From Objective to Subjective Difficulty Evaluation in Video Games

Thomas T.C.C. Constant, Guillaume Levieux, Axel Buendia, Stéphane Natkin

From Objective to Subjective Difficulty Evaluation in Video Games

Thomas Constant, Guillaume Levieux, Axel Buendia, and Stéphane Natkin

Conservatoire National des Arts et Métiers, CNAM-Cédric
292 Rue St Martin, FR-75141 Paris Cedex 03
first.last@lecnam.net

Abstract. This paper describes our research investigating the perception of difficulty in video games, defined as players' estimation of their chances of failure. We discuss our approach as it relates to psychophysical studies of subjective difficulty and to cognitive psychology research into the overconfidence effect. The starting point for our study was the assumption that the strong motivational pull of video games may lead players to become overconfident, and thereby underestimate their chances of failure. We design and implement a method for an experiment using three games, each representing a different type of difficulty, wherein players bet on their capacity to succeed. Our results confirm the existence of a gap between players' actual and self-evaluated chances of failure. Specifically, players seem to underestimate high levels of difficulty. The results do not show any influence on difficulty underestimation from the players gender, feelings of self-efficacy, risk aversion or gaming habits.

Keywords: User Modelling · Affective HCI, Emotion, Motivational Aspects · Tools for Design, Modelling, Evaluation · Fun / Aesthetic Design

1 Introduction

Jesper Juul proposed defining a video game as "*a rule-based formal system with a variable and quantifiable outcome, where different outcomes are assigned different values, the player exerts effort in order to influence the outcome, the player feels attached to the outcome, and the consequences of the activity are optional and negotiable*" [1]. In this definition, the player *exerts effort* to influence the outcome, emphasizing the fact that a video game must have a certain level of difficulty to be considered as such.

Many authors acknowledge challenge as one of the most fundamental aspect of a video game's inherent appeal. Malone proposes three features of computer games that render them captivating: challenge, curiosity and fantasy [2]. In his model, challenge is directly related to game difficulty and corresponds to the uncertainty of the player reaching the game's goals. Lazzaro proposes a four-factor model, where *Hard Fun* relates to the feeling of overcoming difficult tasks [3]. Sweetser et al also see challenge as one of the most important parts of their Game Flow framework [4]. This work builds upon Mihaly Csikszentmihalyi's

Flow Theory [5], which deals with the properties of activities that have a strong, intrinsic ability to motivate. Csikszentmihalyi's research has found that these activities provide *perceived challenges, or opportunities for action, that stretch (neither overmatching nor underutilizing) existing skills* [5]. A study of a large population of players of two commercial games confirmed that players prefer specific levels of difficulty [6]. Ryan et al have also studied intrinsic motivation, applying their Self-Determination Theory to video games. They show how enjoyment relates to the feeling of *competence* - which relies on an optimal level of challenge - and thus to game difficulty [7]. Finally, Jesper Juul provides insight on how failure, and thus difficulty, is one of the core aspects of video game enjoyment and learning progression [8, 9].

In order to foster and maintain a player's' motivation, it is therefore essential to correctly set the difficulty of a video game. To do this, one can either provide different difficulty settings for the player to select or use an algorithm that adapts difficulty in real time to match the game designer's theoretical difficulty curve to the player's player skill level [10–12].

Both methods require a prior evaluation of the game's difficulty. For this the game designer might provide a heuristic which may or may not accurately express the game's difficulty. Alternatively, sensors may be used to estimate workload or affective state, but this method is currently only feasible in a lab setting and its efficacy is still being studied [13, 14]. We could also try to estimate players' chances of failure [15]. Each of these approaches provide insight into a specific aspect of a game's difficulty.

Difficulty is in itself a complex notion. We can draw distinctions between *skill-based* difficulty, *effort-based* difficulty [16], and between *sensory*, *logical* and *motor* difficulty [15, 17]. Moreover, video games are created for an aesthetic purpose, evoking specific emotions in the player [18]. Thus, we must draw a fundamental distinction between *objective difficulty* and *subjective difficulty*. Objective difficulty is estimated directly by observing gameplay variables and events, while subjective difficulty is a psychological construct of the player. When adapting a game's difficulty, especially when using a dynamic difficulty adjustment (DDA) algorithm, we are relying on an objective estimation of difficulty, which may be quite different to what the player actually feels while playing the game.

In this paper we present our work on studying the relationship between subjective and objective difficulty in the context of video games. First, we review various studies on both subjective and objective difficulty estimation, looking first at the psychophysical approach of perceived difficulty, then at cognitive psychology research on overconfidence. We then introduce our own method for measuring objective and subjective difficulty. In this method, objective difficulty is modeled using a mixed effects logistic regression to estimate the player's actual chances of failure for a given challenge. We defined subjective difficulty as the players' estimation of their chances of failure, which we gathered using an in-game betting system. This is followed by a description of the three games we developed for this study that allowed us to separate out logical, motor and sensory gameplay. Lastly, we present and discuss our results.

2 Psychophysical approach to subjective difficulty

Many studies have tried to clarify the link between the subjective and objective difficulty of various tasks: Raven's progressive matrices, number memorization, visual letter search, wire labyrinth tasks [19, 20], Fitts' tapping task [21, 22], throwing darts at a moving target [23], rock climbing [24], reaction time, even while riding a bike [19, 25]. All these experiments take a psychophysical approach, trying to estimate the link between objective difficulty as a stimulus and subjective difficulty as a perception or evaluation of this stimulus.

These studies use various techniques to estimate objective difficulty, and often tend to draw a distinction between objective difficulty and performance. For all of the Fitts' tapping tasks, authors use Fitts's law [26] as a measure of objective difficulty, and time as a measure of performance. When such a law is not available, however, they rely solely on performance, e.g. response time or success frequency [23, 20], or they select a variable highly correlated with perceived difficulty such as, in the case of the rock-climbing experiment, electromyographic data from a specific muscle [24]. In addition, in these studies objective difficulty is never assessed with regard to each subject's abilities, but across all or a few subgroups of subjects. In our research, we do not rely on any specific objective difficulty estimation but follow a more generic approach that allows for cross-game comparisons. We estimate a mapping between the challenge's variables and the player's chance of failing it. We also use a mixed effect model that takes into account each player's abilities.

In the psychophysical studies, subjective difficulty is assessed using a free scale. Very often, a reference value is given to the subject, e.g. a subjective difficulty of 10 for a specific task [21]. Deligniere proposes the DPE-15 scale, a 7-point Likert scale with intermediate values, as a more convenient and comparable measure [23]. In our experiment, we integrated this measure with gameplay and used a specific 7-point scale, as described in section 4. To avoid the problem of subjective interpretation of the notion of difficulty, we concentrated on the success probability, as estimated by the player.

Except in Slifkin & Grilli [21], all subjective evaluations were carried out at the end of each challenge, often after having repeated the challenge many times. We considered that to understand what the player feels during play, rather than while reflecting on a past game session, it might be useful to look at the player's evaluation of current difficulty during each challenge. As our measure of subjective difficulty is an estimation of the chance of failure, it can be integrated into the gameplay and thus be repeated more often without pulling the player out of the game (see section 4).

3 Overconfidence and the Hard/Easy Effect

We define subjective difficulty as the player's own evaluation of their chance of failure. This evaluation is a complex cognitive process, often rushed, based on the interpretation of incomplete information about the game state, based on in-game performance feedback as well as assessments of the player's own knowledge

and skills with respect to a specific challenge. Cognitive psychology research on judgmental heuristics looks at how this kind of reasoning can be biased, and can help us understand how players may have wrongly evaluated their chances of success.

Heuristic approaches to judgment and decision-making have opened up a vast field of research into explaining human behavior in the context of uncertainty. Kahneman & Frederick [27, 28] consider that, when confronted to a complex decision, people substitute one attribute of the decision with a simpler, more accessible one, in order to reduce cognitive effort. In some cases, the use of judgmental heuristics can lead to fundamental errors, called *cognitive biases* by Kahneman & Tversky [29].

The overconfidence effect is one of these biases. Well-studied in the domain of finance, this behavior relies on a surrealistic evaluation of our own knowledge and skills, leading to an overestimation of our abilities or those of others [30–34]. Overconfidence seems particularly useful to study in relation to video games as they are essentially built with the motivation of the player in mind. The self-efficacy theory of motivation states that having a strong feeling of confidence in one's future chances of success is a key aspect of motivation [35]. Video games that feature a well-crafted difficulty curve can manipulate players' perception of their chances of success to keep them motivated.

Overconfidence has already been studied in many games. In a game of bridge, beginners or amateurs players can misjudge both their performances and play outcomes [36]. The same effect has been noticed in other games where novice players show an inferior ability to predict their odds of winning during poker tournaments [37], and games of chess [38], and in gambling games [39, 40]. To summarize: the overconfidence effect appears when the players have a limited knowledge of the game. This applies to any type of game, whether it be a pure game of chance such as a slot machine, or a skill-based game like chess.

There are many situations and cognitive biases that influence a player's overestimation or underestimation of their chances of success. These include: level of expertise [41, 42], the *gambler's fallacy* [43, 44], the *hot hand bias* [45, 44], the *illusion of control* [46, 47] and the *hard/easy effect*. While all these aspects of overconfidence are worth studying in the context of video games, for our research we chose to focus on the *hard/easy effect*. The *hard/easy effect* specifies that for low and high levels of difficulty, decision-makers fail to estimate the true difficulty of a task [48]. For low levels, they underestimate their chances of success; for high levels they overestimate [41, 33].

Using the hard/easy effect as our starting point, our research focused on two main aspects. First, from a methodological point of view, we wanted our experiment to simulate as closely as possible the experience of a real video game. For this reason we used a dynamically adjusted difficulty beginning at a low difficulty level. In addition, instead of evaluating player confidence by explicitly asking them if they felt confident on a percentage scale, we used a betting mechanism integrated into gameplay. In this way we avoided to breaking player immersion. Second, our research distinguished between three types of difficulty.

We used three different games, each focusing on a specific type. We describe our experiment in the following section.

4 Experimentation

As we have previously emphasized, video games feature different types of difficulty. In our experiment we sought to assess them separately, with a view to distinguishing between the various facets of video games. Using Levieux et al's [17, 15] approach we considered three categories of difficulty in games: sensory, logical and motor. Sensory difficulty relates to the effort needed to acquire information about the game state. Logical difficulty corresponds to the effort needed to induce or deduce, from the available information, the solution to a problem in terms of action(s) to perform. Lastly, motor difficulty relates to the physical agility needed to perform these actions. To realize an accurate analysis of the player's behavior for each of these types, the experiment was split between three custom-designed games, all played within a single program.

For this experiment, we chose a general, practical approach wherein we estimated the probability of a player failing a specific challenge relative to their current skill level [15]. Our definition of difficulty builds upon Malone's definition of challenge as a source of uncertainty in video games [2]. Uncertainty in success or failure is what Costikyan also calls *uncertainty of outcomes* [49]. We follow these authors and consider the difficulty as such. We directly ask players to evaluate their success chances, and thus avoid to use the term "difficulty" which has a less accurate meaning. Additionally, we were able to make a distinction between logical, motor and sensory tasks by separating them into three different games (described in section 4.3). In order to maximize player motivation and create an experience that was as close as possible to a real game, the system dynamically adapted difficulty based on analyzing player success or failure. Many games use dynamic difficulty adaptation, including racing games (e.g.rubberbanding in *Mario Kart*), and RPGs (e.g. *Fallout*) or FPSs (e.g. *Unreal Tournament*) where the difficulty in defeating an opponent depends on the player's level. Games without dynamic difficulty adaptation may use a predetermined difficulty curve based on the mean level of the players. Few games use completely random difficulty, but even in these (e.g. *FTL*, *The Binding of Isaac*), there is a global progression. Thus, while randomness would be more convenient for statistical analysis it would be of limited use from a game design perspective.

In addition, to avoid any memory bias on the past challenge and to better monitor the actual feeling of the player, we measured subjective difficulty during the game session rather than with post-experiment questionnaires [50]. To do this without pulling the player out of the game, we used a betting system, which we describe in the following section.

4.1 Measuring Subjective Difficulty

Our proposition takes cognitive psychology tools for measuring overconfidence and integrates them into gameplay. Our goal was to avoid disrupting the game

session in order to maintain a high level of engagement and motivation. The measure is taken before the player's actions, as a pre-evaluation, but after having given them all the elements necessary to make their judgment. We used a betting system based on a 7-point Likert scale, which was integrated into the game progression and in this way tied to the player's score. If the player won, the amount they bet was added to their score; if they lost the amount was subtracted. This motivated the player to think carefully about their self-evaluation. An in-game question served to instruct the player on how to bet and reminded them to take care in assessing their own performance, thus their own confidence.

Measurement of subjective difficulty was based on the player's bet, designated as D_{subj}. With b being the bet value we used the formula $D_{subj} = 1 - \frac{b-1}{6}$ to get the estimated chances of failure.

4.2 Measuring Objective Difficulty

As Levieux et al define it [17, 15], the objective difficulty of a challenge can be estimated based on players' failures and successes in completing it. In order to take into account personal differences, we estimated the objective difficulty for each challenge using a mixed effects logistic regression [51]. The time and difficulty parameters of each challenge (e.g. cursor speed, number of cells) were used as fixed effect parameters, and we added random intercepts. We used a mixed model throughout repeated evaluations of the same subject. The random intercepts gave us a coefficient for each player that we used as a global evaluation of their performance level. The gap between the players' objective difficulty and their evaluations of their odds of success is called the *difficulty estimation error*. The designs of the three games, each one based on a difficulty type - logical, motor or sensory, are detailed in the next section.

4.3 Game Descriptions

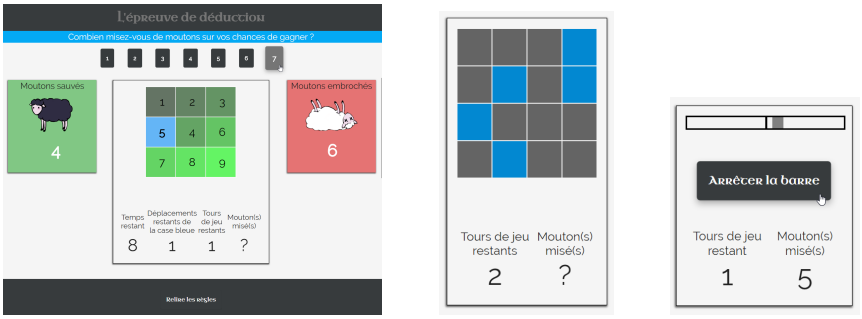
Our experiment was based on the observation of players' betting in relation to the three dimensions of difficulty. Each dimension was represented by a specific game, described below, for which all the adjustment variables for the challenges are pre-established and common for all players. An initial series of playtests was also conducted with the target audience in the same settings used during the experiments for the purpose of gameplay calibration.

A brief story was included in order to enhance player motivation and to provide a narrative justification for the betting system. In the game universe, the player must save citizens of a mysterious kingdom who have been transformed into sheep by a local sorcerer. The player challenges the sorcerer during three tests, one for each kind of difficulty. The player's sole objective is to save as many sheep as possible. In turn,each game is an opportunity for the player to save doomed citizens by betting between one and seven sheep against their odds of winning.

All three games have a common user interface except for the central frame which depends on each sub-game (figure 1). All important information is displayed at the bottom of this frame:the number of remaining turns, the global score, and, in the case of the logical game, the remaining number of actions. Directives are placed below the main title on a colored banner: blue for directives, red for corrective feedbacks. A rules reminder is accessible at the bottom of the screen.

Feedback is provided throughout the game, at the end of each turn. Positive (on green background) and negative (on red) feedback is displayed on both sides of the screen, allowing the player to constantly follow their number of saved and lost sheep. Sound effects accompany this bleating for a saved sheep, a sorcerer's mocking laugh for a lost one. Animations are used to provide a more stimulating in-game interface.

For each game we modified the difficulty using a *difficulty parameter*. This parameter varied from 0 to 1 and was used to interpolate gameplay parameters, which we define in the following section. The difficulty parameter started at 0.2 and increased or decreased by 0.1 after each turn, based on the player's success or failure.



(a) Logical task (b) Sensory task (c) Motor task

Fig. 1: Game interface for the logical, sensory and motor tasks. The logical task is shown using the whole user interface, while for the motor and sensory task only the center frame was used. Screenshots were taken for the easiest levels of difficulty.

Logical difficulty The logical task is based on a well-known sliding-puzzle game. The player must restore the numerical order of a 300-pixel-wide grid composed of 9 squares. The fifth square, originally placed on the middle of the grid, is the only one that can be moved. This square can only be moved by switching its position with an adjacent square (figure 1a). At the beginning of each turn, before displaying the grid, the fifth square is randomly moved several times and the mixed up grid is displayed for 20 seconds before disappearing. The player has all the information required to place a bet: the remaining time is visible and the number of moves is specified. After betting, the grid will reappear

and the player can begin to move the fifth square to restore the numerical order. The difficulty parameter allows us to adapt the difficulty by changing the number of steps during the randomization of the grid linearly from 1 to 11 steps.

Sensory difficulty For sensory difficulty, we designed a 300-pixel-wide grid composed of multiple squares (figure 1b). At the end of a countdown timer, five of them fade out during a limited time that we can approximate as follows, with t being the fade-out time and d as the difficulty parameter: $t = d^2 - 0.24d + 1.2$[1]. The player's task is to find the squares that have faded out by clicking on the grid. These squares are displayed in blue while the others remain in gray in order to avoid any color perception bias. The winning squares are shown after making a bet, over the player's squares selection. By doing this we wanted to induce a near-miss effect, allowing the player to see if they selected all, some or none of the winning squares. The countdown timer is set to 3 seconds. The number of squares varies with the difficulty of the task: when the player wins, the grid gains one square on each side. Meanwhile, the surface of the grid remains the same, meaning the squares become smaller after a winning round. For the maximum difficulty level the grid measures 11 by 11 squares; the minimum level grid size is 4 by 4. The difficulties in between are linearly interpolated using the difficulty parameter. Random locations are used for winning squares to avoid the most simple patterns, thus minimizing pattern-induced variations of difficulty for a specific difficulty parameter value. For example, for a 5x5 grid any adjacent winning squares are forbidden.

Motor difficulty The motor difficulty game is a basic and common reflex-based task. A cursor goes back and forth along a horizontal segment at a linear speed. The player must stop the cursor when it covers a black mark at the center (figure 1c). They can only stop the cursor by clicking on a button. Before they do this, the player must bet on their chance of success. This evaluation is not timed. Difficulty is based on the cursor's speed, which ranges linearly from 100 to 400 pixels per second. The sliding area is 320 pixels wide, the cursor is 15 pixels wide, and the black target 2 pixels wide.

Protocol consistency These three tasks, although different in nature, share a similar protocol and always provide the player with the elements needed to evaluate difficulty. For the motor task, players can observe the moving cursor before betting. For the logical task, the game displays the number of moves and lets the player view the problem for a fixed duration. For the sensory task, in which visual memory is crucial, the player selects tiles to solve the problem, but without any feedback before betting. Initial playtests showed that the task was very frustrating if the player had to stop focusing on the grid for betting without selecting the tiles. Each game has specific gameplay, as each one focuses on a specific dimension of difficulty. Results can thus be compared between games while taking account of gameplay differences.

[1] This equation is a quadratic regression of the fade-out time. In the game, the color is incrementally modified during the game loop, but plotting this equation is much clearer than reading the color update code.

4.4 Procedures

Our experiment was conducted in Paris at the *Cité des sciences de l'industrie*, a national museum dedicated to science and critical thinking, during a school vacation period. The target audience was young volunteers, both gamers and non-gamers. Some who were invited to participate declined saying they lacked gaming experience or were not interested in taking part in a science experiment.

Nine laptops, all with the same configuration, were used in an isolated room. Each one had a mouse and a headset. The main program runs on a web browser, and was developed with JS, HTML5 and CSS. Participants were informed of the game's goal - to save as many sheep as possible - and of the duration of the experiment, approximately 40 minutes, questionnaire included. They were told not to communicate during the session. Before playing, the participants had to fill an online questionnaire used to create several different user profiles:

- **A gaming habits profile**, based on the amount of time that participants spent playing board games, video games (including social games) and gambling games.
- **A self-efficacy profile**, based on General Self-Efficacy scales [52, 53] and adapted to video games situations. This part of the questionnaire was only accessible for the participants who answered yes to the question "*Do you consider yourself as a video game player?*" in the gaming habits section. The purpose of this was to check for any negative or positive effects of the participant's gaming ability on their self-estimation of their confidence.
- **A risk aversion profile**, based on Holt and Laury's Ten-Paired Lottery-Choices [54] in order to evaluate the impact of risk incentive on the player's confidence.

Our three games, each focusing on a different task, are all accessed and experienced within the context of a single software application (the "program"). The program's user flow is the same for all players:

- A **prologue** introduces the story before a random selection of the 3 tasks.
- A specific page presents **the rules** of the task before it starts. Players can take as much time as they want to understand them.
- Each task lasts 33 turns. The first 3 turns are used as a practice phase. At the end of this practice phase the score is reset to 0.
- The **turn progression** is identical for all the tasks. First, players have to observe the current game state in order to evaluate the difficulty. Then, they have to bet from 1 to 7 sheep on whether they will succeed. The same question is always asked of the player: "*How many sheep are you betting on your chances of winning?*". This question allowed us to estimate the player's perception of their chances of failure. By validating the bet, the system unlocks the game and players can try to beat the challenge. The result is presented on screen and the score is updated at the same time. Then a new turn begins with an appropriate adjustment to the difficulty level: when players win, the difficulty increases; when they lose, the difficulty decreases.

- After each task, **a game hub** allows the player to check their progression and score, and to progress to another task.
- After completing the 3 tasks a brief narrative **epilogue** announces the player final score: the total number of sheep won and lost.

To avoid any order effect, task selection is randomized. The best score of the day was written on a board, visible to the players. At the end of each turn the designed difficulty of a challenge, the player's bet and their score, were logged to CSV files.

5 Results

A total of 80 participants played the games. While some left the experiment before the end, we kept the results for all completed games, giving us a total of 6990 observations. For each task we remove outliers, such as players who did not use the betting system to perform a self-assessment, always placed the same bet, or players with outlying performance. A very low score may reflect some user experience issues, and some players took advantage of the adaptive difficulty system in order to maximize their score by deliberately losing with a low bet then by placing a high bet on the next easier challenge and so on. Nine outliers were removed: one from the motor task, three from the perceptive task, and six for logical one. We thus removed 300 observations from the dataset.

5.1 Modeling objective difficulty

As explained in section 4.2, we performed a logit mixed effect regression to evaluate objective difficulty. For each task, we reported the conditional R^2, i.e. using both fixed and random effects [55] and evaluated the model by performing a 10-fold cross-validation, using our model as a binary predictor of the challenge outcome (figure 2).

Parameters / Tasks	**Logical**	**Motor**	**Sensory**
Difficulty parameter	4.88 $(p < 2e-16)$***	3.23 $(p < 2e-16)$***	9.1 $(p < 2e-16)$***
Time	-1 $(p = 2e-6)$***	-0.46 $(p = 0.0051)$**	-0.37 $(p = 0.0454)$*
σ(random intercepts)	1.24	0.83	0.76
R^2	0.48	0.28	0.42
Cross Validation	0.66	0.61	0.69

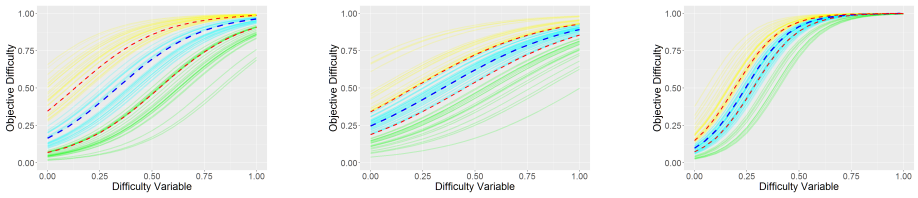
Fig. 2: Modeling objective difficulty for each task: logit mixed effect regression results for difficulty and time over failures.

As can be seen in figure 2, the difficulty parameter is always highly significant, and has the strongest effect on failure probability, especially for the sensory task.

This means that we were indeed manipulating objective difficulty by changing this parameter.

The effect of time is always negative and significant. This means that if the difficulty parameter stays constant, objective difficulty seems to decrease overtime. This might indicate that players are actually learning as their success rate improves overtime for a given difficulty parameter value. The time effect is strongest for the logical task (-1), which is coherent with the fact that the player should learn more from a logical problem than from a purely sensory motor one (respectively, -0.46 and -0.37). Also, it may be noted that we have the highest standard deviation of random intercept for the logical task, which means that inter-individual differences are the highest for this task.

The link between the difficulty parameter and the objective difficulty of the game can be plotted to better understand each challenge difficulty dynamics. We chose to plot objective difficulty over the difficulty parameter at time $t = 0$. We also used the random intercept to separate the player into three groups of levels using k-means (figure 3).



(a) Logical task (b) Motor task (c) Sensory task

Fig. 3: Objective difficulty for each task at $t = 0$. The blue dashed line represents median players, red dashed lines show the first and last quartiles. The least competent players are in yellow, average players in cyan and best players in green.

Curves in figure 3 show information about our design of each task's difficulty. We can see that the logical task is the most balanced, with objective difficulty being the closest to the difficulty parameter value. The motor task is is a bit too hard for low difficulty levels: objective difficulty is around 0.25 where the difficulty parameter is 0. Also, the sensory task should vary more slowly: objective difficulty reaches maximum when the difficulty parameter is only 0.5.

Figure 4 shows the progression of objective difficulty during the game. The curves confirm the balancing of each task and the efficiency of the difficulty adaptation system, as the players reach the average objective difficulty level (0.5) in all cases. The logical task starts at 0.2 for average players and goes up. The motor task is too hard at the beginning, and thus bad players see a decrease in difficulty overtime. The sensory task shows a "wavy" pattern, which may be related to the fact that the difficulty is less stable for this game. Indeed, the difficulty parameter varies by 0.1 of a step for all tasks, but as the maximum

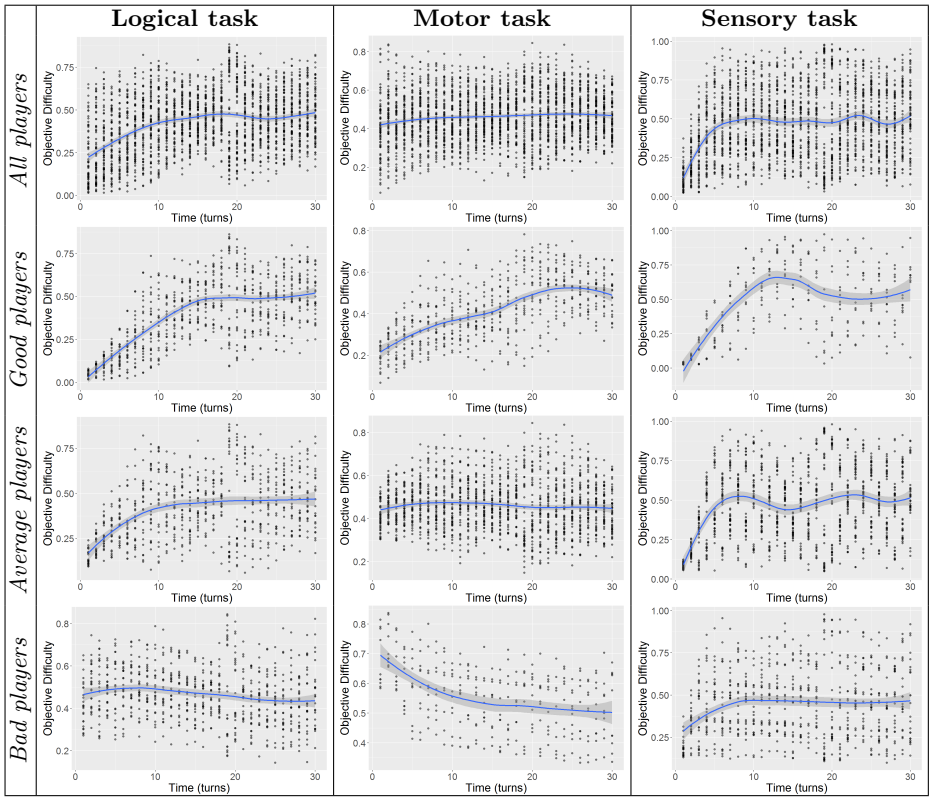


Fig. 4: Progression of objective difficulty overtime for all tasks and players during the entire play session. The blue line represents median players, dots represent the observations for each turn.

objective difficulty is already reached at 0.5 it varies approximately twice as fast as the logical one.

Overall, the objective difficulty model is the weakest for the motor task with a low conditional R^2 (0.28) and the lowest prediction accuracy (0.61). R^2 and prediction accuracy are higher for the logical ($R^2 = 0.48$, $accuracy = 0.66$) and sensory tasks ($R^2 = 0.42$, $accuracy = 0.69$).

5.2 Differences between objective and subjective difficulty

To investigate the differences between objective and subjective difficulty, we separate the data into 16 equally sized bins using the objective difficulty as estimated by the mixed effect model. In each bin, we compute, for each player, the mean subjective difficulty. We thus have only one value by player in the bin, and each observation is thus independent from the others. Then, for each bin, we test the null hypothesis that the bin's median subjective difficulty is equal

to the objective difficulty at the center of the bin's interval. We use a Wilcoxon Signed Rank Test and computed the 95% confidence interval (red bars) and pseudo median (black dot and triangles), plotted in figure 5. We show only the pseudo median and confidence intervals for bins with enough samples to run the Wilcoxon signed rank test. The blue line represents our null hypothesis, where objective difficulty equals subjective difficulty. These results allow us to safely reject the null hypothesis for each median represented by an empty triangle in the plots, where the Wilcoxon signed rank test p-value is lower than 0.05.

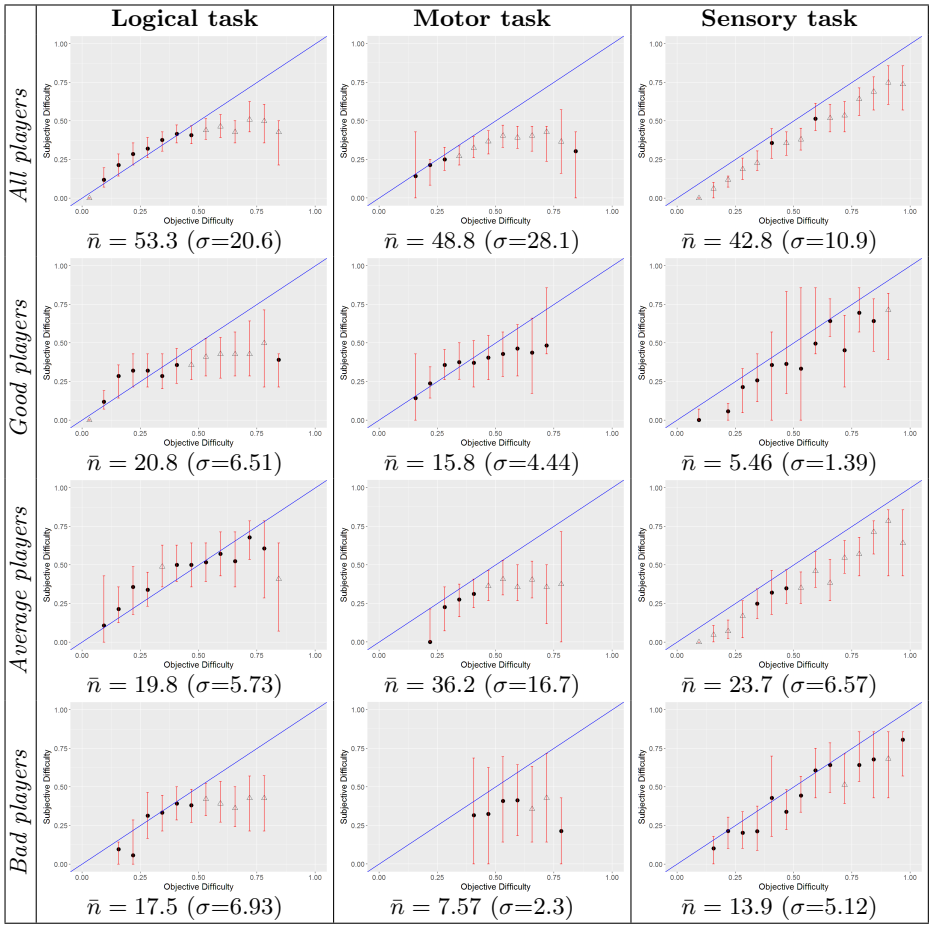


Fig. 5: Subjective and objective difficulty for all tasks and players. $\bar{n}$ is the mean (sd) number of players in each bin for each task and level.

There seems to be a strong *hard effect* for both logical and motor tasks. For the sensory task, players seem to be slightly overconfident for all objective

difficulties. When split by level, the effect seems stable for the motor task, but the relatively low number of bad ($\bar{n} = 7.57$) and good players ($\bar{n} = 15.8$) might mean this result is not significant. The same can be seen in the sensory task, where pseudo medians are always under the calibrated evaluation but the results lose significance with the decreased number of subjects. For the logical task, however, while bin sizes are equivalent for the three conditions, average (i.e. averagely-performing) players seem better calibrated. This result should be investigated further within a specific experiment to provide more in-depth results.

5.3 Influence of Participants' Profiles on Subjective Difficulty

We conducted several tests in order to analyze whether gender, gaming habits, assessment of self-efficacy and risk aversion have an impact on player level and the difficulty-estimation error. We took the random intercept of the objective difficulty model as each player's level.

Out of 80 participants, 57 were male and 23 were female. 49 of them play video games daily, and 12 weekly. 31 play board games monthly, and 36 almost never. 58 are risk-averse, and for the 46 of the participants who answered the self-efficacy questionnaire, 28 tended to see themselves as competent players and estimated themselves superior to an average player.

First, we tested gender influence on player levels and difficulty-estimation error using a Wilcoxon rank sum test. The null hypothesis is that both of them are derived from the same distribution for each gender. The test was only significant for player level. Female players seemed to perform less well on the motor game ($W = 255$, $p = 2.6^{e-5}$) with a difference in location of -0.67, and on the logical game ($W = 341$, $p < 0.01$) with a difference in location of -0.82.

We tested how gaming habits, self-efficacy and risk aversion impact on level and difficulty-estimation error using Kendall's rank-based correlation test. The test was only significant for the influence of risk aversion on player level for the sensory game ($z = 3.3093$, $p < 0.001$) with $\tau = 0.29$ and for the logical game ($z = 3.2974$, $p < 0.001$) with $\tau = 0.28$, meaning that for both these games risk averse players tend to perform better. Thus, in our experiment, we did not detect any impact of gender, playing habits, assessment of self-efficacy and risk aversion on difficulty-estimation error, but only on players' actual performance.

6 Discussion

6.1 Influence of Difficulty and the Hard Effect

We observed that the players estimation of difficulty is always below the actual objective difficulty, except for the logical and motor tasks on the easiest difficulty levels. More precisely, motor and logical tasks show the existence of a strong *hard effect* - that is, an overestimation of the player's chances of success at the hardest levels of difficulty (figure 5). Contrary to studies related to overconfidence, in addition to the *hard effect*, nothing seems to indicate any *easy effect* - that is, an underestimation of the chances of success for the easiest tasks [48, 56].

The presence of a *hard effect* and absence of an *easy effect* might be explained by the players' confidence in the game designers: games are rarely impossible to finish. What makes games different from many other tasks is that difficulty is artificially created for entertainment: players know that, given enough time, they are almost always supposed to eventually win. This may lead players to feel overconfident in their chances of success.

Moreover, player overconfidence and the hard effect may be stronger in our games than in previous cognitive psychology studies due to player progression. Indeed, our games allow players to experiment and learn from their failures, thereby improving their performance. This feeling of progression and mastery may help players to become more confident in their chances of success. In cognitive psychology studies, where general knowledge questionnaires are very often used, this might not be the case.

Players' global confidence towards the game and their feelings of progression and mastery are also enhanced by the use of the DDA algorithm. By presenting players with challenges that are adapted to their current level, the game is neither too boring nor too frustrating, allowing them to stay motivated and to believe in the fairness of the game.

In addition, in our experiment, we note that objective difficulty starts below 0.5, meaning that players face easier challenges at the beginning, when they are unfamiliar with the game, than at the end. Previous studies on the *hard/easy effect* rely on general knowledge questions, potentially enabling players to assess their knowledge and their chances of winning from the very first question. Therefore we may imagine that players' assessment of easy challenges is biased by their ignorance of the gameplay. However, the motor task has an almost flat progression curve, showing no evidence of an easy effect. Also, though both the sensory and logical tasks have easier challenges at the beginning of the session, for the logical task we seem to be nearing a small easy effect while for the sensory we see the opposite occurring, with players showing overconfidence for easy challenges. Here, therefore, oversampling easy challenges at the beginning of the session does not show a clear impact on the easy effect.

We may explain the differences in results between the sensory and both the logical and motor tasks by considering the nature of the subjective difficulty the players are asked to assess. As defined in our method in section 4.3, the betting system focuses on the player's estimation of their performance, and this estimation is not always performed under the exact same conditions. For the sensory game, the player can select the squares before betting, thus experiencing some gameplay before interpreting their chances of failure. While they are not yet aware of their actual performance, they do go one step further toward the completion of the challenge than for the two other games. By assessing their chances of failure after having experienced the exercise they may have a more accurate feeling about the quality of their answer. For the two other games they perform no interaction and must guess the tasks' next steps. This design choice for the sensory task was made because we did not want to focus on memorization, but on the sensory aspect of detecting blinking squares.

Note that our results differ from those of psychophysical studies on subjective difficulty, where perceived difficulty seems to never reach a plateau and shows a more linear or exponential curve. We think that this is mainly because we ask player to predict the difficulty of a challenge rather than to evaluate it after many repetitions. Our approach, which more closely resembles those used in cognitive psychology, may be closer to what a player really feels while playing.

The motor task is the one where the quality of our model is the lowest ($R^2 = 0.28$). It is the fastest game to play - participants can complete quickly one turn after the other - and this may explain the higher objective difficulty variability. However, this feature is typical of action games. While slowing the game's pace may produce stable results, the experiment would be less representative.

6.2 Impact of the Player's Profile

We did not find any evidence of the influence of the players' profile on their estimations of difficulty. This appears to contradict studies on overconfidence. Certain aspects of our experiment may be responsible for this, however.

In a field study conducted within the profession of financial analysis, Barber & Odean [57] looked at whether overconfidence can explain the difference in trading performance based on gender. They concluded that men have a tendency to be more overconfident and less risk averse than women. We did not observe this in our experiment. This can be attributed to differences in experimentation protocol between our study and theirs. First, the median age of our participants was 15 while theirs was 50. Secondly, their participants held a certain degree of expertise in investment, whereas ours were ignorant of the content of the games before playing them. Finally, it may be that as our tasks are very abstract they are less prone to culturally induced gender differences.

Risk-aversion is also a determinant of excessive confidence [57, 34]. However, we did not find any influence of risk on difficulty estimation error. In contrast to Barber's and Johnson's studies, the age of our participants was quite young. Also, as our questionnaire relies on mental calculus and probability assessments it may be less effective on adolescents.

Stone [58] shows that initial and positive self-efficacy assessment may reinforce participants' confidence and modify their performance. This was not evident in our study. In Stone's experiment, however, self-efficacy was assessed in relation to a given task, i.e. participants were asked to estimate their performance. In our study, we estimated self-efficacy using a general self-efficacy questionnaire [52, 53]. However, if we use players' mean bet as a measure of self-efficacy, there is a clear relationship between self-efficacy (how high the mean bet is) and overconfidence (how high mean bet minus mean actual result is). This is not surprising, as objective difficulty is adapted to 0.5 for each player. In addition, we found no link between the mean bet and player performance.

6.3 Limitations of the Experiment

There are some limitations into our approach, particularly in the betting system.

The bet system Our approach is based on the use of a betting system to measure the difficulty estimation error of players. This approach is limited to specific tasks, where the rhythm of interaction can be combined with a recurrent question addressed to the player. Also important to note is the fact that betting is not strictly related to confidence as measured in cognitive psychology studies. For our games, the optimal strategy is to bet 7 when $D_{\text{objective}} > 0.5$, and 1 when $D_{\text{objective}} < 0.5$. Therefore our evaluation maybe less accurate than confidence scales. Moreover, as we said in section 6.1, the betting system does not allow us to clearly distinguish between effort-based and skill-based subjective difficulties. Future experiments could improve the separation between them.

Dynamic Difficulty Adjustment DDA is representative of how video games are designed, and should have a notable impact on the hard/easy effect. Such an adjusted curve should allow players to feel more confident in their chances of success, allowing us to observe a weaker easy effect and a stronger hard effect than in a purely random experiment. Our experiment shows than when using DDA, players do develop a strong feeling of confidence in two of the three tasks. Nevertheless, to be able to attribute this overconfidence to DDA we would need an A/B experiment comparing our results with results derived from the use of a random difficulty system.

Motivational influences The actual performance of a player depends on both task difficulty and players effort. If a player is not motivated enough, they may make a correct assessment of difficulty but play less well because they do not want to make the effort. Video game players experience various states of emotion [59, 8], including boredom and anxiety. As such, these emotions should be taken account of for future experiments. We must also note that only the sensory and motor tasks induced a *near-miss effect*, while players were unaware of whether or not they were almost successful in the logical task. The *near-miss effect* may have convinced players that they were almost winning, leading them to overestimate their chances of success for the next turn [39, 40].

7 Conclusion and Perspectives

In this article we described our study investigating player perception of difficulty. Our work builds upon previous psychophysical and cognitive psychology research by proposing a method to evaluate objective difficulty, focusing on video games.

First, results demonstrate the efficacy of our method for objective difficulty estimation. The mixed effect model allows us to easily take into account differences between players. Results show a predictive accuracy ranging from 61% for the motor task, to almost 70% for the other tasks. Estimated objective difficulty is consistent with DDA, showing a convergence of objective difficulty to 0.5 for all groups and levels. We were also able to see a learning effect, as a negative effect of time on objective difficulty for a given difficulty parameter value. This learning effect is relative to the nature of the tasks, with a higher learning effect for the logical task.

These results confirm the existence of an unrealistic evaluation of players' actual chances of failure. More specifically, players were always overconfident, except at low levels of difficulty in the motor and logical tasks. A strong *hard effect* was present for the motor and logical tasks, with no significant *easy effect* for all tasks.

We suggest that this strong overconfidence might be attributable the fact that our tasks are video games. First, players know that games are designed to be eventually mastered. Second, games allow players to improve, developing a sense of progression and mastery. Furthermore, the use of DDA would reinforce both these aspects. The absence of a hard effect on the sensory task may be understood by considering its design: the difficulty evaluation was performed after players had started the task, thereby potentially gaining additional insight into their performance.

Further experiments will be conducted in order to improve our understanding of difficulty perception in video games. In order to validate the impact of DDA on the hard/easy effect we plan to compare our results with a second experiment that uses a random difficulty curve. In addition, we plan to investigate the influence of previous turns on the player perception of difficulty. DDA creates a temporal relationship between the difficulty of subsequent turns thereby preventing us from performing this analysis on our experiment.

We also plan to investigate the impact of feedback on players' assessment of difficulty. Constant feedback about the decision process makes participants re-evaluate their judgments during the task, attaining a higher level of accuracy [60]. Giving users continuous feedback on their progress is a key feature of human computer interaction in general and video games in particular. It requires distinguishing between positive and negative feedback and testing the influence of its accuracy. Video games adopt various types of feedback, both positive and negative, designed to affect players in terms of increasing the uncertainty of outcomes enhancing enjoyment [59, 49].

From a game design perspective, the presence of a hard effect has both benefits and disadvantages. The hard effect is a positive consequence of the game's motivational mechanics: if the player believes in their chances of success they may be motivated to play. However, having players believe that a challenge is easier than it is, particularly where difficulty is high, may also cause frustration because players will fail challenges they thought they could succeed in. The motivational aspects of the discrepancies between subjective and objective difficulty seem therefore worthy of further investigation.

Finally, we plan to expand our approach with other measures of mental effort like eye-tracking methods that have been used to assess cognitive load related to computer interface [61], specially about memory and logical related tasks [62].

Acknowledgment

Authors would like to thank Daniel Andler, Jean Baratgin, Lauren Quiniou, and Laurence Battais & Hélène Malcuit from *Carrefour Numérique*.

References

1. Juul, J.: The game, the player, the world: Looking for a heart of gameness. In Raessens, J., ed.: Level Up: Digital Games Research Conference Proceedings. Volume 1. (2003) 30–45
2. Malone, T.W.: Heuristics for designing enjoyable user interfaces: Lessons from computer games. Proceedings of the 1982 conference on Human factors in computing systems (1982) 63–68
3. Lazzaro, N.: Why we play games: Four keys to more emotion without story. In: Game Developers Conference. (March 2004)
4. Sweetser, P., Wyeth, P.: Gameflow: a model for evaluating player enjoyment in games. Computers in Entertainment (CIE) **3**(3) (2005) 3–3
5. Nakamura, J., Csikszentmihalyi, M.: The concept of flow. In: Flow and the foundations of positive psychology. Springer (2014) 239–263
6. Allart, T., Levieux, G., Pierfitte, M., Guilloux, A., Natkin, S.: Difficulty Influence on Motivation over Time in Video Games using Survival Analysis. In: Proceedings of Foundation of Digital Games, Cap Cod, MA, USA (2017)
7. Ryan, R.M., Rigby, C.S., Przybylski, A.: The motivational pull of video games: A self-determination theory approach. Motivation and emotion **30**(4) (2006) 344–360
8. Juul, J.: A Casual Revolution: Reinventing Video Games and Their Players. Mit press edn., Cambridge, USA (2009)
9. Juul, J.: The Art of Failure. 1 edn. The MIT Press, Cambridge, USA (2013)
10. Hunicke, R.: The case for dynamic difficulty adjustment in games. In: Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology, ACM (2005) 429–433
11. Andrade, G., Ramalho, G., Santana, H., Corruble, V.: Extending reinforcement learning to provide dynamic game balancing. In: Proceedings of the Workshop on Reasoning, Representation, and Learning in Computer Games, 19th International Joint Conference on Artificial Intelligence (IJCAI). (2005) 7–12
12. Vicencio-Moreira, R., Mandryk, R.L., Gutwin, C.: Now you can compete with anyone: Balancing players of different skill levels in a first-person shooter game. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM (2015) 2255–2264
13. Rani, P., Sarkar, N., Liu, C.: Maintaining optimal challenge in computer games through real-time physiological feedback. In: Proceedings of the 11th international conference on human computer interaction. Volume 58. (2005)
14. Afergan, D., Peck, E.M., Solovey, E.T., Jenkins, A., Hincks, S.W., Brown, E.T., Chang, R., Jacob, R.J.K.: Dynamic Difficulty Using Brain Metrics of Workload. In Jones, M., Palanque, P., eds.: CHI '14 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Toronto, Ontario, Canada, ACM New York, NY, USA (2014) 3797–3806
15. Aponte, M.V., Levieux, G., Natkin, S.: Difficulty in Videogames: An Experimental Validation of a Formal Definition. In Romão, T., ed.: Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology, ACE 2011, Lisbon, Portugal, ACM New York, NY, USA (2011) 1–18
16. Passyn, K., Sujan, M.: Skill-based versus effort-based task difficulty: A task-analysis approach to the role of specific emotions in motivating difficult actions. Journal of Consumer Psychology **22**(3) (2012) 461–468
17. Levieux, G.: Mesure de la difficulté dans les jeux vidéo. Thèse, Conservatoire National des Arts et Métiers CNAM Paris (2011)

18. Hunicke, R., LeBlanc, M., Zubeck, R.: MDA: A Formal Approach to Game Design and Game Research. In: Proceedings of the AAAI Workshop on Challenges in Game AI, San Jose, CA, USA, AAAI Press (2004)
19. Delignières, D., Famose, J.: Perception de la difficulté et nature de la tâche. Science et motricité **23** (1994) 39–47
20. Borg, G., Bratfisch, O., Dorni'c, S.: On the problems of perceived difficulty. Scandinavian journal of psychology **12**(1) (1971) 249–260
21. Slifkin, A.B., Grilli, S.M.: Aiming for the future: prospective action difficulty, prescribed difficulty, and fitts law. Experimental Brain Research **174**(4) (2006) 746–753
22. Delignières, D., Famose, J.P.: Perception de la difficulté, entropie et performance. Science & sports **7**(4) (1992) 245–252
23. Delignières, D., Famose, J.P., Genty, J.: Validation d'une échelle de catégories pour la perception de la difficulté. Revue STAPS **34** (1994) 77–88
24. Delignières, D., Famose, J.P., Thépaut-Mathieu, C., Fleurance, P., et al.: A psychophysical study of difficulty rating in rock climbing. International Journal of Sport Psychology **24** (1993) 404–404
25. Delignières, D., Brisswalter, J., Legros, P.: Influence of physical exercise on choice reaction time in sports experts: the mediating role of resource allocation. Journal of Human Movement Studies **27**(4) (1994) 173–188
26. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. Journal of experimental psychology **47**(6) (1954) 381
27. Kahneman, D., Frederick, S.: A model of heuristic judgment. In Holyoak, K.J., Morrison, R.G., eds.: The Cambridge Handbook of Thinking and Reasoning. 1 edn. Cambridge University Press, Cambridge, UK (2005) 267–293
28. Shah, A.K., Oppenheimer, D.M.: Heuristics made easy: an effort-reduction framework. Psychological bulletin **134**(2) (mar 2008) 207–22
29. Kahneman, D., Tversky, A.: Judgment under Uncertainty: Heuristics and Biases. Science (New York, N.Y.) **185**(4157) (sep 1974) 1124–31
30. Russo, J.E., Schoemaker, P.J.H.: Managing overconfidence. Sloan Management Review **33**(2) (1992) 7–17
31. Bessière, V.: Excès de confiance des dirigeants et décisions financières: une synthèse. Finance Contrôle Stratégie **10** (2007) 39–66
32. Moore, D.A., Healy, P.J.: The Trouble with Overconfidence. Psychological review **115**(2) (apr 2008) 502–17
33. Griffin, D., Tversky, A.: The weighing of evidence and the determinants of confidence. Cognitive psychology **411435** (1992) 411–435
34. Johnson, D.D.P., Fowler, J.H.: The evolution of overconfidence. Nature **477**(7364) (sep 2011) 317–20
35. Bandura, A.: Self-efficacy: Toward a Unifying Theory of Behavioral Change. Psychological Review **84**(2) (1977) 191–215
36. Keren, G.: Facing uncertainty in the game of bridge: A calibration study. Organizational Behavior and Human Decision Processes **39**(1) (feb 1987) 98–114
37. Linnet, J., Gebauer, L., Shaffer, H., Mouridsen, K., Møller, A.: Experienced poker players differ from inexperienced poker players in estimation bias and decision bias. Journal of Gambling Issues (24) (2010) 86–100
38. Park, Y.J., Santos-Pinto, L.: Overconfidence in tournaments: Evidence from the field. Theory and Decision **69**(1) (2010) 143–166
39. Sundali, J., Croson, R.: Biases in casino betting : The hot hand and the gambler ' s fallacy. Judgment and Decision Making **1**(1) (2006) 1–12

40. Parke, J., Griffiths, M.: The psychology of the fruit machine: The role of structural characteristics (revisited). International Journal of Mental Health and Addiction **4**(2) (2006) 151–179
41. Lichtenstein, S., Fischhoff, B.: Do those who know more also know more about how much they know? Organizational Behavior and Human Performance **20** (1977) 159–183
42. Klayman, J., Soll, J.B.: Overconfidence: It depends on how, what, and whom you ask. Organizational Behavior and Human Decision Processes **79**(3) (1999) 216–247
43. Kahneman, D., Tversky, A.: Subjective probability: A judgment of representativeness. Cognitive Psychology **3**(3) (1972) 430–454
44. Croson, R., Sundali, J.: The gambler's fallacy and the hot hand: Empirical data from casinos. Journal of Risk and Uncertainty **30**(3) (2005) 195–209
45. Gilovich, T., Vallone, R., Tversky, A.: The hot hand in basketball: On the misperception of random sequences. Cognitive Psychology **17**(3) (jul 1985) 295–314
46. Langer, E.J.: The illusion of control. Journal of personality and social psychology **32**(2) (1975) 311–328
47. Goodie, A.S.: The role of perceived control and overconfidence in pathological gambling. Journal of Gambling Studies **21**(4) (2005) 481–502
48. Pulford, B.D., Colman, A.M.: Overconfidence: Feedback and item difficulty effects. Personality and Individual Differences **23**(1) (jul 1997) 125–133
49. Costikyan, G.: Uncertainty in Games. 1 edn. MIT Press, Cambridge, USA (2013)
50. Lankoski, P., Björk, S.: Game Research Methods: An Overview. 1 edn. ETC Press (2015)
51. Bates, D., Mächler, M., Bolker, B.M., Walker, S.C.: Fitting linear mixed-effects models using lme4. Journal of Statistical Software **67**(1) (2015) 1–48
52. Chen, G., Gully, S.M., Eden, D.: Validation of a New General Self-Efficacy Scale. Organizational Research Methods **4**(1) (2001) 62–83
53. Bandura, A.: Guide for constructing self-efficacy scales. In Urdan, T., Pajares, F., eds.: Self-efficacy beliefs of adolescents. 1 edn. Information Age Publishing, Charlotte, USA (2006) 307–337
54. Holt, C.A., Laury, S.K.: Risk aversion and incentive effects. The American Economic Review **92**(5) (2002) 1644–1655
55. Nakagawa, S., Schielzeth, H.: A general and simple method for obtaining r2 from generalized linear mixed-effects models. Methods in Ecology and Evolution **4**(2) (2013) 133–142
56. Keren, G.: On The Calibration of Probability Judgments: Some Critical Comments and Alternative Perspectives. Journal of Behavioral Decision Making **10**(3) (sep 1997) 269–278
57. Barber, B.M., Odean, T.: Boys will be boys: Gender, overconfidence, and common stock investment. Quarterly Journal of Economics **116**(1) (2001) 261–292
58. Stone, D.N.: Overconfidence in Initial Self-Efficacy Judgments: Effects on Decision Processes and Performance. Organizational Behavior and Human Decision Processes **59**(3) (1994) 452–474
59. Caillois, R.: Les jeux et les hommes : le masque et le vertige. 2 edn. Gallimard, Paris, France (1958)
60. Arkes, H.R., Christensen, C., Lai, C., Blumer, C.: Two methods of reducing overconfidence. Organizational Behavior and Human Decision Processes **39** (1987) 133–144

61. Goldberg, J.H., Kotval, X.: Computer interface evaluation using eye movements : Methods and constructs Computer interface evaluation using eye movements : methods and constructs. International Journal of Industrial Ergonomics **24**(November 2015) (1999) 631–645
62. Klingner, J., Tversky, B., Hanrahan, P.: Effects of visual and verbal presentation on cognitive load in vigiolance, memory, and arithmetic tasks. Psychophysiology **48** (2011) 323–332