Objective Difficulty-Skill Balance Impacts Perceived Balance but Not Behaviour: A Test of Flow and Self-Determination Theory Predictions

SEBASTIAN DETERDING, Imperial College London, UK
JOE CUTTING, University of York, UK

Flow and self-determination theory predict that game difficulty in balance with player skill maximises enjoyment and engagement, mediated by attentive absorption or competence. Yet recent evidence and methodological concerns are challenging this view, and key theoretical predictions have remained untested, importantly which *objective* difficulty-skill ratio is *perceived* as most balanced. To test these, we ran a preregistered study (n=309) using a Go-like 2-player game with an AI opponent, randomly assigning players to one of three objective difficulty-skill ratios (AI plays to win, draw, or lose) over five matches. The AI successfully manipulated objective balance, with the draw condition perceived as most balanced. However, balance did not impact play behaviour, nor did we find the predicted uniform 'inverted-U' between balance and positive play experiences. Importantly, we found both theories too underspecified to severely test. We conclude that balance and competence likely matter less for behavioural engagement than commonly held. We propose alternative factors such as player appraisals, novelty, and progress, and debate the value and challenges of theory-testing work in games HCI.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **Empirical studies in interaction design**; • **Applied computing → Computer games**.

Additional Key Words and Phrases: Balance, Challenge, Difficulty, Skill, Flow, Engagement, Competence, Self-Determination Theory

ACM Reference Format:
Sebastian Deterding and Joe Cutting. 2023. Objective Difficulty-Skill Balance Impacts Perceived Balance but Not Behaviour: A Test of Flow and Self-Determination Theory Predictions. *Proc. ACM Hum.-Comput. Interact.* 7, CHI PLAY, Article 419 (November 2023), 27 pages. https://doi.org/10.1145/3611065

1 INTRODUCTION

Difficulty or challenge are widely recognised to impact how we experience games and interactive systems [11, 44, 76]. Challenge is the most frequently studied antecedent of player experience in games HCI [51] and sometimes framed as "the most important aspect of good game design." [78] Grounded in flow theory [17], early game and media entertainment theory suggested that "it is a balance between individual differences in cognitive abilities and media message challenges that explains enjoyment of media use" [71]. Good games are seen to keep players in a so-called flow channel, where challenges adapt to and grow in lockstep with each player's ability [13]. This *balanced challenge* view predicts an inverted-U relation: Where player skill significantly exceeds

Authors' addresses: Sebastian Deterding, s.deterding@imperial.ac.uk, Imperial College London, 25 Exhibition Road, South Kensington, London, London, UK, SW7 2DB; Joe Cutting, joe.cutting@york.ac.uk, University of York, Deramore Lane, Heslington, York , York, UK, YO10 5GH.

2573-0142/2023/11-ART419 $15.00
https://doi.org/10.1145/3611065

difficulty, play is boring. Where difficulty exceeds skill, players feel anxious and frustrated. The closer we move from either end to a goldilocks middle, the more enjoyable and absorbing play becomes.

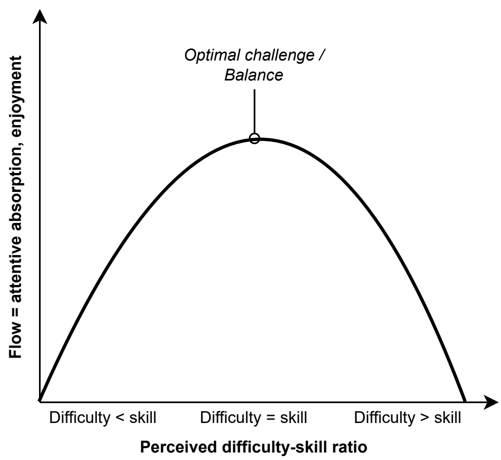


Fig. 1. The inverted-U of difficulty-skill balance

In games HCI, *flow theory* has informed the widely adopted GameFlow model, originally developed for strategy games but since widened into a "general model of player enjoyment" [77]. GameFlow posits "Challenge: Games should be sufficiently challenging and match the player's skill level" as one of eight core elements of player enjoyment [77]. This is repeated in games research grounded in *self-determination theory* (SDT), claiming that "optimal challenges" maximise competence experiences [63]. The balanced challenge view features prominently in game design textbooks [31, 49, 69, 86] and informs game design practices like balancing [70], matchmaking [34, 52], or rational level design [50], as well as game AI applications like dynamic difficulty adjustment [85] or procedural content generation [75, 84].

Yet the balanced challenge view is coming into question. Millions of players are deeply engaged by Idle games (games of economic accumulation that self-play without player input, often characterised as having no obvious challenge) [6, 21] as well as Soulslike games, a subgenre of action role-playing games notorious for their high difficulty and frequent player deaths [60]. These contradict – at least at first sight – the idea that enjoyment and engagement cannot be had without balanced challenges. Evidence does not uniformly support theory [30]. Recent conceptual analyses predict a linear not U-shaped relation between difficulty and competence [26]. Empirical work on balanced challenge often suffers from methodological shortcomings [20] and has mostly studied *perceived* balance, which is less useful for commercial dynamic difficulty adjustment systems or AI directors that operate with objective difficulty measures. The few empirical studies on *objective* difficulty-skill balance present conflicting evidence. Importantly, they did not directly test flow and SDT predictions on proximal psychological outcomes of competence and attentive absorption; nor what *objective* difficulty-skill ratio is actually *perceived* as the most balanced.

All this makes game balance an excellent case in point for recent calls in (games) HCI for more careful use, testing, and extension of psychological theories [7] such as flow or SDT [8, 79], and more rigorous replication work [14, 27, 82, 83] in response to the so-called replication crisis. Standard operating procedures in the quantitative social and behavioural sciences have been found to be prone to generating false-positive findings that don't replicate when put to the test [72]. Researchers have therefore been advancing methodological reforms for more severe, transparent, and reproducible testing of hypotheses [54, 57]. *Severe testing* here means designing studies "highly capable of demonstrating a claim is false" [46], which can be furthered with strategies like *preregistration*: publicly specifying tested hypotheses, research designs, and analysis plans before collecting and analysing data.

In addressing important open questions around difficulty-skill balance, we also want to take a step toward such more 'severe' theory-testing and -developing games HCI. Our aim is thus *to severely test flow theory and SDT predictions on the impact of objective difficulty-skill ratios on* *proximal player experiences (perceived balance, competence, attentive absorption) and their distal outcomes (enjoyment and behavioural engagement).* For this, we use a novel experimental paradigm [20] with an AI opponent to unobtrusively induce three different objective difficulty-skill ratios: AI playing weaker than the human player (easy), equally strong (balanced), or stronger (hard). In our preregistered online experiment, participants (n=309) were randomly assigned to one condition for five matches before self-reporting their play experience and then freely choosing to continue playing or end the study.

Our results make several contributions to games HCI: First, we find that an equally strong opponent is perceived as most balanced, supporting this common but previously untested intuition. Second, objective and perceived balance did not significantly impact behaviour, nor did we find a uniform inverted-U. This contradicts standard views that balance is "the most important aspect of good game design" [78] and suggests that designers of single-player games should not unquestioningly rely on perceived balance measures or matched strength as a given optimum. Third, our data contradicts several claims in flow theory and SDT, suggesting at a minimum previously unrecognised limitations or moderators. Finally, we find that both theories are not well-specified enough to allow actual severe testing; future theoretical work is needed to formalise optima and claimed scopes of generality.

2 BACKGROUND

To set up our paper, we will first clarify terminology, then articulate what specific claims flow theory and SDT make about difficulty-skill balance, followed by a review of empirical work.

2.1 Terminology: Difficulty, Challenge, and Balance

There are no strong consensus definitions or operationalisations for "difficulty" or "challenge" [25]. The two terms are often used interchangeably. We here focus on difficulty and challenge in the narrow sense of the effort and ability required to accomplish a given task and resultant likelihood of attainment (as opposed to more recent broader framings such as "emotional challenge" [15]).

A key consideration is whether we are interested in the first-person, subjective experience of a player (as Denisova and colleagues [24] do), or some third-person, intersubjective quality of a game (as game developers do when they talk about, e.g., the difficulty level of a game). We will refer to the former as *perceived*, and the latter as *objective* difficulty or challenge.

Objective difficulty in turn can be approached in at least two ways. One is 'raw' difficulty *as the attribute of a task, game, or match on its own*, without respect to a player (group). This could be operationalised, e.g., as the number of enemies in a level, the size of a target, or the TrueSkill or ELO rating of an opponent – two common player rating systems used in multiplayer games [28, 35]. More often, we are interested in difficulty as a *relational property between game task and player*. Such *difficulty-skill ratios* can be operationalised as an antecedent, e.g., as the delta between the ELO ratings of two players at the start of a game, or post-hoc as a player's observed final score, win/loss ratio, etc. In this paper, we use "difficulty" to refer to the raw properties of a game task, and "challenge" or "difficulty-skill ratio" for player-relative difficulty: how attainable or effortful a given task is for a given player. Thus, *objective challenge* refers to intersubjectively measurable difficulty-skill ratios, and *perceived challenge* to a given player's felt sense of challenge facing a given task.

"Balancing" or "balance", finally, commonly refer to a broad range of aspects of a game that designers may tune to accomplish a desired play experience [9, 69, 70]. We here focus on *balance* as closeness to an assumed *optimal* difficulty-skill ratio: the closer to this optimum, the more balanced. *Perceived balance* will refer to player's subjective sense that the difficulty-skill ratio of a game "feels right" for them.

2.2 Difficulty-Skill Ratios in Games HCI Theory

Flow theory is probably the most-frequently used theory to study difficulty-skill ratios. It equates enjoyable flow states or optimal experience with "intense experiential involvement in moment-to-moment activity" [17, p. 230]. More recent work by Abduhamdeh and Csikszentmihalyi suggest that any impacts of balance on enjoyment are fully mediated by this *attentional involvement* [3] (see also [2, 38]).

Flow theory frames difficulty-skill balance or "when environmental challenges are in balance with the person's skills" [19, p. 187] as one signature or antecedent of flow [17, 56]. Out of this comes the oft-reprinted "flow channel" diagram ([18, pp. 147-8], [17, p. 74]), which can be remapped into an inverted-U graph with flow, attentive absorption, or enjoyment as the x-axis and difficulty-skill ratio on the y-axis, and flow peaking where the difficulty-skill ratio is "in the neighbourhood of a 50/50 balance" [19, p. 187] (see Figure 1). Csikszentmihalyi variously insisted that this balance refers to "*subjective challenges and subjective skills, not objective ones*" ([56, p. 91], emphasis in original). However, he also suggested that objective circumstances like two Chess players with equal UCSF strength ratings could manifest balance [18, p. 148]. In later years, he co-authored studies that operationalised objective balance as equal Chess ratings of opponents or equal opponent scores mid-game [4, 5]. This implies that Csikszentmihalyi equated flow-optimal perceived balance to an objective challenge of facing opponents of play strength equal to oneself and thus, equal odds of winning for each party. This assertion has been embraced as a truism in, e.g., matchmaking [34], but not actually been tested to our knowledge. Importantly, matchmaking is used in multiplayer games, where equal opponent strength ties into additional player concerns like fairness, which don't necessarily obtain in single-player contexts.

Next to flow theory, self-determination theory (SDT [67]) has become a key theory in games HCI [79]. It proposes that activities like video game play are intrinsically motivating and enjoyable because and when they satisfy basic psychological needs like competence – "feeling effective in one's interactions with the social environment, that is, experiencing opportunities and support for the exercise, expansion, and expression of one's capacities and talents"' [67, p. 86]. While less commonly connected to difficulty-skill balance than flow theory, SDT nevertheless makes similar claims about balance impacting motivation and enjoyment. SDT creators Deci and Ryan have consistently asserted that competence needs are satisfied by seeking-and-overcoming *optimal challenges*: "the needs for competence and self-determination keep people involved in ongoing cycles of seeking and conquering optimal challenges" [22, p. 33] (see also [63, 67]). At the same time, SDT predicts that competence satisfaction largely arises from positive competence feedback, also labelled the "perceived competence effect" [67, p. 156]. "Within SDT, then, optimal challenge means facing demands that most often one can master, rather than ones that are continuously at the leading edge of one's capabilities. That type of high difficulty challenge should, however, be an intermittent element, in which case it can enhance and heighten intrinsic motivation" [67, p. 153]. As this quote from the current state-of-the-art SDT articulation shows, unlike flow theory, SDT refers to *objective* not perceived challenge. Yet it makes somewhat incongruent and ambiguous predictions: One, competence satisfaction is maximised under *optimal challenges* with frequent but not guaranteed success, not specifying what precise success ratio or odds are optimal and why. Two, competence satisfaction is driven by *positive competence feedback* – which should logically be greater the more frequent and extensive success occurs. Ryan and Deci acknowledge that this incongruence holds empirically: optimal challenge predicts a competence-enhancing effect from occasional negative competence feedback, yet empirically they observed only a perceived competence effect to date [67, p. 156]. Nevertheless, SDT's optimal challenge model is underwriting current industry design consultancy [62, 63], has informed theory and empirical work on games [42, 68], and related scales like PENS, which entails a dedicated balance item in its competence subscale [68].

2.3 Prior empirical work

Based on flow theory and SDT, difficulty-skill balance has been widely researched (for reviews, see [30], [67, pp. 153-157]). Results are not uniformly supporting theory. The most recent meta-analysis of 28 eligible studies found difficulty-skill balance to have a moderate relation with flow state (Fisher's $z = 0.58$), and a small relation with intrinsic motivation ($z = 0.24$) [30].

2.3.1 Methodological issues. Unfortunately, the majority of existing studies suffer from various methodological issues [20, 30]. The first is *questionable measurement practices* [29]: studies have used such a variety of scales, measures, and calculations for challenge and balance that the cited meta-analysis found them simply "too heterogeneous to meaningfully aggregate" [30]. These often include unvalidated single-item measures. Other studies use as their independent variable the challenge subscale of the same flow scale that is then used to measure flow as dependent variable, risking spurious correlations or the jangle fallacy [29]: assessing the same thing by different names. In other cases, balance is operationalised as 'using an industry dynamic difficulty adjustment system,' with no insight into how said system quantifies balance [42].

Many studies also lack construct validity, as they manipulate *cohort-level absolute difficulty*, while flow theory and SDT make claims about *individual-level difficulty-skill ratios*. This operational proxy is defensible under random assignment provided that (a) skill in the sample is normally distributed and (b) said normal distribution peaks on whatever is considered a 'balanced' match to the difficulty of the game under study – for instance, player skill is normally distributed such that the median player would most likely arrive at a draw. Under these conditions, the fact that each individual player brings a different skill level to an assigned task and may therefore face a different difficulty-skill ratio that deviates from the ideal (median) of the assigned condition gets aggregated out as random noise. But the moment either assumption is violated, such study designs no longer allow valid inferences to the effect of individual-level difficulty-skill ratios. To our knowledge, no study relying on these assumptions actually tested whether they hold.

Furthermore, existing studies did not preregister hypotheses and analysis plans [54, 74]. We therefore cannot say with certainty whether or how frequently researchers chose not to report harder to publish null results, or explored data and analysis strategies until they found some statistically significant inverted-U to report. We don't imply that any given researcher did engage in such practices. But we know empirically that introducing preregistration substantially increases the proportion of reported null results in a given body of literature [40].

2.3.2 Perceived versus objective challenge. Importantly for practice, the vast majority of prior work assessed *perceived* difficulty-skill balance. While this is true to flow theory, it doesn't match SDT's claims about objective difficulty-skill balance. And it is of little help to game designers wanting to balance their games using AI directors, PCG systems, dynamic difficulty adjustment, matchmaking, or similar, as these systems overwhelmingly operate with targeted and observed *objective* challenge metrics like TrueSkill rankings, frequency or odds of player death, needed attempts per level, positions in a race, and the like [34, 85]. (Affect recognition-based systems exist but have remained research prototypes. We are not aware of mass-market difficulty balancing system that use self-report or biosignals as input.) Furthermore, neither flow theory nor SDT make explicit claims when, why, and how perceived and objective challenge *should* diverge. If the two track closely, this would support industry practice of focusing on easier to measure and manipulate objective balance. If they diverge, and perceived balance drives engagement, industry practice would be flawed.

2.3.3 Conflicting results on objective challenge. Based on a prior systematic review covering work published until 2014 and our own unsystematic literature search from 2014 onward [30], we identified seven studies that directly test relations between an objective, individual-level skill-difficulty ratio and fully separate enjoyment or engagement measures. The first found an inverted-U between challenge and enjoyment using online Chess matches, operationalising challenge as relative pawn strength at a given board state [4]. A second used a 'Wizard of Oz' manipulation of a pseudo-AI opponent in a digital dexterity game, measuring challenge as player:opponent score ratios; it found an inverted-U between challenge and enjoyment, and a positive linear relation between challenge and competence [5]. In contrast, a large-scale online maths game study for children found a positive linear relation between success rate and behavioural engagement [47]. This was replicated in a follow-up study [48]. A study on a serious game for literacy education manipulated difficulty as the proportion of successful trials, but found no difference in engagement between a 60% success rate and an 80% success rate [65]. A final recent study [20] used an AI opponent to manipulate individual-level, objective difficulty skill ratios with three conditions (player win, draw, loss). Surprisingly, itfound no significant difference in enjoyment or behavioural engagement across conditions.

2.3.4 Lack of direct theory-testing. The studies we found by and large haven't directly tested the proximal predictions of either theory. Thus, none of the above studies nor the studies included in Fong et al.'s systematic review [30] tested which objective difficulty-skill ratio was perceived as most balanced, leaving a core question for applying flow theory open. This also holds for the theories core mediation claims: they predict challenge to impact enjoyment and engagement mediated by attentive absorption (flow theory) or perceived competence (SDT). Yet only Abduhamdeh and colleagues [5] included a perceived competence measure in their study. All other studies assessed enjoyment or behavioural engagement as distal dependent variables.

2.3.5 Summary. In short, objective challenge is widely used in game balancing systems and practice, and the idea of an inverted-U between objective challenge and enjoyment/engagement is widely taken as given. Yet prior research has chiefly focused on perceived not objective challenge and suffers from methodological shortcomings. We currently don't know what objective challenge *does* maximise perceived balance. What's more, the two major underpinning theories offer little unambiguous predictions about what objective challenge *should* maximise perceived balance, enjoyment, and behavioural engagement. Csikszentmihalyi variously suggested that it might be maximised under 50% success odds. SDT offers no similar quantified specification of optimal challenge as "demands that most often one can master", yet makes the competing claim that competence is maximised under frequent positive competence feedback. Rigorous empirical work on objective difficulty-skill balance variously finds inverted-Us, linear relations, or no significant relations. Flow theory and SDT make direct predictions about balance and proximal psychological outcomes (attentive absorption and competence), yet only one study directly tested these (finding a linear not U-shaped relation between success rate and competence).

2.4 Aims and hypotheses of the present study

Given the above unclear picture, the present study sets out to *severely test key flow theory and SDT propositions on how objective difficulty-skill ratios impact play experience and behaviour*. Both theories suggest that objective difficulty affects player experience (H1), specifically enjoyment (H2) and subsequent intrinsically motivated behaviour (H3), mediated by enjoyment (H5). Flow theory predicts an inverted-U where a *perceived* mid-point of optimal balance affords maximal attentive absorption (H1d) , which drives enjoyment (H2a) and subsequent engagement (H3a). This perceived optimum balance is proposed to sit around an objective challenge mid-point of 50% success odds

(H1a). SDT predicts an inverted-U around an objective optimal challenge of high but not guaranteed success odds, which maximises competence satisfaction (H1c but *not* H1b), driving enjoyment (H2a) and engagement (H3a). Yet SDT *also* predicts that more positive competence feedback results in higher competence satisfaction, suggesting that the higher the success rate, the more competence satisfaction we see (H1b *and* H1c). Finally, as a manipulation check, we predicted that our AI would create the desired objective difficulty-skill ratios (H4, H4a). Our full set of preregistered hypotheses is thus:

- **H1**: There will be a significant difference in experience between the three conditions (hard, balanced, easy), measured as a difference in at least one of the three constructs of perceived balance, competence, and attentive absorption.
- **H1a**: Perceived balance will be significantly higher in the Balanced than in the Hard or Easy condition.
- **H1b**: Competence will be significantly higher in the Easy than in the Balanced or Hard condition.
- **H1c**: Competence will be significantly higher in the Balanced than in the Hard condition.
- **H1d**: Attentive absorption will be significantly higher in the Balanced than in the Easy or Hard condition.
- **H2**: There will be a significant difference between levels of experiential enjoyment in the three different game conditions.
- **H2a**: The Balanced condition will have a significantly higher experiential enjoyment than either other condition.
- **H3**: There will be a significant difference between levels of behavioural engagement in the three different game conditions.
- **H3a**: The Balanced condition will have significantly higher behavioural engagement than either other condition.
- **H4**: There will be a significant difference in player win margins between conditions in the final demanded game that participants play.
- **H4a**: There will be a significant difference in player win margins between each condition in the direction that the AI was aiming for.
- **H5**: There will be a significant correlation between experiential enjoyment and behavioural engagement.

3 METHODS

To address the methodological shortcomings of prior work, Cutting et al. [20] recently developed an experimental paradigm that allows to precisely and transparently manipulate individual-level, objective difficulty-skill ratios. Their Go-like game, *Explorers vs Owls* (Figure 2), uses an AI opponent that turn-by-turn chooses the move most likely to produce a particular target outcome in the form of a score delta between human player and AI. Thus, the 'strength' of each move is adjusted to the individual player's strength toward a target difficulty-skill ratio. In a preregistered study (n=311), they found that the AI produced the desired difficulty-skill ratios without players noticing the manipulation, supporting the validity of their approach.

We here adopted this paradigm and used the materials and procedures of Cutting et al. [20] (see below for detailed description), with two amendments: First, one important limitation of their study was that the average play session lasted only seven minutes (two matches); hence, enjoyable novelty could have drowned out weaker yet true balance effects. To address this, our study increased play duration to five matches (avg. 16 minutes total playtime) before measuring experience and engagement. Second, we added additional post-play survey items to assess relevant player experience constructs. Our study received ethical approval from the University of York. Pre-registration, data, analysis code, and study materials can be found at https://osf.io/8wtve/ (folder Experiment 2). We did not deviate from the preregistered design, sampling, and analysis, with two exceptions: For the sake of brevity, we don't report all proposed exploratory analyses of PXI measures, and we excluded 18 participants who took longer than 30 minutes to complete participation (see below).

3.1 Materials

We asked participants to play *Explorers vs Owls*, a custom-built turn-based, tactical 2-player game (see Figure 2. The game is based on the multi-award winning commercial encircling board game *Hey That's My Fish!*, [16], which has also been published as a casual single-player mobile game [32]. It is rated as an accessible family game (age 8+) on the community site boardgamegeek [16]. A match of the single-player version takes only a few minutes to play. *Explorers vs Owls* fully emulates the written rules of *Hey, That's My Fish*, but changes the original theming and operates with a fixed 7x7 board. Turn-based, with no hidden information, game states fully and predictably determined by player moves, and with no randomness (save the initial board generation), game outcomes are fully dependent on player/AI tactical skill. Reviewers on boardgamegeek likened it to games like Go and Chess in that it invites careful calculating of moves and counter-moves several turns into the future. Informal playtests showed that our version retains the approachability and tactical depth of the original.

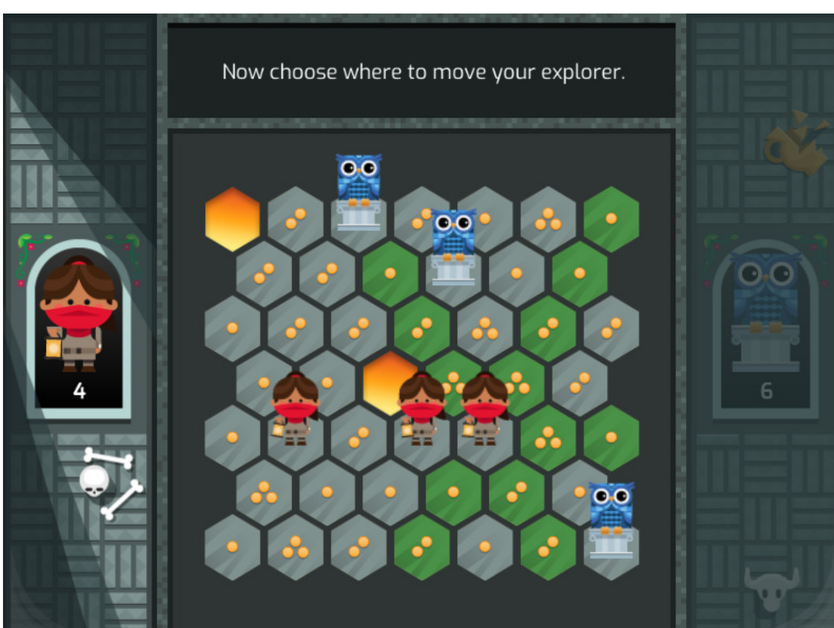


Fig. 2. In *Explorers vs Owls*, the opponents start alternatively placing three figures on a 7x7 grid of hexagonal tiles carrying a random distribution of one to three coins each. They then take turns moving their figures along straight unobstructed lines to a free tile of their choice an collect the coins thereon. Figures obstruct movement on or through them, and any tile that is moved away from is replaced with lava, obstructing movement through it. A figure with no available movements (encircled by lava and/or board edges) is removed from the game, until no figure remains. The opponent with the most coins at the end wins the game.

To manipulate objective difficulty-skill ratios, we used an AI opponent implementing True Outcome Sensitive Action Selection (True POSAS) [23], a variant of the Monte Carlo Tree Search (MCTS) algorithm designed for difficulty adjustment in turn-based games. MCTS algorithms build a branching decision tree to some n steps into the future to determine the optimal decision per its reward function [12]. MCTS probabilistically samples branches deeper that promise higher rewards. This allows it to search select branches many steps ahead, which is particularly useful in games like Go where outcomes cannot be fully determined until the end of the game. MCTS-based systems (such as *AlphaGo* [73]) have beaten top-ranked human Go players.

True POSAS uses a graded reward function that gives a higher reward the closer the final score is to a given target score or difficulty level. In our game, human and AI player collect coins, where the party with the highest tally wins. Thus, the target score can be expressed as *Player Win Margin (PWM)*, calculated as the human player score minus the AI opponent score. The theoretically possible PWM span of *Explorers vs Owls* is 162, from -81 (AI wins by 81 points) to 81 (human wins by 81 points). As conditions, we set three different target scores that pilots had shown to feel markedly different to players and are median PWMs in human-human play:

- **Easy**: Human win (PWM=15) – an AI opponent playing consistently weaker than the player
- **Balanced**: Draw (PWM=0) – an AI opponent playing consistently on par with the player
- **Hard**: AI win (PWM=-15) – an AI opponent playing consistently stronger than the player

In games like *Explorers vs Owls*, the first mover has an advantage, and in our implementation, the human moves first. To counterbalance this, True POSAS uses a 'flattened cone' reward function where the highest reward value stretches from the target score to a score that is slightly more in favor of the AI (as second mover) (see Figure 3). The more iterations (complete playthroughs) the

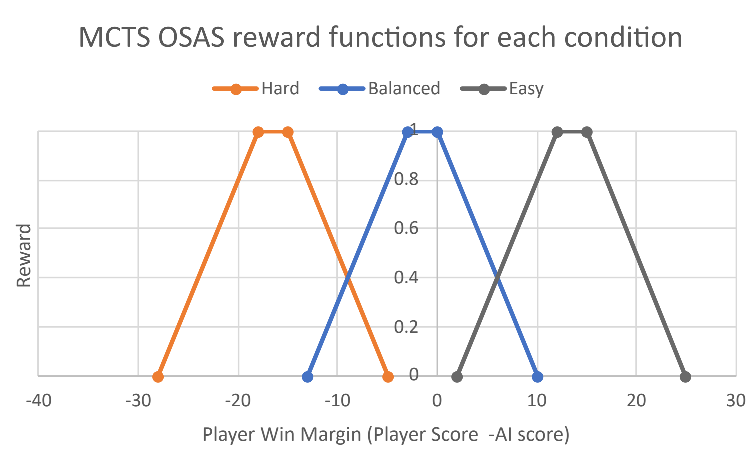


Fig. 3. Reward function of MCTS OSAS (True POSAS variant) for Easy, Balanced, and Hard game conditions, with a target "flattened cone" of maximal reward spanning PWM 0 to -3 for the Balanced condition, PWM 13 to 18 for Easy, and PWM -15 to -18 for Hard.

AI can simulate, the more likely it will achieve its target score. However, each iteration requires computing time. If the AI takes too long, this could confound player experience [55]. Participants access *Explorers vs Owls* as an online browser application on their own computers, resulting in a wide variety of computing performance. Thus, we need to determine how many iterations afford good-enough AI performance without creating undue player wait time on home computers, and ensure that much slower computers don't confound the results. Following Cutting and colleagues [20], we set our AI to 7,500 iterations. We also replicated their screenings to reject participants whose computers were too slow. As a pre-study screen, participants' computers are tested to create 100,000 memory objects in less than 0.17 seconds; those who fail are prevented from starting the study. As a post-study screen, we reject participants whose computers take more than 6 seconds for any one move. Note that the AI takes longest for the first moves and then speeds up rapidly as the decision tree shrinks. Hence, a maximum move time of 6 seconds corresponds to an average move time of around 2.5 seconds. Human move animations take around 1-2 seconds, during which the AI already computes. Our 6 second cut-off thus results in an average perceived wait time of less than a second.

In practical terms, at each turn, the AI selected its next move based on the current board state to make it most likely that the game would end as closely as possible to the target outcome (PWM 15, 0, or -15). It thus adjusted its play strength move-by-move relative to each individual player's strength toward a target objective difficulty-skill *ratio*, operationalised as the target PWM. If a good player makes consistently 'good' moves, and the target outcome is set to AI win, the AI will consistently choose 'even better' moves; if the target is set to draw, the AI will consistently choose 'evenly matched' moves. This way, our system operationalised different objective difficulty-skill *ratios* (rather than fixed levels of difficulty), and did not have to rely on any assumed randomised normal distribution of player skill across and within the three conditions (although we did randomise player assignment to conditions). By directly manipulating and measuring difficulty-skill *ratios* (in the form of PWMs), we similarly avoid the need to measure or control for game difficulty and player skill separately. And unlike prior work using 'black box' commercial dynamic difficulty adjustment, this system is fully transparent and manipulable.

3.2 Procedure

Participants were recruited via the Prolific online recruitment platform and paid £2.90. On signing in to the Prolific study, each participant is allocated to one condition in turn: the first is assigned Easy, the second Balanced, and so on. The study then runs the described pre-study speed test. Participants who pass it provide consent and complete a demographic questionnaire, and go through a short tutorial of *Explorers vs Owls* with one piece per player and a smaller board (4x4 tiles), where the AI aims for a draw. Afterward, participants play five full matches, where the AI aims for the assigned PWM condition in each match. After the last match, participants complete the questionnaire embedded into the game and report any bugs or technical issues. Participants are then informed that they have done all that is needed to receive payment for the study but may play another round of the game if they wish, which they can do.

3.3 Participants

To determine desired power, we chose a smallest effect size size of interest of $d = 0.3$, an alpha of 0.05 and power of 90% for our Bonferroni-corrected test (H1). Simulations using a custom R script (see Supplementary Materials) indicated that 285 participants would give 90% power for our Bonferroni-corrected test and 95% power for all other, individual tests. We rounded this up to a target of 300 participants in total, with 100 participants per condition.

350 participants completed the experiment. In addition to the stated pre- and post-study speed screens, we preregistered three exclusion criteria: (1) We exclude any participant who fails our attention check. (2) We exclude participants who take less than half (<22) the mean number of moves per match from pilot studies (44) in any match, which may suggest they intentionally played to lose. (3) We discard any participant who self-report bugs or performance issues that would have significantly changed their game experience. We rejected one participant in speed screening, one

who made less than 22 moves in a match, ten who failed the attention check, and 11 who reported technical issues. Although not in the pre-registration, we also rejected 18 who took more than 30 minutes to perform the experiment, as we were concerned that they may have taken breaks during the game which would confound the game experience. This left 309 valid participants, 100 in the easy condition, 109 in the balanced condition and 100 in the hard condition. Age ranged from 18-56 (M = 30.4). In response to our free-text prompt "Gender", 130 participants identified as women, 174 as men, two as non-binary and three preferred not to give a gender.

3.4 Measures

Replicating [20], we measured **enjoyment** using the interest/enjoyment scale of the IMI questionnaire [66], widely used in games research for this purpose [51]. We measured **behavioural engagement** using the proportion of players per condition who chose to play the game again voluntarily after they were told that they have completed enough of the study to be paid, emulating common voluntary time on task measures for intrinsically motivated engagement [66]. To measure objective challenge, we calculated the **player win margin (PWM)** as the human player's final score minus the AI opponent's final score in the last demanded game.

For descriptive results and exploratory analyses, we also calculated the percentage of players winning, losing, or drawing in the last match per condition, and **absolute PWM** as the absolute score distance from a draw; e.g., PWMs of -12 (player loses) or 12 (player wins) would both have an absolute PWM of 12. We also calculated **PWM improvement** by subtracting the PWM of a player's first match from the PWM of their last demanded match.

For proximal player experiences, we used the Player Experience Inventory (PXI) as the most well-validated, publicly available, and short recent player experience instrument [1]. It clearly separates psychological antecedents (like challenge) from consequential experiences, and we found that the PXI scales and items had the best construct and face validity for our constructs. **Perceived balance** was assessed with the PXI Challenge scale. Its items actually measure perceived balance (e.g. "The challenges in the game were at the right level of difficulty for me"), such that higher scores indicate greater perceived balance. To avoid confusion, we therefore label it perceived balance in the following. The **mastery** PXI subscale (e.g., "I felt I was good at playing this game") was used to test SDT-predicted relations between objective challenge and competence. It shows strong criterion validity for game-related competence, measured as co-variance with the PENS competence subscale [68], but avoids some of the psychometric issues of PENS [39] and is freely available. Notably, the PENS competence subscale has an item measuring balance, "My ability to play the game is well matched with the game's challenges". Using PENS to measure the impact of balance on competence would thus have been tautological. The **immersion** PXI subscale was used to test predictions related to attentive absorption in flow theory. While there are well-established general flow state scales with a separable dimension for attentive absorption (like "Concentration on Task at Hand" in the FSS-2 scale [37]), these are not freely publicly available. The PXI immersion items (e.g., "I was fully focused on the game") had strong face validity for attentive absorption and mirrored items of comparable flow scales (e.g., "I am completely focused on the task at hand" in the FSS-2 "Concentration on Task at Hand" dimension, [36]). For potential future exploratory analyses, we also recorded the remaining PXI subscales (meaning, autonomy, curiosity, ease of control, progress feedback, audiovisual appeal, goals and rules).

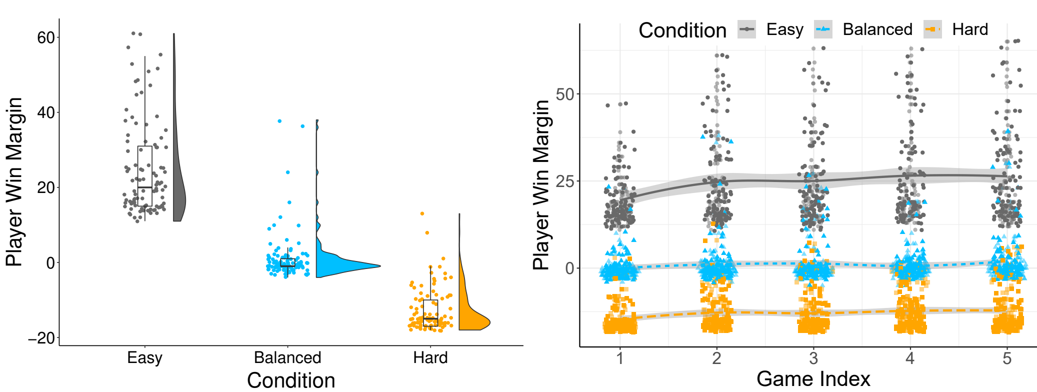
4 RESULTS

All analyses were performed using Jamovi 2.25. We label all preregistered analyses here as such. All other analyses are not labelled or labelled "exploratory". Descriptive results can be seen in Table 1.

Condition	Easy	Balanced	Hard
N	100	109	100
Target Player Win Margin (PWM)	15	0	-15
Actual game outcomes (last game)			
PWM Mean (SD)**	26.4 (13.7)	2.02 (6.81)	-12.2 (6.29)
PWM Median	21.5	0	-15
Player win % *C*	100	37.6	6
Draw % *C*	0	23.9	1
Player lose % *C*	0	38.5	93
PWM Improvement Mean (SD)	6.91 (15.5)	2.03 (7.98)	2.56 (7.23)
PWM Improvement Median	2	1	1
Participant experience and behaviour			
Enjoyment Mean (SD)*	36.7 (9.57)	37.7 (7.70)	34.1 (10.1)
Behavioural Engagement*** *C*	15	19.2	24
Perceived Balance Mean (SD)**	0.07 (-1.51)	1.27 (-1.12)	0.133 (-1.46)
Mastery Mean (SD)**	1.83 (-1.12)	0.57 (-1.5)	-1.067 (-1.40)
Immersion Mean (SD)*	1 (-1.32)	1.33 (-1.02)	0.8 (1.39)

Table 1. Descriptive results. *C* indicates categorical value without standard deviation, * indicates p<.05, ** indicates p<0.001, *** % of players who whose to play again

Exploratory scatter plots include LOESS local regression curves purely for illustration; we did not conduct exploratory tests on optimal fit with various (e.g., linear versus quadratic) functions.



(a) Player win margin in the last game. Conditions clearly differed as intended.

(b) PWM over all 5 games. Game Index 1-5 corresponds to the order that the games were played in. Our system kept game results to the target score regardless of any improvement in player ability.

Fig. 4. Manipulation check of achieved player win margin.

4.1 Manipulation Check of Objective Challenge

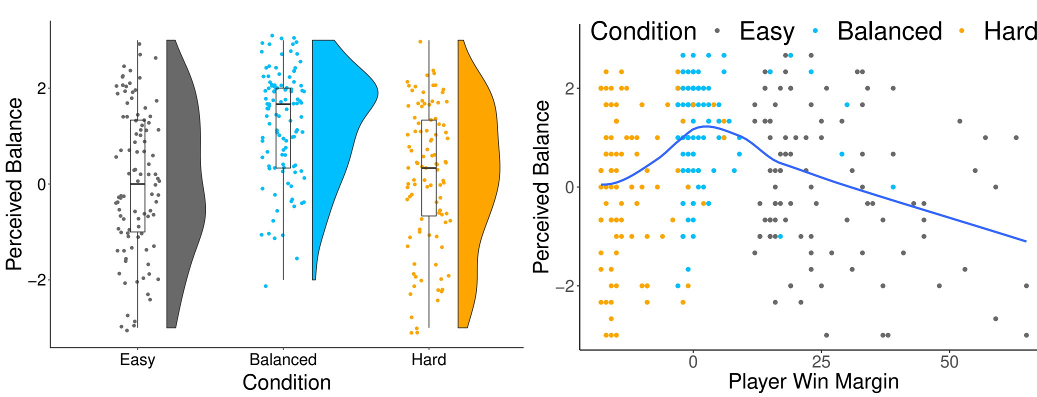
We first want to see whether our system induced objective challenge (PWM) as intended, as stated in H4 and H4a:

- ***H4***: *There will be a significant difference in player win margins between conditions in the final demanded game that participants play.*
- ***H4a***: *There will be a significant difference between each condition in the direction that the AI was aiming for.*

Our system did. In our preregistered analyses, PWMs significantly differed between conditions with an extremely large effect size (H(2)=247, p<.001, $\epsilon^2 = 0.801$); pairwise comparisons (p<.001) show a significant difference between all conditions, with median PWMs per condition very close to the target: 21.5 for Easy (target 15), 0 for Balanced (target 0), Hard -15 (target -15) (Figure 4a). This supports **H4** and **H4a** and replicates [20]. As an exploratory manipulation check, we tested whether the AI maintained target PWMs over the five matches, that is, whether it successfully adjusted to any potential player learning. This was the case, with minimal median improvements of 2 (out of 162) points in the Easy condition, 1 in Draw, and 1 in Hard (Figure 4b).

4.2 Impact of Objective Challenge on Player Experience

Next, we want to know whether objective challenge impacts proximate player experience constructs at all: ***H1***: *There will be a significant difference in game experience between the three game conditions (hard, balanced, easy), measured as a difference in at least one of the three constructs of perceived balance, competence, and attentive absorption.* This was true for two of our three constructs, supporting **H1**. After a Bonferroni correction, our preregistered analyses found both perceived balance (H(2)=44.67, p<.001, $\epsilon^2 = 0.145$) and mastery (H(2)=134, p<.001, $\epsilon^2 = 0.434$) to differ significantly between conditions, but not immersion, which differed between conditions (H(2)=7.2, p=0.027,$\epsilon^2 = 0.434$), but with an alpha above the Bonferroni-corrected threshold of significance ($p = 0.016$).

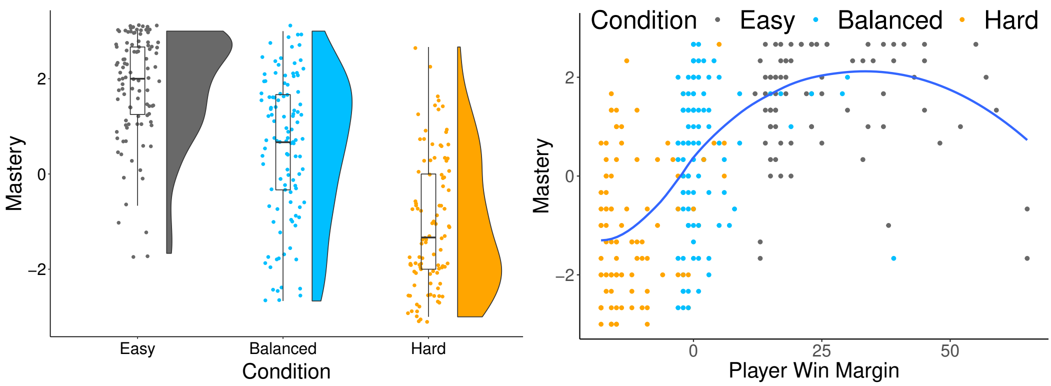


(a) Perceived Balance per condition (b) Perceived Balance and PWM across conditions

Fig. 5. Objective Challenge and Perceived Balance.

4.3 Impact of Objective Challenge on Perceived Balance

A main open question is whether an opponent of matching strength is indeed perceived as most balanced: ***H1a***: *Perceived balance will be significantly higher in the Balanced than in the Hard or Easy condition.* Supporting this hypothesis, our preregistered pair-wise comparison showed that perceived balance was significantly higher (p<.001) in the Balanced condition ($M = 1.27$) than in the Easy ($M = 0.07$) or Hard ($M = 0.133$) one (see figure 5a). A scatter plot of PWM against perceived balance (figure 5b) shows the expected inverted-U pattern peaking around PWM=0. A further exploratory analysis found that absolute PWM is significantly negatively correlated with perceived balance (r=-0.332, p<0.001). Taken together, this supports that our objective challenge manipulation induces concurrent differences in perceived balance. Players perceive matches with results close to a draw to be most balanced.



(a) Mastery per condition (b) Mastery and PWM across conditions

Fig. 6. Objective Challenge and Mastery.

4.4 Impact of Objective Challenge on Mastery

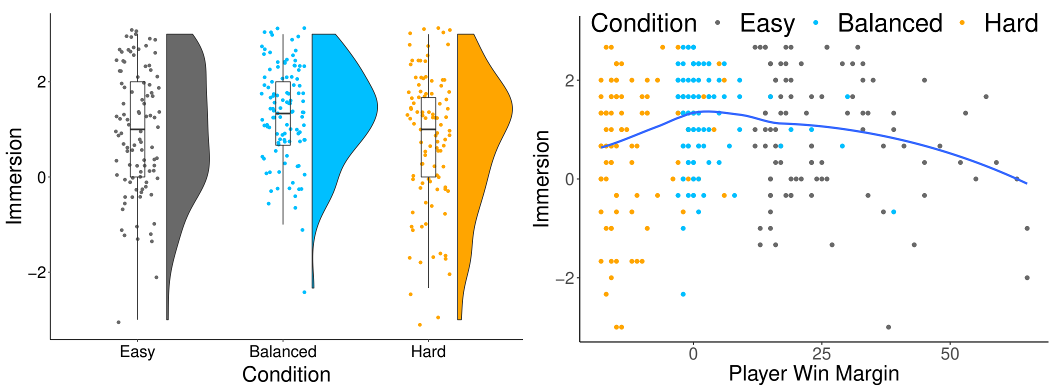
Our next two hypotheses tested two competing proximal claims of SDT about competence, here measured as mastery:

- ***H1b***: *Competence will be significantly higher in the Easy than in the Balanced or Hard condition.*
- ***H1c***: *Competence will be significantly higher in the Balanced than in the Hard condition.*

The optimal challenge claim (competence peaks at balanced challenge) would predict H1b but not H1c. The perceived competence claim would predict H1b *and* H1c. **H1b** and **H1c** were supported as our preregistered pairwise comparison showed mastery to be significantly higher in the Easy condition ($M = 1.83$) than the Balanced ($M = 0.57$) or Hard ($M = -1.067$) conditions ($p < .001$), and significantly higher in the Balanced than Hard condition ($p < .001$) (see Figure 6a). Exploratory analyses showed a large correlation between PWM and mastery across conditions ($r = 0.584, p < 0.001$). The scatter plot (Figure 6b) suggests a linear relation between mastery and PWM that peaks and then 'trails off' at around a PWM of +35 into an inverted-U. However, this is based on visual inspection only, with few, highly varied data points. This aligns with SDT's perceived competence effect, but contradicts its optimal challenge account.

4.5 Impact of Objective Challenge on Immersion

Moving on to flow theory, we predicted: ***H1d*** *Attentive absorption will be significantly higher in the Balanced than in the Easy or Hard condition*, measuring absorption with the PXI immersion scale. **H1d** was not supported as our preregistered pairwise comparison found the Balanced condition to have significantly higher immersion ($M = 1.33$) than the Hard ($M = 0.8$) condition ($p = 0.024$), but not than the Easy ($M = 1.0$) condition ($p = 0.175$), see Figure 7a. Exploratory analyses found a weak negative correlation across conditions between absolute PWM and immersion ($r = -0.163, p < 0.05$), and no significant correlation between actual PWM and immersion ($r = -0.01, p > 0.05$). The scatter plot (Figure 7b) suggests a 'trailing off' of Immersion at very high PWMs, but again, this is a purely visual observation based on few, highly varied data points in the Easy condition. Overall, this suggests that objective challenge had some minor impact on immersion, but not with an inverted-U.



(a) Immersion per condition

(b) Immersion and PWM across conditions

Fig. 7. Objective Challenge and Immersion.

4.6 Impact of Objective Challenge on Enjoyment

Both flow theory and SDT predict that challenge impacts enjoyment such that it is maximised under perceived balanced or 'optimal´ challenge:

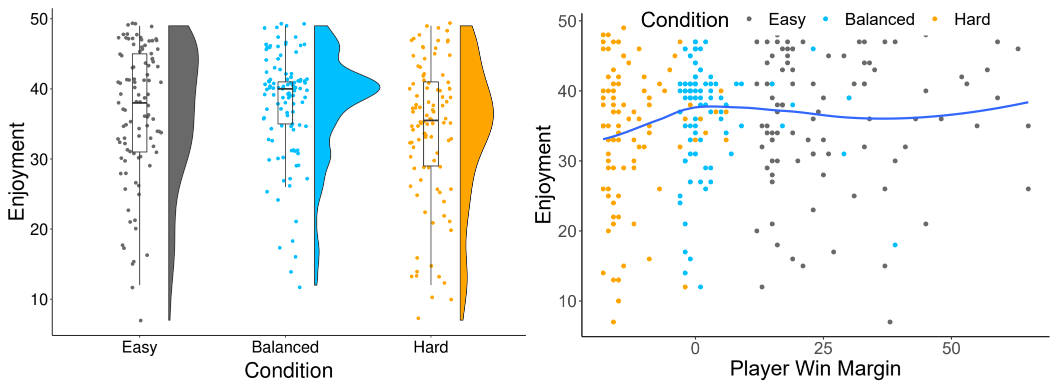
- ***H2***: *There will be a significant difference between levels of experiential enjoyment in the three different game conditions.*
- ***H2a***: *The Balanced condition will have a significantly higher experiential enjoyment than either other condition.*

H2 was supported as our preregistered analysis found a significant difference ($F(2) = 4.31, p = 0.014, \eta_p^2 = 0.027.$) in enjoyment between conditions (Figure 8a). That said, we did not see a significant inverted-U, rejecting **H2a**: the Balanced condition ($M = 37.7$) had significantly ($p = 0.012$) higher enjoyment than the Hard condition ($M = 34.1$), but did not significantly ($p = 0.678$) differ from the Easy condition ($M = 36.7$). This partially replicates [20], who found enjoyment to not significantly differ between conditions. Exploratory analyses similarly found no significant correlations between enjoyment and PWM ($r = 0.087, p > 0.05$), nor enjoyment and absolute PWM ($r = -0.046, p > 0.05$), mirrored in the scatter plot (Figure 8b).

4.7 Impact of Objective Challenge and Enjoyment on Behavioural Engagement

Finally, both theories predict that challenge should impact behavioural engagement in the form of intrinsically motivated behaviour, mediated by enjoyment:

- ***H3***: *There will be a significant difference between levels of behavioural engagement in the three different game conditions.*
- ***H3a***: *The Balanced condition will have significantly higher behavioural engagement than either other condition.*
- ***H5***: *There will be a significant correlation between experiential enjoyment and behavioural engagement.*

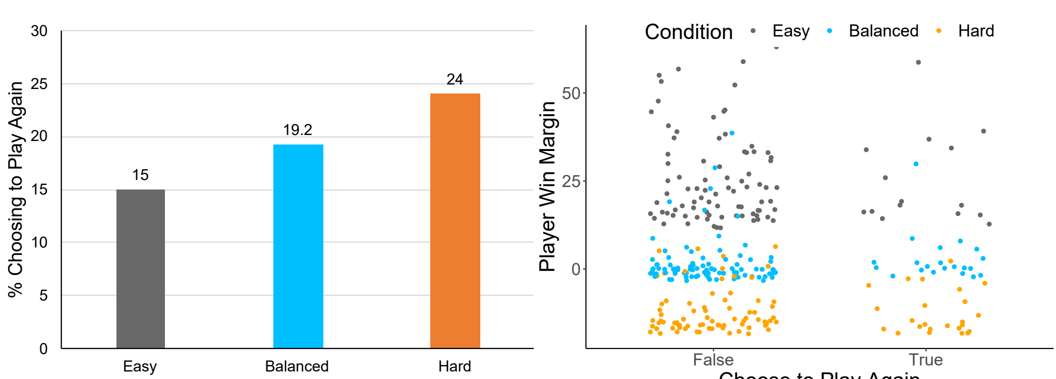


(a) Enjoyment per condition

(b) Enjoyment and PWM across conditions

Fig. 8. Objective Challenge and Enjoyment.

Rejecting **H3** and **H3a**, Easy, Balanced, and Hard conditions did not significantly differ in the percentage of participants who chose to play the game again ($\chi^2(2, 309) = 2.59, p = .274$, Figure 9a). In absolute numbers, 'the harder' the condition, the more participants chose to play again: 24 in the Hard condition, versus 19.2 in the Balanced and 15 in the Easy condition. This data pattern mirrors Cutting et al. [20], who found 20 choosing to play again in Easy, 29 in Balanced and 30 in Hard. However, in both studies, this did not pass the chosen threshold of statistical significance. We highlight this data pattern chiefly because it goes *against* the patterns for perceived balance, mastery, and immersion. Additional exploratory analyses found similarly no significant correlations between PWM across conditions and engagement ($r^2 = 0.021, p = .16$, Figure 10), nor between absolute PWM across conditions and engagement ($r^2 = 0.002, p = 0.415$). In contrast, our pre-registered binomial logistic regression found a weak but significant correlation between enjoyment and engagement (McFadden's $r^2 = 0.0799, p < .001$), supporting **H5** and replicating [20].



(a) Engagement per condition

(b) PWM per condition split by choice to continue engaging

Fig. 9. Objective challenge and behavioural engagement.

Correlation	Easy	Balanced	Hard	All
Perceived Balance and Immersion	0.457**	0.370**	0.378**	0.430**
Perceived Balance and Mastery	0.246*	0.463**	0.615**	0.313**
Perceived Balance and Enjoyment	0.546**	0.461**	0.539**	0.508**
Perceived Balance and Engagement B	0.021!	0.02!	0.008!	0.031!
Player Win Margin and Perceived Balance	-0.253*	-0.010!	0.108!	-0.120*
Player Win Margin and Immersion	-0.172!	-0.122!	0.104!	-0.010!
Player Win Margin and Mastery	0.002!	0.127!	0.258*	0.584**
Player Win Margin and Enjoyment	0.032!	-0.0093!	0.007!	0.087!
Player Win Margin and Engagement B	0.002!	0.004!	0.000!	0.006!
Immersion and Enjoyment	0.676**	0.629**	0.482**	0.659**
Immersion and Engagement B	0.092*	0.059*	0.100*	0.076**
Mastery and Enjoyment	0.432**	0.629**	0.482**	0.456**
Mastery and Engagement B	0.000!	0.020!	0.000!	0.000!
Enjoyment and Engagement B	0.088*	0.048*	0.137**	0.080**
Absolute Player Win Margin and Perceived Balance	-0.253*	-0.035!	-0.067!	-0.332**
Absolute Player Win Margin and Immersion	-0.172!	-0.150!	-0.0.098!	-0.163*
Absolute Player Win Margin and Mastery	0.002!	0.032!	-0.185!	0.233**
Absolute Player Win Margin and Enjoyment	0.032!	-0.150!	-0.081!	-0.046!
Absolute Player Win Margin and Engagement B	0.002!	0.000!	0.000!	0.002!

Table 2. Exploratory correlations for each condition and in aggregate across conditions (All), Pearson's r unless otherwise specified. B indicates Binomial Logistic Regression (McFadden's r^2), * indicates p<0.05, ** indicates p<0.001, ! indicates p>0.05.

4.8 Exploratory Analyses of Perceived Balance and Player Experience

We calculated exploratory correlations between objective antedecents and consequences (PWM, absolute PWM, engagement) and perceived balance, enjoyment, immersion, and mastery, which are shown in Table 2 and Figure 10. Perceived balance and player experience constructs (mastery, immersion, enjoyment) significantly correlated with each other to a small to moderate extent, within and across conditions. In contrast, only mastery showed a significant moderate correlation with PWM and absolute PWM, and only across conditions, with immersion significantly but weakly correlating with absolute PWM across conditions. Only immersion had a significant (though small) correlation with engagement.

5 DISCUSSION

Our findings broadly replicate Cutting et al. in that the MCTS system produced desired objective difficulty-skill ratios, but these do not impact behaviour. And although the balanced objective challenge condition was perceived as most balanced, other player experience measures did not chart a uniform inverted-U around objective challenge. To unpack these findings, we first discuss their ramifications for flow and SDT, then the disconnect between objective and self-report measures and the need for alternative explanations for our findings. We close with a reflection on our attempt at theory-testing work in games HCI.

5.1 Implications for Flow Theory

Our findings show a robust relation between objective and perceived challenge as suggested by Csikszentmihalyi: an objective difficulty-skill ratio of 1:1 (the Balanced condition, with PWMs close

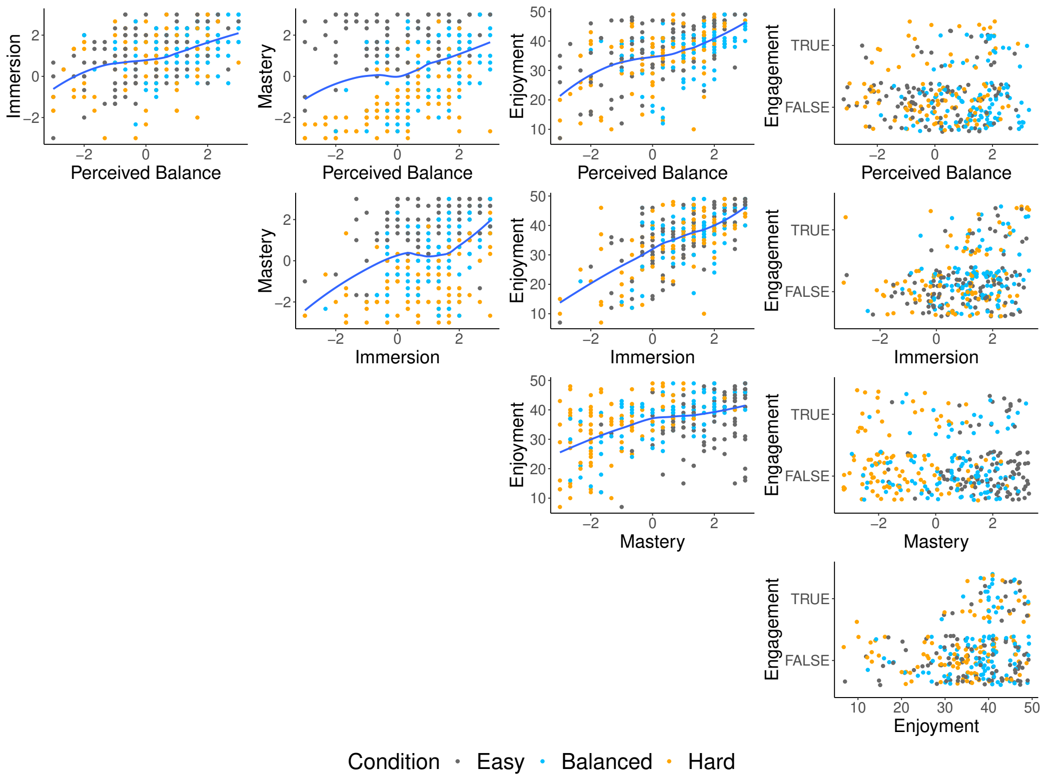


Fig. 10. Scatter plots of exploratory correlations.

to 0) is perceived as most balanced. Absolute PWM (how far away the objective end result is from a draw) is negatively correlated with perceived balance. While this result may be intuitive, it is the first direct test (to our knowledge) of this relation between objective and perceived balance. Our exploratory analyses also found perceived balance to correlate significantly with immersion and enjoyment. All this aligns with flow theory.

Contradicting theory, objectively balanced games were not significantly more enjoyable or immersive than easy games, even though they were perceived as more balanced. We also found no statistically significant links between absolute PWM and enjoyment, and neither objective nor perceived balance correlated with actual play behaviour. For flow theory, this is a puzzling result. If anything, repeated claims that flow only occurs under conditions of matched *high* skill and *high* difficulty [56] would suggest balanced and *hard* conditions to be equally high in enjoyment and immersion, not balanced and *easy*. And we would expect significant correlations of immersion and enjoyment with objective balance (absolute PWM) in the Hard condition. Yet this is not the case (see Table 2). So we are left with a dual puzzle: objective balance induces perceived balance, but neither makes an appreciable difference for immersion and enjoyment in easy vs balanced games; and neither impacts play behaviour.

One possible counter is that we simply did not manipulate objective difficulty-skill ratios strongly enough or long enough: inverted Us and behavioural differences would show if we targeted much larger PWMs, e.g., -50 and +50, and longer play duration. That is possible, though pilot testing, results for mastery, and the fact that the hard condition induced a significantly and uniformly worse player experience suggest that the current conditions already noticeably differ for players. Thus, our findings suggest at a minimum that the impact of objective balance on enjoyment and behaviour, mediated by perceived challenge and attentional absorption, is moderated in unaccounted ways.

Another possible counter is that flow theory refers to 'live', moment-to-moment experience, while we measured and manipulated summative post-hoc game outcomes and player experiences. Evidence suggests that the latter are not a mere faithful mirror of the former [45]. The two flow studies we know using mid-game experience sampling [4, 5] did find inverted-Us. However, we think that our method evades this counter at least for our manipulation, as it manipulates objective challenge moment-to-moment: every turn, the AI assess the current game state at that time and chooses a move anew that creates a game state most likely leading to a player win, AI win, or draw, depending. Thus, a player *should* perceive, turn-by-turn, a weaker, evenly matched, or superior opponent. But we grant that this is an assumption we need to confirm with, e.g., a comparable mid-game experience sampling study. Still, if flow predictions only hold moment-to-moment, that would have appreciable implications for research and practice, where summative post-hoc assessments and metrics focusing summative game outcomes are still common. We are not aware of existing specified models or predictions in flow theory whether it only applies to lived momentary or also post-hoc remembered experience, or how the former shapes the latter. Absent such specification, we cannot strictly say whether our findings contradict flow theory – but they may show an unacknowledged limit to its generalisability.

5.2 Implications for Self-Determination Theory

Our data showed a linear relation that fits SDT's *perceived competence effect* [67, p. 156] prediction, recent conceptual analyses [26] and was also found in other studies [5]: the easier the condition or the greater the outcome PWM, the more mastery players reported. Because SDT does not formally specify its verbal account of *optimal challenge* as "demands that most often one can master" [67, p. 153], it is hard to say whether our findings falsified it. They do contradict the detail claim that "[a]ctivities that are trivial or simple and therefore provide no challenge are not expected to be intrinsically interesting even if the person perceived him- or herself to be extremely competent." [22, p. 58]. After all, the Easy condition, where all players won 100 percent of the time, five matches in a row, induced the most mastery and enjoyment equal to the balanced condition. While not statistically significant, mastery visually 'peaked' around a PWM of +35 in our data. This suggests that in terms of competence satisfaction, *optimal challenge*, if such a thing exists, could sit around a point of objectively highly frequent and extensive success far above the point players perceive as optimally balanced.

More surprisingly, while mastery correlated with enjoyment (as predicted by SDT), we found no significant correlation between PWM or mastery and behavioural engagement, also contrasting Lomas' [47] finding that players persisted longer the easier the game was. This does not even fit a perceived competence effect logic that lower challenge leads to higher competence, which leads to higher enjoyment, leading to higher behavioural engagement. Not only did more mastery *not* lead to more play. Although not significant, both Cutting et al. [20] and our data showed that the harder the condition, the more participants chose to play again.

We thus need to assume that major *other* variables impacted enjoyment or engagement, or moderated their link with competence. This is a far cry from 'disproving' SDT. But it holds three ramifications: One, between the two competing SDT predictions of optimal challenge and perceived competence effect, our data only fits the latter. Two, competence experience, and difficulty-skill ratios inducing it, matter less (or less directly, pending unknown moderators) for play behaviour than previously thought. Three, current verbal articulations of SDT are underspecified to guide games HCI research and practice on difficulty-skill balance.

5.3 The Disconnect of Challenge, Experience, and Behaviour

Maybe the most interesting data pattern is the partial disconnect between challenge, player experience, and behaviour. No measure of objective or perceived challenge significantly correlated with behavioural engagement. And yet all self-report measures – perceived balance, mastery, immersion, enjoyment – significantly cross-correlate with each other. The objectively hard condition was significantly *less* enjoyable, immersive, mastery-inducing, and yet the condition where the most players chose to play more (though not significantly so). How can we explain this?

One possibility is that the causal arrow between player experience and perceived balance is reversed: Whatever afforded positive experiences, when players had one, this led them to post-hoc report that, to quote one PXI Challenge item, "The challenges in the game were at the right level of difficulty for me." We are not aware of existing studies testing this plausible alternative explanation. This would indicate that objective challenge is even less directly relevant for player experience.

A strong version of this contention is that our self-report-based analyses suffer from unconsidered crud (all variables in correlational research in the social sciences necessarily correlate to some extent [59]), or jangle fallacies [29]: No matter if we label it 'mastery', 'immersion', 'enjoyment', 'challenge', etc., we are effectively giving different names (and scales) to a single phenomenological entity of 'positive player experience' – or at best, verbally-analytically but not empirically separable dimensions thereof. Our data could then be explained by multiple factors causing or moderating omnibus positive player experience, while the same (or other) factors separately cause behavioural engagement. Objective challenge then becomes one (among many) antecedent for perceived balance and mastery. Enjoyment as a commonly used omnibus construct for positive player experience would then have *some* but little *direct* causal bearing on player behaviour, explaining its small ($r = 0.08$) correlation with engagement.

If this were the case, one upshot for research on difficulty-skill ratios would be to manipulate objective ratios and measure actual play behaviour alongside player experience (as done here) to tease out whether, when, and how much player experience (and its factors) make a difference.

5.4 Alternative Explanations

Another upshot would be to start looking for uncontrolled factors. First, player appraisals could impact behaviour: there is some evidence that players selectively choose game tasks based on how they assess their own skill (and thus, relative success odds) [48]. Most behavioural evidence on optimal challenge in SDT comes from paradigms where participants voluntarily chose to work on tasks labelled moderately difficult over those labelled easy or hard, not paradigms where perceived enjoyment or competence at t predicted task choice at $t + 1$ [67, p. 153]. However, if game selection were based on *expected* enjoyment or immersion or mastery as a consequence of *appraised* skill and difficulty, given a learning history of five matches where participants consistently won, we would still expect participants in our Easy condition to have chosen playing again most, which was not the case. So what could players be sensitive to instead?

Several researchers have proposed curiosity and its collative variables [5, 26, 48]. Following this broadly information-theoretical view, people seek out and enjoy optimal levels of stimulus novelty, uncertainty, complexity, or incongruity as that which maximises information gain [10, 41]. This would also fit our data on immersion or attentive absorption, as attention is strongly steered by curiosity [33]. Following this line, *Explorers vs Owls* featured such interesting (novel, complex) gameplay that this overshadowed any true challenge impacts on engagement, while (appraised and expected) outcome uncertainty was so low for very strong players in the Easy condition that they became less attentively involved and engaged. This would align with Sherry's broader media flow theory [71], suggesting that media 'challenge' cognitive capacities also by way of novelty, complexity, etc.

A final recent explanation is that players enjoy and seek out not absolute success, novelty, or uncertainty, but dynamic *progress*: doing better, that is improving, learning, or reducing uncertainty faster than expected [26]. This could easily incorporate the above-mentioned novelty of gameplay. Now our MCTS system actively suppressed objective match-to-match progress in terms of PWMs. Yet players may still have made objective and perceived progress in understanding the game and learning and executing tactics. Players could have discovered and tried, e.g., an encircling strategy and saw it 'work' and improve their scoring or board position two to three moves in. This would register as progress in the same way that *Dark Souls* players report learning progress even in dying repeatedly against a boss monster [60]. In our study, the more challenging the condition, the more it might have required and made salient such move-level progress, whereas the Easy condition allowed players to 'coast' and not learn, or strong players to 'grok' the game over the five matches, leaving no actual and expected opportunities for progress or doing better than expected. The Hard condition in contrast may have offered more perceived-resolvable uncertainty at the time of measurement. While players in aggregate dislike losing repeatedly (= making zero progress), they may still have expected that they might win the next game after this losing streak, which would thus mark massive progress. This would explain why the Hard condition had a significantly worse player experience and yet marginally higher voluntary play engagement.

5.5 Implications for Theory-Testing Games HCI

We positioned this paper as an attempt at more severe hypothesis-testing work guided by and developing theory in games HCI. On reflection, we consider it a mixed success. We followed many practices promoted in open science and scholarship to increase the severity of hypothesis-testing, such as preregistering hypotheses and data collection and analysis plans. We chose an experimental paradigm that allows for a transparent, replicable, and formally quantified manipulation. However, we quickly ran into a major issue: Neither of our two tested theories was specified enough to easily derive predictions that could falsify theory. This was most obviously the case for objective balance itself: neither theory offered an unambiguous, quantifiable, easy-to-operationalise prediction of 'optimal' objective difficulty-skill ratios that was logically derived from theory and could thus falsify theory if found not to hold. Second, both theories contain alternative specifications. As we saw, SDT proposes two *contradictory* relations between objective challenge and competence (optimal challenge and perceived competence effect). To make SDT testable (and applicable), we need theoretical work formally respecifying its account of challenge.

Flow theory is currently similarly used in at least two variants, one only requiring perceived balance [19, p. 187], the other perceived matching *and* 'high' difficulty and skill [56] without specifying 'high' further. The wide range of difficulty-skill ratio operationalisations in the literature mirror this underspecification [30]. We did not account for this duality when we designed our study and hypotheses. Future work here needs to be specific which variant it concerns, or explicitly test both to infer to the better explanation. We also need work formally or empirically specifying what objective difficulty and skill levels are perceived as high.

Another major shared underspecification arguably holds for a wider range of games HCI work: do they only apply to concurrent, moment-to-moment play experience, or also to post-hoc summative remembered experience? And in either case, on what scales of time and generality? Games HCI regularly studies moment-to-moment play, a player's most recent remembered play episode, their overall experience with a game, or their overall experience playing games over extended periods. Theory and evidence variously suggest that these levels of time and generalisation scales and lived versus remembered experience differ [43, 53, 80]. So if we find a data pattern that doesn't fit a theory at a given scale or with a given data source (remembered, not lived), when and why does that 'merely' limit its generalisability, and when can it be considered a falsification? We frankly don't know. But to us, it brings into sharp and humbling relief that if we want to engage in theory-testing and developing games HCI, we need to start articulating theory and evidence mindful of this complexity.

Now this issue of theoretical underspecification is not new or unique to games HCI: in the social and behavioural sciences, the so-called theory crisis discourse [58, 61, 81] concurrently notes that increased methodological rigour will not advance knowledge generation without matching rigour in specifying theories that allow us to identify relevant and testable hypotheses. And as it turns out, many existing social and psychological theories are seen to lack sufficient (formal) specification [58, 81]. On a practical note, if we were to re-run the present study, we would likely deliberately choose and specify such boundary conditions in advance, guided by some principle of maximal charity: What would be the *optimal* conditions for a predicted effect to show? For instance, we could have tested flow theory using experience sampling of lived experience, as flow theory proponents would likely grant that this is an optimal setup.

Our findings show that difficulty-skill balance as modelled in flow theory and SDT is not sufficient to predict behavioural engagement and enjoyment, at least with novel games over relatively short play periods. That directs us toward searching for alternative factors – and to a final reflection. In the theory-*using* mode prevalent in HCI [64], if we find a data pattern not matching expectations, we can be tempted to search for and switch to a *different* explanation or *new* boundary condition taken from another theory. After all, we want to *use* theory to *explain* data (or gaming phenomena). Player appraisal, novelty and curiosity, or sense of progress could explain our data (see above). The important question is: what happens then? Do we leave it at that and flexibly use a now slightly broader selection of theory to explain the next phenomenon – or do we steer toward follow-on re-specifying and re-testing of the original theory? Flow theory claims to be a general theory of enjoyment and intrinsic motivation. Thus, introducing novelty or attenuation time would amount to a significant moderator, extension, or limitation, with further significant implications. For instance, we should control for novelty when testing for the impact of challenge, or consider only working with games our participants are well-familiar with. We may want to rearticulate and re-test theory on how novelty impacts attentive absorption in addition to challenge. We should establish when a game becomes familiar enough that novelty no longer matters. This is a much slower, more incremental, less novelty-prone style of work than we ourselves are certainly used to. It requires agreeing on and using consistent and reliable operationalisations and measures. And it spans many coordinated studies that form a wider empirical and theory cycle, where both theoretical and qualitative work have important roles to play [81]. While the present singular paper obviously cannot deliver on this, we see its contribution in foregrounding some concrete dimensions of specificity and generalisation that games HCI will need to grapple with when tackling such a programme of work.

5.6 Implications for Design?

In the same spirit, we hesitate to jump to strong practice implications. If our results replicate, player enjoyment and engagement may be far more forgiving to 'poor' difficulty-skill balance than thought. This would not devalue balancing as a more holistic practice [70], but direct discretionary design attention to factors other than difficulty-skill ratios alone. Where designers do manual or automated difficulty adjustment, our findings presently caution against using only self-report measures, especially of perceived balance. Until psychological mediators between objective balance and play behaviour are better established, behavioural metrics may be a prudent additive or alternative. We would also caution that summative post-hoc assessments and moment-to-moment assessments of balance and its impacts of player experience may diverge. For dynamic difficulty adjustment systems [85], our findings suggest that the oft-proposed optimum of 50% winning odds is empirically not a given (see similarly [47, 48]); developers would do well to agnostically explore which difficulty-skill ratios maximise engagement for their game audience. For matchmaking systems [34, 52], we don't want to derive recommendations, as human-v-human play involves important other player concerns like fairness.

5.7 Limitations

One limitation we already flagged: future work should test whether our assumption holds that we *de facto* also manipulated moment-to-moment objective and perceived balance, e.g. via mid-game experience sampling and board state scoring.

Second, we used a paid online sample, with questionable ecological validity for unpaid, leisurely gameplay. Our attention checks indicate that our sample responded with some care. Still, participants may have chiefly wanted to maximise their hourly payment and minimise time on task, extinguishing player experience impacts on engagement that would have shown in leisurely play.

Third, we measured engagement with a dichotomous variable: choosing to play another match (or not). This is conceptually similar to voluntary time on task measures, but ratio variables (time spent) may be more sensitive. Similarly, while we consider the chosen IMI and PXI subscales to have good validity for the constructs of interest, critical readers may wish to see replications using immediately SDT- or flow-derived scales like FSS-2.

Fourth, we tested a single-player, turn-based, tactical game with relatively short play, a single game session, and a representative population sample. Our findings may not generalise to longer sessions, repeat engagement, multiplayer settings, other genres, or player communities. Balance may particularly impact extensive repeat play and multiplayer matches more, where it may affect additional mediators like perceived fairness.

Finally, our pre-registered tests were limited to three pre-specified conditions, based on the assumption that matches close to a draw manifest the peak of the theorised inverted-U. This was borne out by our data, but one could argue that there are likely inverted-U patterns, only at different points and testable with different analyses. Future work could, e.g., purposely evenly populate a wider PWM range and test whether quadratic functions offer the comparatively best fit with the data.

6 CONCLUSIONS

Difficulty-skill balance is a long-standing concern of games HCI, where both flow theory and SDT predict an inverted-U pattern of positive player experience and engagement around an 'optimal' midpoint. Yet when it comes to objective not perceived balance, prior work paints a mixed picture and has not directly tested key theoretical propositions. In response, we conducted a large-scale preregistered study using a game with AI-controlled difficulty to severely test flow theory and SDT predictions.

Our findings make several contributions to games HCI. First, we show that players perceive opponents with objectively matching play strength as the most balanced. This common intuition was not empirically tested before, and can anchor future work on objective difficulty-skill balance. Second, we show that over shorter play duration, objective and perceived balance do not significantly impact behavioural engagement. This contradicts standard views that balance is "the most important aspect of good game design" [78]; we identify player expectations, novelty and curiosity, and sense of progress as factors potentially explaining engagement for future research. Connected, we don't find a uniform inverted-U pattern of positive player experience and engagement peaking around the midpoint of an equally strong opponent. This cautions designers to not treat difficulty-skill balance as the single most important factor of game enjoyment and engagement, nor the often-floated 50% success odds as a well-evidenced optimum, at least in single-player games. We also note a partial disconnect between objective game features and play behaviour and self-reported player experience, encouraging more empirical research manipulating and measuring objective design variants and play behaviour to guard against potential self-report crud and jangle fallacies. Third, we find partial support for flow theory in that attentive absorption correlates with perceived balance, enjoyment, and weakly but significantly with engagement, but also partial conflicting evidence (easy and balanced condition do not significantly differ in absorption, enjoyment, or engagement). At a minimum, this suggests potential, to-be-tested limits of generalisation for flow theory, namely that balance effects may only manifest in well-familiar games or lived, moment-to-moment experience. Fourth, our data challenges the SDT account of difficulty-skill balance: while it aligns with SDT's perceived competence effect, it contradicts its optimal challenge account, and shows no significant relation between competence satisfaction and behavioural engagement. Finally, we found both flow theory and SDT not well-specified enough to allow severe testing, as they make several contradicting claims and lack formally derivable quantification of key constructs. This suggests that the need for more formal theory work highlighted in the wider theory crisis debate [58, 81] may also apply games HCI.

ACKNOWLEDGMENTS

This work was supported by the Digital Creativity Labs jointly funded by EPSRC/ AHRC/InnovateUK under grant no EP/M023265/1 and a University of York EPSRC Impact Accelerator Account Proof of Concept Award. S.D. declares that the work has no relation with his current position and work at Amazon UK Services Ltd.

REFERENCES

[1] Vero Vanden Abeele, Katta Spiel, Lennart Nacke, Daniel Johnson, and Kathrin Gerling. 2020. Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences. *International Journal of Human-Computer Studies* 135 (2020), 102370.

[2] Sami Abuhamdeh. 2020. Investigating the "flow" experience: Key conceptual and operational issues. *Frontiers in psychology* 11 (2020), 158.

[3] Sami Abuhamdeh and Mihaly Csikszentmihalyi. 2012. Attentional involvement and intrinsic motivation. *Motivation and Emotion* 36, 3 (2012), 257–267. https://doi.org/10.1007/s11031-011-9252-7

[4] Sami Abuhamdeh and Mihaly Csikszentmihalyi. 2012. The Importance of Challenge for the Enjoyment of Intrinsically Motivated, Goal-Directed Activities. *Personality and Social Psychology Bulletin* 38, 3 (2012), 317–330. https://doi.org/10.1177/0146167211427147

[5] Sami Abuhamdeh, Mihaly Csikszentmihalyi, and Baland Jalal. 2015. Enjoying the possibility of defeat: Outcome uncertainty, suspense, and intrinsic motivation. *Motivation and Emotion* 39, 1 (2015), 1–10.

[6] Sultan A Alharthi, Olaa Alsaedi, Phoebe O. Toups, Theresa Jean Tanenbaum, and Jessica Hammer. 2018. Playing to wait: A taxonomy of idle games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–15.

[7] Nick Ballou. 2023. A Manifesto for More Productive Psychological Games Research. *Games: Research and Practice* 1, 1 (2023).

[8] Nick Ballou, Sebastian Deterding, April Tyack, Elisa D Mekler, Rafael A Calvo, Dorian Peters, Gabriela Villalobos-Zúñiga, and Selen Turkay. 2022. Self-Determination Theory in HCI: Shaping a Research Agenda. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–6.

[9] Alexander Becker and Daniel Görlich. 2020. What is Game Balancing? - An Examination of Concepts. *ParadigmPlus* 1, 1 (2020), 22–41. https://doi.org/10.55969/paradigmplus.v1n1a2

[10] D. E. Berlyne. 1960. *Conflict, arousal, and curiosity*. McGraw-Hill, New York.

[11] Mark Blythe and Marc Hassenzahl. 2003. The semantics of fun: Differentiating enjoyable experiences. In *Funology*. Springer, 91–100.

[12] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. 2012. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games* 4, 1 (2012), 1–43.
[13] Jenova Chen. 2007. Flow in games (and everything else). *Commun. ACM* 50, 4 (2007), 31–34.
[14] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. Hark no more: On the preregistration of chi experiments. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–12.
[15] Tom Cole, Paul Cairns, and Marco Gillies. 2015. Emotional and functional challenge in core and avant-garde games. In *Proceedings of the 2015 annual symposium on computer-human interaction in play*. 121–126.
[16] Günter Cornett and Alvydas Jakeliunas. 2003. *Hey, That's My Fish!* https://boardgamegeek.com/boardgame/8203/hey-s-my-fish
[17] Mihaly Csikszentmihalyi. 1991. *Flow: The psychology of optimal experience*. HarperPerennial New York.
[18] Mihaly Csikszentmihalyi. 2014. *Flow and the Foundations of Positive Psychology*. Springer Netherlands, Dordrecht.
[19] Mihaly Csikszentmihalyi and Jeanne Nakamura. 2010. Effortless attention in everyday life: A systematic phenomenology. *Effortless attention: A new perspective in the cognitive science of attention and action* (2010), 179–189.
[20] Joe Cutting, Sebastian Deterding, Simon Demediuk, and Nick Sephton. 2023. Difficulty-skill balance does not affect engagement and enjoyment: a pre-registered study using artificial intelligence-controlled difficulty. *Royal Society Open Science* 10, 2 (2023), 220274.
[21] Joe Cutting, David Gundry, and Paul Cairns. 2019. Busy doing nothing? What do players do in idle games? *International journal of human-computer studies* 122 (2019), 133–144.
[22] Edward L. Deci and Richard M. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior*. Plenum Press, New York.
[23] Simon Demediuk, Marco Tamassia, William L Raffe, Fabio Zambetta, Xiaodong Li, and Florian Mueller. 2017. Monte carlo tree search based algorithms for dynamic difficulty adjustment. In *2017 IEEE conference on computational intelligence and games (CIG)*. IEEE, 53–59.
[24] Alena Denisova, Paul Cairns, Christian Guckelsberger, and David Zendle. 2020. Measuring perceived challenge in digital games: Development & validation of the challenge originating from recent gameplay interaction scale (CORGIS). *International Journal of Human-Computer Studies* 137 (2020), 102383. https://doi.org/10.1016/j.ijhcs.2019.102383
[25] Alena Denisova, Christian Guckelsberger, and David Zendle. 2017. Challenge in Digital Games: Towards Developing a Measurement Tool. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 2511–2519. https://doi.org/10.1145/3027063.3053209
[26] Sebastian Deterding, Marc Malmdorf Andersen, Julian Kiverstein, and Mark Miller. 2022. Mastering uncertainty: A predictive processing account of enjoying uncertain success in video game play. *Frontiers in Psychology* 13 (2022), 924953. https://doi.org/10.3389/fpsyg.2022.924953
[27] Florian Echtler and Maximilian Häußler. 2018. Open source, open science, and the replication crisis in HCI. In *Extended abstracts of the 2018 CHI conference on human factors in computing systems*. 1–8.
[28] Arpad E Elo. 1978. *The rating of chessplayers, past and present*. Arco Pub.
[29] Jessica Kay Flake and Eiko I Fried. 2020. Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science* 3, 4 (2020), 456–465. https://doi.org/doi.org/10.1177/25152459209523
[30] Carlton J. Fong, Diana J. Zaleski, and Jennifer Kay Leach. 2015. The challenge–skill balance and antecedents of flow: A meta-analytic investigation. *The Journal of Positive Psychology* 10, 5 (2015), 425–446. https://doi.org/10.1080/17439760.2014.967799
[31] Tracy Fullerton. 2008. *Game Design Workshop: A Playcentric Approach to Creating Innovative Games*. Morgan Kaufman, Amsterdam et al.
[32] Fantasy Flight Games. 2011. *Hey, That's My Fish!* https://www.fantasyflightgames.com/en/products/hey-thats-my-fish-digital-device-application/
[33] Jacqueline Gottlieb, Pierre Yves Oudeyer, Manuel Lopes, and Adrien Baranes. 2013. Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences* 17, 11 (2013), 585–593. https://doi.org/10.1016/j.tics.2013.09.001
[34] Thore Graepel and Ralf Herbrich. 2006. Ranking and matchmaking. *Game Developer Magazine* 25 (2006), 34.
[35] Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill™ : A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems*, Vol. 19. MIT Press.
[36] Susan A Jackson and Robert C Eklund. 2002. Assessing flow in physical activity: the flow state scale-2 and dispositional flow scale-2. *Journal of sport & exercise psychology* 24, 2 (2002).
[37] Susan A Jackson, Andrew J Martin, and Robert C Eklund. 2008. Long and short measures of flow: The construct validity of the FSS-2, DFS-2, and new brief counterparts. *Journal of Sport and Exercise Psychology* 30, 5 (2008), 561–587.

[38] Kyros Jalife, Casper Harteveld, and Christoffer Holmgård. 2021. From Flow to Fuse: A Cognitive Perspective. *Proceedings of the ACM on Human-Computer Interaction* 5, CHI PLAY (2021), 1–30.
[39] Daniel Johnson, M John Gardner, and Ryan Perry. 2018. Validation of two game experience scales: the player experience of need satisfaction (PENS) and game experience questionnaire (GEQ). *International Journal of Human-Computer Studies* 118 (2018), 38–46.
[40] Robert M. Kaplan and Veronica L. Irvin. 2015. Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. *PLOS ONE* 10, 8 (2015), e0132382. https://doi.org/10.1371/journal.pone.0132382
[41] Celeste Kidd and Benjamin Y. Hayden. 2015. The Psychology and Neuroscience of Curiosity. *Neuron* 88, 3 (2015), 449–460.
[42] Madison Klarkowski, Daniel Johnson, Peta Wyeth, Mitchell McEwan, Cody Phillips, and Simon Smith. 2016. Operationalising and evaluating sub-optimal and optimal play experiences through challenge-skill manipulation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5583–5594.
[43] Christoph Klimmt. 2006. *Computerspielen als Handlung: Dimensionen und Determinanten des Erlebens interaktiver Unterhaltungsangebote*. Vol. 2. Herbert von Halem Verlag.
[44] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User modeling and user-adapted interaction* 22, 4 (2012), 441–504.
[45] Shringi Kumari, Sebastian Deterding, and Jonathan Freeman. 2019. The Role of Uncertainty in Moment-to-Moment Player Motivation : A Grounded Theory. In *CHI PLAY '19 Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. ACM Press, New York. https://doi.org/10.1145/3311350.3347148
[46] Daniel Lakens. 2019. The value of preregistration for psychological science: A conceptual analysis. 62, 3 (2019), 221–230.
[47] Derek Lomas, Kishan Patel, Jodi L Forlizzi, and Kenneth R Koedinger. 2013. Optimizing challenge in an educational game using large-scale design experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 89–98.
[48] J Derek Lomas, Kenneth Koedinger, Nirmal Patel, Sharan Shodhan, Nikhil Poonwala, and Jodi L Forlizzi. 2017. Is difficulty overrated?: The effects of choice, novelty and suspense on intrinsic motivation in educational games. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. ACM, 1028–1039.
[49] Colleen Macklin and John Sharp. 2016. *Games, Design and Play: A Detailed Approach to Iterative Game Design*. Addison-Wesley, Boston, MA.
[50] Chris McEntee. 2012. Rational Design: The Core of Rayman Origins. http://www.gamasutra.com/view/feature/167214/rational_design_the_core_of_.php
[51] Elisa D Mekler, Julia Ayumi Bopp, Alexandre N Tuch, and Klaus Opwis. 2014. A systematic review of quantitative studies on the enjoyment of digital entertainment games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 927–936.
[52] Josh Menke. 2016. Skill, Matchmaking, and Ranking Systems Design. https://www.gdcvault.com/play/1023014/Skill-Matchmaking-and-Ranking-Systems
[53] Talya Miron-Shatz, Arthur Stone, and Daniel Kahneman. 2009. Memories of yesterday's emotions: Does the valence of experience affect the memory-experience gap? *Emotion* 9, 6 (2009), 885.
[54] Marcus R. Munafò, Brian A. Nosek, Dorothy V.M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric Jan Wagenmakers, Jennifer J. Ware, and John P.A. Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1, 1 (2017), 1–9. https://doi.org/10.1038/s41562-016-0021
[55] Fiona Fui-Hoon Nah. 2004. A study on tolerable waiting time: how long are web users willing to wait? *Behaviour & Information Technology* 23, 3 (2004), 153–163.
[56] Nakamura, Jeanne and Csikszentmihalyi, Mihaly. 2002. The Concept of Flow. In *Handbook of positive psychology*. Oxford University Press, Oxford, 89–105.
[57] Leif D Nelson, Joseph Simmons, and Uri Simonsohn. 2018. Psychology's renaissance. *Annual review of psychology* 69 (2018), 511–534.
[58] Klaus Oberauer and Stephan Lewandowsky. 2019. Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review* 26, 5 (2019), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2
[59] Amy Orben and Daniël Lakens. 2020. Crud (Re)Defined. *Advances in Methods and Practices in Psychological Science* 3, 2 (2020), 238–247. https://doi.org/10.1177/2515245920917961
[60] Serge Petralito, Florian Brühlmann, Glena Iten, Elisa D Mekler, and Klaus Opwis. 2017. A good reason to die: how avatar death and high challenges enable positive experiences. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 5087–5097.
[61] Travis Proulx and Richard D Morey. 2021. Beyond statistical ritual: Theory in psychological science. *Perspectives on Psychological Science* 16, 4 (2021), 671–681.

[62] C. Scott Rigby and Richard M Ryan. 2007. *The Player Experience of Need Satisfaction (PENS).* Technical Report. Immersyve.
[63] Scott Rigby and Richard M. Ryan. 2011. *Glued to Games: How Video Games Draw Us in and Hold Us Spellbound.* Praeger, Santa Barbara, Denver, Oxford.
[64] Yvonne Rogers. 2012. *HCI Theory: Classical, Modern, and Contemporary.* Morgan & Claypool.
[65] Miia Ronimus, Janne Kujala, Asko Tolvanen, and Heikki Lyytinen. 2014. Children's engagement during digital game-based learning of reading: The effects of time, rewards, and challenge. *Computers & Education* 71 (2014), 237–246.
[66] Richard M Ryan. 1982. Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of personality and social psychology* 43, 3 (1982), 450.
[67] Richard M. Ryan and Edward L. Deci. 2017. *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness.* Guilford Press, New York.
[68] Richard M Ryan, C Scott Rigby, and Andrew Przybylski. 2006. The motivational pull of video games: A self-determination theory approach. *Motivation and emotion* 30, 4 (2006), 344–360.
[69] Jesse Schell. 2009. *The Art of Game Design: A Book of Lenses.* Morgan Kaufman, Amsterdam et al.
[70] Ian Schreiber and Brenda Romero. 2021. *Game Balance.* CRC Press, Boca Raton, London, New York.
[71] J. L. Sherry. 2004. Flow and Media Enjoyment. *Communication Theory* 14, 4 (Nov. 2004), 328–347. https://doi.org/10.1093/ct/14.4.328
[72] Patrick E Shrout and Joseph L Rodgers. 2018. Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual review of psychology* 69 (2018), 487–510.
[73] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 7587 (2016), 484–489.
[74] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* 22, 11 (2011), 1359–1366. https://doi.org/10.1177/0956797611417632
[75] Nathan Sorenson, Philippe Pasquier, and Steve DiPaola. 2011. A generic approach to challenge modeling for the procedural creation of video game levels. *IEEE Transactions on Computational Intelligence and AI in Games* 3, 3 (2011), 229–244.
[76] Alistair Sutcliffe. 2009. Designing for user engagement: Aesthetic and attractive user interfaces. *Synthesis lectures on human-centered informatics* 2, 1 (2009), 1–55.
[77] Penelope Sweetser, Daniel Johnson, Peta Wyeth, Aiman Anwar, Yan Meng, and Anne Ozdowska. 2017. GameFlow in Different Game Genres and Platforms. *Computers in Entertainment* 15, 3 (April 2017), 1–24. https://doi.org/10.1145/3034780
[78] Penelope Sweetser and Peta Wyeth. 2005. GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)* 3, 3 (2005), 3–3.
[79] April Tyack and Elisa D Mekler. 2020. Self-Determination Theory in HCI Games Research: Current Uses and Open Questions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–22.
[80] Robert J Vallerand and Catherine F Ratelle. 2002. Intrinsic and extrinsic motivation: A hierarchical model. *Handbook of self-determination research* 128 (2002), 37–63.
[81] Iris van Rooij and Giosuè Baggio. 2021. Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science* 16, 4 (2021), 682–697.
[82] Max L Wilson, Ed H Chi, Stuart Reeves, and David Coyle. 2014. RepliCHI: the workshop II. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems.* 33–36.
[83] Max L Wilson, Wendy Mackay, Ed Chi, Michael Bernstein, Dan Russell, and Harold Thimbleby. 2011. RepliCHI-CHI should be replicating and validating results more: discuss. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems.* 463–466.
[84] Georgios N Yannakakis and Julian Togelius. 2011. Experience-driven procedural content generation. *IEEE Transactions on Affective Computing* 2, 3 (2011), 147–161.
[85] Mohammad Zohaib. 2018. Dynamic difficulty adjustment (DDA) in computer games: A review. *Advances in Human-Computer Interaction* 2018 (2018).
[86] Robert Zubek. 2020. *Elements of Game Design.* MIT Press.

Received 2023-02-21; accepted 2023-07-07