome experimentally. It would be interesting to apply the experimental paradigm to less interesting or more routine tasks to study the effects of a "game frame" in other environments. Future studies could vary the amount of feedback given by the system to examine effects of different degrees of feedback in varying conditions. Autonomy did not decrease in any of the conditions; thus, it can be said that the received feedback was not of controlling nature (Deci & Ryan, 2004). However, it would be interesting to examine the effect of controlling feedback in various contexts because it might affect autonomy differently.

It would be worthwhile to study the effects of different contexts and failure in a social environment. The study of relatedness, the third need in SDT, was omitted in this study. It is possible that the awareness of others could affect performance and intrinsic motivation in

varying contexts differently. One promising approach would be the theory of social facilitation (e.g., Geen, 1989).

While no differences between the conditions in terms of gender and age were found, it could still be possible that demography of the sample might affect the perception of the different conditions. For example, Koivisto and Hamari (2014) found that social benefits of gamification were more effective for women. With the broad range of age and the approximately equal number of female and male participants, further analyses would be possible with the collected data.

To reduce the potential influence of a novelty effect, which was found to influence the perception of gamified applications (Koivisto & Hamari, 2014), it would be beneficial to study the effects of the different conditions over a longer time span. Additionally, it would be worthwhile to study whether autonomy supporting and value conveying gamification can shift external regulation towards internal regulation over time and subsequently increase intrinsic motivation.

Conclusion

Framing an activity as a game did not reduce the negative effects of failing in an activity when compared to a simple task, a test, and a task with additional meaning. Participants reported decreased positive affect in all conditions of Experiment 1, but significantly more loss was observed in the task condition. This means that more information about a task can reduce some of the negative effects. Additionally, loss in competence was only associated in the test condition with loss in intrinsic motivation, suggesting that in activities tied to one's self-image, support for competence might be able to reduce the negative impact of failure on intrinsic motivation. When the participants were paid, perceived value of the activity was found to be a weak but significant predictor for additional rounds played. Without pay, perceived autonomy was a relatively strong predictor for prolonged engagement even after the experiment. This shows that perception of value and particularly autonomy need satisfaction in gamification are important and can be linked to persisting behavior.

References

Abuhamdeh, S., Csikszentmihalyi, M., & Jalal, B. (2015). Enjoying the possibility of defeat: Outcome uncertainty, suspense, and intrinsic motivation. *Motivation and Emotion*, *39*(1), 1-10. doi: 10.1007/s11031-014-9425-2

Anderson, C. A., & Bushman, B. J. (2001). Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: A meta-analytic review of the scientific literature. *Psychological Science*, *12*(5), 353-359. doi: 10.1111/1467-9280.00366

Bavelier, D., Green, C. S., Han, D. H., Renshaw, P. F., Merzenich, M. M., & Gentile, D. A. (2011, 12). Brains on video games. *Nature Reviews Neuroscience*, *12*(12), 763–768. doi: 10.1038/nrn3135

Brühlmann, F. (2013). *Gamification from the perspective of self-determination theory and flow.* Retrieved from `https://bruehlmann.io/pub/fbruehlmann_bachelorthesis_final.pdf`

Burgers, C., Eden, A., van Engelenburg, M. D., & Buningh, S. (2015). How feedback boosts motivation and play in a brain-training game. *Computers in Human Behavior*, *48*(0), 94 - 103. doi: 10.1016/j.chb.2015.01.038

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin*, *140*(4), 980. doi: 10.1037/a0035661

Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, *90*, 123–133. doi: 10.1016/j.jebo.2013.03.003

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. London: Routledge.

Csikszentmihalyi, M. (1990). *The psychology of optimal experience*. New York: Harper Collins.

Deci, E., Koestner, R., & Ryan, R. (1999). A meta-analytic review of experiments examining

the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, *125*(6), 627. doi: 10.1037/0033-2909.125.6.627

Deci, E., Koestner, R., & Ryan, R. (2001). Extrinsic rewards and intrinsic motivation in education: reconsidered once again. *Review of Educational Research*, *71*(1), 1–27. doi: 10.3102/00346543071001001

Deci, E., & Ryan, R. (2004). *Handbook of self-determination research*. Rochester: University Rochester Press.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification". In *Proceedings of the 15th international academic mindtrek conference: Envisioning future media environments* (pp. 9–15). New York, NY, USA: ACM. doi: 10.1145/2181037.2181040

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

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using r*. London: SAGE Publications.

Filzmoser, P. (2005). Identification of multivariate outliers: a performance study. *Austrian Journal of Statistics*, *32*, 127–138.

Filzmoser, P., & Gschwandtner, M. (2015). mvoutlier: Multivariate outlier detection based on robust methods [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=mvoutlier` (R package version 2.0.6)

Geen, R. G. (1989). Alternative conceptions of social facilitation. In P. B. Paulus (Ed.), *Psychology of group influence* (pp. 15–51). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? – literature review of empirical studies on gamification. In *Proceedings of the 47th hawaii international conference on system sciences.* doi: 10.1109/HICSS.2014.377

Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and

academic performance. *Computers & Education*, *80*, 152 - 161. doi: 10.1016/j.compedu.2014.08.019

Hassenzahl, M., & Tractinsky, N. (2006). User experience - a research agenda. *Behaviour & Information Technology*, *25*(2), 91–97. doi: 10.1080/01449290500330331

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), pp. 65-70.

Juul, J. (2013). *The art of failure: An essay on the pain of playing video games*. Cambridge, Mass: MIT Press.

Kaiser Family Foundation. (2010). *Generation m2: Media in the lives of 8- to 18-year-olds.* Retrieved from `http://www.kff.org/entmedia/upload/mh012010presentL.pdf`

Koivisto, J., & Hamari, J. (2014). Demographic differences in perceived benefits from gamification. *Computers in Human Behavior*, *35*, 179–188. doi: 10.1016/j.chb.2014.03.007

Lepper, M., Sagotsky, G., Dafoe, J., & Greene, D. (1982). Consequences of superfluous social constraints: Effects on young children's social inferences and subsequent intrinsic interest. *Journal of Personality and Social Psychology*, *42*(1), 51. doi: 10.1037/0022-3514.42.1.51

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

Lindquist, K. A., Satpute, A. B., Wager, T. D., Weber, J., & Barrett, L. F. (2015). The brain basis of positive and negative affect: Evidence from a meta-analysis of the human neuroimaging literature. *Cerebral Cortex*. doi: 10.1093/cercor/bhv001

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*(3), pp. 217-224. doi: 10.1080/00031305.2000.10474549

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive science*, *5*(4), 333–369. doi: 10.1207/s15516709cog0504_2

May, D. R., Gilson, R. L., & Harter, L. M. (2004). The psychological conditions of

meaningfulness, safety and availability and the engagement of the human spirit at work. *Journal of Occupational and Organizational Psychology*, *77*(1), 11–37. doi: 10.1348/096317904322915892

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods*, *17*(3), 437. doi: 10.1037/a0028085

Mekler, E., Brühlmann, F., Opwis, K., & Tuch, A. (2013a). Disassembling gamification: The effects of points and meaning on user motivation and performance. In *Proceedings of the 2013 annual conference on human factors in computing systems (chi ea '13).* ACM, New York, NY, USA. doi: 10.1145/2468356.2468559

Mekler, E., Brühlmann, F., Opwis, K., & Tuch, A. (2013b). Do points, levels and leaderboards harm intrinsic motivation? an empirical analysis of common gamification elements. In *Proceedings of the first international conference on gameful design, research, and applications.* Stratford, Ontario, Canada. doi: 10.1145/2583008.2583017

Mekler, E., Brühlmann, F., Tuch, A., & Opwis, K. (in press). Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior*. doi: 10.1016/j.chb.2015.08.048

Nicholson, S. (2012). A user-centered theoretical framework for meaningful gamification. *Games+Learning+Society 8.0.*

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological methods*, *8*(4), 434. doi: 10.1037/1082-989X.8.4.434

Reeve, J., & Deci, E. L. (1996). Elements of the competitive situation that affect intrinsic motivation. *Personality and Social Psychology Bulletin*, *22*, 24–33. doi: 10.1177/0146167296221003

Reinecke, L., Tamborini, R., Grizzard, M., Lewis, R., Eden, A., & David Bowman, N. (2012). Characterizing mood management as need satisfaction: The effects of intrinsic needs on selective exposure and mood repair. *Journal of Communication*, *62*(3), 437–453. doi: 10.1111/j.1460-2466.2012.01649.x

Rigby, S., & Ryan, R. M. (2011). *Glued to games: How video games draw us in and hold us*

spellbound: How video games draw us in and hold us spellbound. Santa Barbara, CA: ABC-CLIO.

Ryan, R. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of personality and social psychology*, *43*(3), 450. doi: 10.1037/0022-3514.43.3.450

Ryan, R., & Deci, E. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, *25*(1), 54–67. doi: 10.1006/ceps.1999.1020

Ryan, R., Koestner, R., & Deci, E. (1991). Ego-involved persistence: When free-choice behavior is not intrinsically motivated. *Motivation and Emotion*, *15*(3), 185-205. doi: 10.1007/BF00995170

Ryan, R., Rigby, C., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion*, *30*(4), 344–360. doi: 10.1007/s11031-006-9051-8

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? reinvestigating the robustness of anova against violations of the normal distribution assumption. *Methodology*, *6*(4), 147-151. doi: 10.1027/1614-2241/a000016

Schmierbach, M., Chung, M.-Y., Wu, M., & Kim, K. (2014). No one likes to lose. *Journal of Media Psychology*, *26*(3), 105-110. doi: 10.1027/1864-1105/a000120

Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, *74*, 14 - 31. doi: 10.1016/j.ijhcs.2014.09.006

Thom, J., Millen, D., & DiMicco, J. (2012). Removing gamification from an enterprise sns. In *Proceedings of the acm 2012 conference on computer supported cooperative work* (pp. 1067–1070). doi: 10.1145/2145204.2145362

Thompson, E. R. (2007). Development and validation of an internationally reliable short-form of the positive and negative affect schedule (panas). *Journal of cross-cultural psychology*, *38*(2), 227–242. doi: 10.1177/0022022106297301

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass.: Addison-Wesley.

Valadez, J. J., & Ferguson, C. J. (2012). Just a game after all: Violent video game exposure

and time spent playing effects on hostile feelings, depression, and visuospatial cognition. *Computers in Human Behavior*, *28*(2), 608 - 616. doi: 10.1016/j.chb.2011.11.006

Vansteenkiste, M., & Deci, E. L. (2003). Competitively contingent rewards and intrinsic motivation: Can losers remain motivated? *Motivation and emotion*, *27*(4), 273–299. doi: 10.1023/A:1026259005264