Computers in Human Behavior 124 (2021) 106891

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

The link between flow and performance is moderated by task experience

Jussi Palomäki [a,d,*], Tuisku Tammi [a], Noora Lehtonen [a], Niina Seittenranta [a], Michael Laakasuo [a], Sami Abuhamdeh [b], Otto Lappi [a], Benjamin Ultan Cowley [a,c]

[a] Cognitive Science, Department of Digital Humanities, Faculty of Arts, University of Helsinki, Siltavuorenpenger 1 A, 00012, Helsinki, Finland
[b] Department of Psychology, Marmara University, Istanbul, Turkey
[c] Faculty of Educational Sciences, University of Helsinki, Finland
[d] Helsinki University Hospital, Problem Gambling Clinic (Peliklinikka), Finland

ARTICLE INFO

Keywords:
Flow
Learning
Game
Experience
Performance

ABSTRACT

Flow is an intrinsically motivating (i.e. 'autotelic') psychological state of complete absorption in moment-to-moment activity that can occur when one performs a task whose demands match one's skill-level. Flow theory proposes that Flow causally leads to better performance, but empirical evidence for this assumption is mixed. Recent evidence suggests that self-reported Flow may not be linked to performance-levels *per se*, but instead to deviations from anticipated performance (the so-called flow deviation, or $F{\sim}d$ effect). We aimed to replicate and extend these results by employing a high-speed steering game (CogCarSim) to elicit Flow, and specifically focused on the moderating effects of learning and task experience on the $F{\sim}d$ effect. In a longitudinal design, 18 participants each played CogCarSim for 40 trials across eight sessions, totaling 720 measurements across participants. CogCarSim reliably elicited Flow, and learning to play the game fit well to a power-law model. We successfully replicated the $F{\sim}d$ effect: self-reported Flow was much more strongly associated with deviation-from-expected performance than with objective performance levels. We also found that the $F{\sim}d$ effect grew stronger with increasing task experience, thus demonstrating an effect of learning on Flow. We discuss the implications of our findings for contemporary theories of Flow.

1. Introduction

The state of Flow is a well-documented phenomenon referring to intrinsically rewarding, or *autotelic*, experiences of total involvement that people report across a wide range of activities, such as music performance, rock climbing, playing chess, high-performance driving, and sports, to name a few. Flow typically arises when skill and task challenge match, the goals of the task are clear, and performance feedback is immediate and unambiguous. Cognitively, the Flow state is characterized by complete absorption in the task, a sense of control, but a lack of self-awareness and a lack of a sense of effort, often accompanied with a loss of the sense of time (Nakamura & Csikszentmihalyi, 2002). Since its initial conceptualization by Csíkszentmihályi in 1975 (Csikszentmihalyi, 1975), Flow has been the focus of hundreds of empirical studies from a vast diversity of fields, including human-computer interaction, game design, high-performance cognition, and many others (Abuhamdeh, 2020; Cowley et al., 2020; Emerson, 1998; Swann, Keegan, Piggott, & Crust, 2012).

The widespread appeal of Flow may stem partly from how it applies directly to performance. From the beginning, it has been assumed that peak experience (in the context of goal-directed activities) also implied peak performance. This is not surprising, as the tell-tale characteristics of Flow – intense concentration on the task at hand, a lack of anxiety, etc. – would seem to be benefit performance. The 'feelings of control' that represent a core aspect of the experience were proposed to stem, at least in part, from one's objective performance, and, in a competitive setting, from the ability to outperform the opponent (Csikszentmihalyi, 1975, p. 51). This assumption that Flow was associated with heightened performance was supported by the interviews Csíkszentmihályi and his colleagues conducted with athletes, chess players, etc. who regularly

* Corresponding author. Cognitive Science, Department of Digital Humanities, Faculty of Arts, University of Helsinki, Siltavuorenpenger 1 A, 00012, Helsinki, Finland.
E-mail address: jussi.palomaki@helsinki.fi (J. Palomäki).

https://doi.org/10.1016/j.chb.2021.106891
Received 28 October 2020; Received in revised form 10 April 2021; Accepted 9 May 2021
Available online 9 June 2021

experienced Flow (Csikszentmihalyi, 1975). [1]

Despite the clear rationale for assuming a strong link between Flow and performance, however, empirical investigations of the relationship have yielded results which, taken together, defy easy interpretation. Although a number of studies found a moderate or strong positive relationship (e.g. in the context of academic performance, computer games, or sports: Chen & Sun, 2016; Engeser & Rheinberg, 2008; Schüler, 2007; Sumaya & Darling, 2018; Jackson, Thomas, Marsh, & Smethurst, 2001; Stavrou, Jackson, Zervas, & Karteroliotis, 2007), other findings suggest that the relationship might not be so simple as 'Flow implies peak performance'. For example, some studies (e.g. Garcia et al., 2019; Jin, 2012; Schüler & Brunner, 2009, in study 2), have found only a weak positive relationship, while others (e.g. Keller & Bless, 2008; Delrue et al., 2016; Engeser & Rheinberg, 2008; Keller & Blomann, 2008; Schüler & Brunner, 2009, studies 1 and 3) have found no relationship at all. Notably, a systematic review and meta-analysis by Harris, Allen, Vine, and Wilson (2020) [2] found a "consistent medium-sized relationship between flow experience and task performance" across 22 studies, yet also found that "current evidence is unable to determine the exact nature of the flow-performance relationship". Such varied results might be because, as suggested by Linden, Tops, and Bakker (2020), different kinds of tasks elicit different types of Flow; we further propose that the same task at different stages of skill development might affect the practitioner's experience of Flow differently.

When the relationship between one construct and another varies across studies (even if the relationship is typically positive), this suggests the existence of other constructs, which may moderate the relationship but have not yet been accounted for (Baron & Kenny, 1986). A recent exploratory investigation by Cowley et al. (2019) suggested one such construct – task experience, i.e. learning the skill of performing the task. Despite the fact that Flow is fundamentally associated with skilled activity, relatively little work has directly examined how Flow relates to learning to acquire a skill, or task experience. In other words, the relationship between the process of explicitly training to learn a skill, and likelihood to experience Flow, is unclear.

Cowley et al. (2019) showed that Flow is predicted by *expected* performance in a task, hinting that the process of learning a skill may play a role in how we experience Flow from exercising that skill. In Cowley et al. (2019), nine participants played a high-speed steering game designed to induce Flow by continually matching the task demand to skill. After learning to play the game for 40 trials over eight playing sessions (about three weeks), results showed that performance followed a power-law learning curve,[3] improving rapidly at first before settling to a consistent level. However, performance itself did not increase participants' self-reported Flow: instead, Flow correlated with deviation from the power-law curve. Participants reported higher Flow whenever they performed better than would be expected from their learning curve, and vice-versa for worse-than-expected performance (below, we call this the flow deviation, or *F*~*d* effect). The authors proposed that Flow in this task is not related to gained experience and skill *per se*, but to *anticipated performance* (i.e. performance *expectancy*). Flow is experienced more if performance 'exceeds expectations' set by a person's skill level, at least when skill is changing over time. However, since Cowley et al. (2019) was a pilot study, more evidence and deeper analysis is needed to develop and test this hypothesis further.

In Cowley et al. (2019), the relationship between Flow self-reports and performance was assessed by relating Flow to a learning curve obtained using performance data from *all* trials. This model includes information that is not available to the participant *in any given trial* for forming their subjective expectancy. To obtain a more accurate picture of how self-reported Flow may relate to positive or negative deviations from subjective performance expectancy (the *F*~*d* effect), we must model the relationship using information drawn only from the subjective *past*, i.e. not including performance data from any trials after a given self-report. Skilled performers differ from novices not only in their level of ability but also in what they expect (or can reasonably expect) about their performance. Similarly, when individuals develop their skills by repeatedly engaging in a task (such as playing a steering game as in Cowley et al., 2019), they also become more familiar with the task and may develop more accurate estimation of their own abilities. Thus, learning to perform a task may lead to more accurate and precise performance expectancy, and we predict this change will also alter the relationship between expectation and self-reported Flow.

To investigate the relationship between training to learn a skill and likelihood to experience Flow, we conducted a replication of Cowley et al. (2019) and a more comprehensive analysis using the pooled datasets (from Cowley et al. (2019) and its replication). The replication tested the main findings of Cowley et al. (2019) whereby a) learning to play a high-speed Flow-inducing steering game follows a power-law model, and b) Flow is linked to performance *expectancy* more so than *absolute* performance (the *F*~*d* effect). Since Cowley et al. (2019) was an exploratory study and presented no hypotheses, here we ratify their findings as the following hypotheses, to be tested on the novel dataset:

H1a. Learning to play a high-speed Flow-inducing steering game is best fit by a power-law-of-practice model (Newell & Rosenbloom, 1982).

H1b. The *F*~*d* effect: better (or worse) than expected performance is associated with more (or less) Flow.

Additionally, as explained above, we expect the evidenced task learning (H1a) to also affect the accuracy of expectation (H1b), thus leading to an interaction of task experience with *F*~*d* effect. We analyse the pooled datasets to obtain more statistical power to extend the H1 findings by proposing the following hypothesis:

H2. The *F*~*d* effect is moderated by task experience (i.e. learning) so that the effect grows stronger with increasing task experience.

2. Methods

To achieve replication, the procedure of the current study was practically identical to that of Cowley et al. (2019). This section describes the new experiment and dataset, and also the combined datasets from the current study and from Cowley et al. (2019). The experiment was carried out in accordance with the code of ethics of the world medical association (Declaration of Helsinki) for experiments involving humans. Informed consent was obtained from all participants.

2.1. Participants

A convenience sample of nine participants were recruited via emails sent to student mailing lists at the University of Helsinki (n = 9, 5 females; mean age = 24.8, range: 21–28). All reported normal or corrected-to-normal visual acuity and no history of neurological or psychiatric illness. Eight participants had a driving license and self-reported lifetime driving experience between 1000 and 300000 km (see Table 1 for details). Six participants had relatively little gaming experience; one reported playing games for 1–3 h per month, and two

[1] For example, a well-known composer described his most rewarding moments as follows: *"My hand seems devoid of myself, and I have nothing to do with what is happening. I just sit there watching it in a state of awe and wonderment."*. In a similar vein, the sense of control during Flow is echoed in one chess player's deliberation: *"I get a tyrannical sense of power. I feel immensely strong, as tho (sic) I have the fate of another human in my grasp."*(ibid., p. 51).

[2] We cite here a preprint to a paper in press; the preprint has been updated to reflect revisions to reviewers' comments and is thus peer-reviewed: Harris, D. personal communication.

[3] That is, a curve of the form $f(x) = c - ax^{\frac{1}{k}}$, where reciprocal of the exponent k produces a declining curve, c is 'performance at trial 1' (typically between 300 and 400sec), and a the scaling factor. Such curves have been shown to fit to a wide range of data in visuomotor skill-learning tasks, and termed the power-law-of-practice model (Newell & Rosenbloom, 1982).

Table 1
Participant background information.

ID	Age	Gender	Driving experience (km)	Driving license	Gaming experience
1	>=25	M	1000-10000	Yes	> one hour per week
2	<=25	M	10000-30000	Yes	> one hour per week
3	<=25	F	0-1000	No	1-3 hours per month
4	>=25	F	0-1000	Yes	> one hour per week
5	>=25	M	30000-100000	Yes	> one hour per week
6	<=25	M	30000-100000	Yes	1-3 hours per month
7	>=25	F	10000-30000	Yes	Little / none
8	>=25	M	100000-300000	Yes	1-3 hours per month
9	>=25	M	10000-30000	Yes	> one hour per week
10	>=25	F	1000-10000	Yes	Little / none
11	<=25	F	30000-100000	Yes	Little / none
12	<=25	M	10000-30000	Yes	> one hour per week
13	>=25	M	100000-300000	Yes	Little / none
14	>=25	M	10000-30000	Yes	> one hour per week
15	<=25	M	1000-10000	Yes	1-3 hours per month
16	>=25	F	0-1000	No	Little / none
17	<=25	F	1000-10000	Yes	A few times per year
18	<=25	F	30000-100000	Yes	Little / none

Note. Participant IDs 1-9 were collected in 2017 (results reported in Cowley et al. (2019)). IDs 10-18 (highlighted cells) are the new dataset collected in 2019. Age is obfuscated for anonymity.

reported playing games for at least an hour each week.

Table 1 presents these data alongside the same datapoints gathered during Cowley et al. (2019). The sample profile by and large matched the sample profile from Cowley et al. (2019) with the exception of having a better gender-balance. The combined datasets of the current study and Cowley et al. (2019) contain 18 participants (N = 18, 8 females; mean age = 25.9, range = 21–38).

During recruitment, participants were told the experiment was about gaming experience and skill development; thus, they were naïve to the study's true aims. Participants were motivated by the offer of nine culture vouchers (exchangeable in a variety of Finnish service businesses, worth 5€ each) for attending all sessions, plus two more vouchers contingent on whether they managed to improve their task performance by the end of the experiment. In fact, participants were awarded all 11 vouchers (total worth 55€) after the last session, regardless of contingencies.

2.2. Design

After signing informed consent, participants chose eight suitable session times on different days, between 8:00 and 20:00. Dates were chosen to minimise gaps between consecutive sessions (grand average gap [SD] = 2.46 days [0.66], range of participant-wise average gap: 1.28–4.14; see Fig. A.7 in the Appendix for measurement dates across all participants). All sessions took place in the same laboratory at the University of Helsinki Traffic Research Unit.

One session consisted of playing the game for five trials (see game details below). A trial consisted of one run in the game through a track of fixed length, followed by a screen presenting feedback on performance, and then self-report measures (the translated Flow Short Scale, FSS; Cowley et al. (2019); Engeser and Rheinberg (2008)) in pen-and-paper form (details in Tables A.1 in the Appendix). Run duration depended on how fast the participants could drive through the track; run durations ranged from about 2.5 to 7 min (mean = 3.1, median = 3.0). At the end of a run, the game displayed a summary containing total run duration, the number of collisions, and the run durations of the participant's ten best runs so far, sorted by run duration. Fig. 1 depicts the study procedure.

Participants' physiological signals (pupillometry, electrodermal activity, blood volume pulse, and electrooculography) were recorded on the 1st and 5-8th sessions. A 5-min baseline measure for physiological data was recorded before these sessions, and duration of sessions 1, 5–8 was about 1 h. No physiological signals were recorded during sessions 2–4, thus acting as a within-subjects control for whether the physiology sensor setup affected performance or self-reports; these sessions lasted approximately 30 min (the study of physiological measures is not the focus of the current paper and will be reported elsewhere).

2.3. Procedure

Each session was supervised by two experimenters, who stayed behind a partition wall during the trials.

At the beginning of the first session, the participants were introduced to the game and the study procedures, and asked to fill in a questionnaire on their background (age, gender, previous gaming and driving experience, and whether they have been diagnosed with disorders of, or take medication affecting, the central nervous system or eyes). Before each session participants also reported: whether they were wearing contact lenses; had taken medication, caffeine or nicotine; and level of alertness on a scale from 1 = "I do not feel rested at all" to 5 = "I feel completely rested".

In the physiological measurement sessions (1 and 5 to 8), participants were then seated on a driving seat in a quiet, dimly lit room and both physiological sensors and an eye-tracking headset were attached. After that, they were asked to sit still in the driving seat for 5 min, looking at a screen showing a dark blue color while physiological baseline was recorded. Thereafter the participants played the game for five runs, and filled the FSS in each run. The procedure for sessions 2 to 4 was otherwise identical but without physiological measures.

2.4. Game – CogCarSim

The game task (CogCarSim) is a custom high-speed steering game designed to induce Flow, as has been validated in the prior study by Cowley et al. (2019). Participants use a steering wheel to control a blue cube, whose horizontal position is directly proportional to wheel angle. The cube travels along a straight track with red/yellow stationary obstacles, as depicted in Fig. 2 .[4] The cube continually accelerates at a steady rate and slows down if an obstacle is hit. Thus, the better the participant is able to avoid obstacles, the faster the run. Participants were instructed to avoid obstacles, and thus complete the run as fast as possible.

The virtual camera had a field of view angle of horizontal 60° and vertical 32°. The camera position was 1 unit behind the cube at 4 units eye-height, and pointed forward (parallel to the plane of travel). The track was 25 units wide, the cube and obstacles all 2 units high and wide.

[4] Video of gameplay: https://figshare.com/articles/CogCarSim_game_play_video/7269395/1.

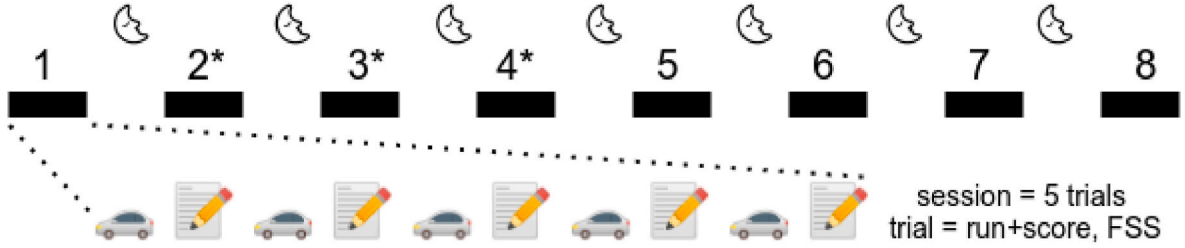

Fig. 1. The sessions took place on eight different days. Sessions without physiological measures are marked with an asterisk. One session consisted of five trials, which consist of one run of Cogcarsim, feedback on scores, and self-report questionnaire (FSS).

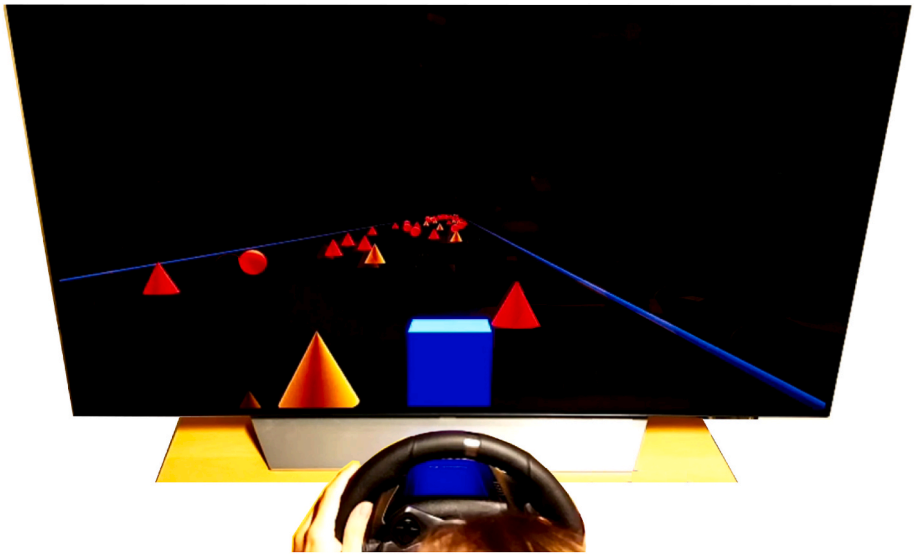

Fig. 2. The high-speed steering task. The participant steers the blue cube to avoid conical/spherical obstacles on the track, which is bounded to each side by dark blue parallel lines. The game was designed to continually adapt the difficulty level (speed) to the participant's skill (obstacle collisions). Such balance is considered one of the key antecedents of Flow. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

The track edges could not be crossed.

Track length was roughly 24200 units. For each run, a total of 2000 obstacles were placed on the track at periodic intervals of 12 units, with random lateral placement. This system ensured there was always a path through any group of obstacles.

The speed of the cube was initially set to 1.6 units per time step (96 units per second), and acceleration at 0.0012 units/step/step). The speed drop resulting from a collision with an obstacle was 0.102 units/step. Collisions that caused a speed drop were followed by an immunity period of 100 steps, wherein additional collisions had no effect. Speed drops were visually signalled with a brief screen flash.

Control of the game dynamics was deliberately simple: one degree linear holonomic control. We piloted extensively to adjust steering wheel sensitivity (steering ratio and damping).

The participants started each run by pressing a button on the steering wheel when they felt ready. At the end of each run, the elapsed time and number of collisions were displayed, along with a high score of the participant's ten best runs so far.

Data collected by CogCarSim included the positions, shape, and color of obstacles on the track; run-level aggregated performance data (run duration, number of collisions, average velocity); and within-run time series data (steering wheel and cube position, speed, registered collisions at each time step).

2.5. Flow Short Scale

To operationalize Flow, we used a modified Finnish version of a commonly used measure – the Flow Short Scale (FSS, originally developed in German (Rheinberg, Vollmeyer, & Engeser, 2007), but adapted by us from the English (Engeser & Rheinberg, 2008)). The FSS is a 10-item scale composed of two sub-scales – one intended to measure "absorption" (4 items; e.g. "I do not notice time passing"), and the other intended to measure performance "fluency" (6 items; e.g. "My thoughts/activities run fluidly and smoothly"). Answers to the questions are given on a 7-point Likert scale, varying from 'Not at all' to 'Very much'. An additional 3 item measure of perceived importance is administered with the FSS to determine the experienced importance of the given task (e.g. "I must not make any mistakes here"). Concern about failing in the task may cause anxious feelings, and therefore the perceived importance factor enables control of the possible states of anxiety. See Appendix Table A.1 for full English text, Finnish translation (modified to fit the steering game task), and back-translated English text

(to clarify for non-Finnish speakers the idiomatic meaning of the Finnish items).

In addition to the FSS and perceived importance items asked after every trial, participants were asked at the end of every session to report 3 more items measuring the fit of skills and demands of the task (Cowley et al., 2019), including a measure of perceived level of competence. These items also had 7-point scales, e.g.,: "For me personally, the current demands are … (too low—just right—too high)."

There is no consensus on how to operationalize the Flow construct, and commonly-used operationalizations have all been criticised (Abuhamdeh, 2020; Moneta, 2012; Swann, Piggott, Schweickle, & Vella, 2018). We view the FSS as a workable solution for our multi-trial design, but cognisant of the possible issues, nevertheless we withheld a priori assumptions regarding its validity. Thus, although Engeser and Rheinberg (2008) suggest using the 10-item scale as a measure of experienced Flow (as was done in Cowley et al. (2019)), we conducted an extensive validation study (N = 200) on the psychometric properties of the Finnish translation of the FSS to decide on which items to use in the scale. These analyses included a Mokken scale analysis, Parallel Analysis, Very Simple Structures analysis and a standard Confirmatory Factor Analysis. The full details of this study are reported in https://psyarxiv.com/8er92. In sum, the analyses suggested that items 1 and 3 (both from the Absorption subscale) needed to be dropped; the resulting 8-item FSS scale version had satisfactory psychometric properties.

The Cronbach's alpha for the 8-item scale (excluding perceived importance) was calculated separately for each run, and ranged between 0.70 and 0.98, thus demonstrating excellent internal consistency; See Fig. A. 6 in the Appendix for histograms of Cronbach's alphas across runs. Note that our main results were essentially unchanged even if the original 10-item version of FSS was used.[5]

2.6. Physiological measurements

Several physiological signals were recorded on a Lenovo Y520-15IKBN laptop running Ubuntu 18.04. Eye images were measured using a Pupil Labs Binocular 120 Hz eye tracking headset. Also, electrooculography (EOG), electrodermal activity, and blood volume pulse were recorded at 128 Hz sampling rate using NeXus-10 (Mind Media B. V, Roermond-Herten, The Netherlands) connected to the laptop via Bluetooth. Signals were acquired with Trusas open access software (https://github.com/jampekka/trusas-nexus).

Compared to the procedure used for Cowley et al. (2019), two minor changes to the physiology setup were introduced: (a) we added two electrodes to measure EOG and thus more accurately assess blinks, and (b) we removed the eye tracking calibration step since movements of gaze are not of interest in this task. Since changes in the procedure do not interfere with task performance or self-reports, and the added electrodes are very non-intrusive, the differences to Cowley et al. (2019) should not impact the performance and self-report data reported in this paper.

2.7. Playing equipment

The game ran on a Corsair Anne Bonny desktop with Intel i7 7700 k processor and Nvidia GeForce GTX 1080 GPU, with Windows 10 operating system. Participants were seated in a Playseat Evolution Alcantara, laterally centered in front of a 55″ LG 55UF85 monitor. The game was controlled by a Logitech G920 Driving Force steering wheel. The distance between the seat and the wheel was adjusted by the participant so that playing was comfortable. Because the distance of the seat from the monitor was adjusted to fit the player, the viewing distance was approximately the length of the player's arm reaching the wheel comfortably, plus 50 cm from wheel to monitor.

2.8. Statistical methods

All data were processed and analyzed within R platform for statistical computing (v. 3.5.2) (RCoreTeam, 2013). All data and R syntax used in the analyses are available from an open online repository which provides instructions for use https://doi.org/10.6084/m9.figshare.13567409.

To investigate questions of replication of the results of Cowley et al. (2019), we used linear mixed modelling (LMM) with the lme4-package (Bates, Mächler, Bolker, & Walker, 2014). Linear mixed modeling, also known as multilevel modeling, is a well-known method in game research to analyse data with multiple measurements from each participant (Kosunen et al., 2018; Ravaja, Saari, Salminen, Laarni, & Kallinen, 2006). LMMs have many advantages over traditional repeated measures analyses of variance, such as a better ability to deal with missing values (due to the partial pooling method).

First, for H1a (regarding power-law curves for learning) we fit a power-law of practice model: Log-transformed run duration was used as the dependent variable (DV), and log-transformed number of cumulative runs as the independent variable (IV). A participant identifier (ranging from 1 to 9) was used as a random effect allowing for variability in the intercepts and slopes. For H1b (relating Flow to observed deviation from the power-law curve, the $F{\sim}d$ effect), participant-wise deviation scores were obtained by subtracting predicted run durations, i.e. fitted values of the model in H1a, from observed run durations. These deviation scores (essentially power-law model residuals), were used as the IV in an LMM with non-standardized Flow scores as the DV, and participant identifier (1–9) as a random effect allowing intercepts and slopes to vary. These analyses for H1a and H1b are fully identical to those in Cowley et al. (2019). For additional analyses on participants' background variables, see Figs. A.3, A.4, and A.5 in the Appendix.

For H2 (how experience alters the $F{\sim}d$ effect), to get more statistical power we pooled our novel data and the data from Cowley et al. (2019), as justified by data-gathering replication. We first fit an LMM with non-standardized Flow scores as the DV, and deviation scores, number of cumulative runs, and the interaction between the two as IVs. Participant identity (1–18) was used as a random effect allowing intercepts and slopes to vary for deviation scores, cumulative runs, and their interaction. Additionally, we fit a slightly simpler model allowing only the intercepts to vary, to avoid model non-convergence. These models allow for evaluating whether the effect of deviation scores on Flow scores, i.e. $F{\sim}d$, depends on the level of cumulative runs; or, in other words, whether the $F{\sim}d$ effect grows stronger or weaker as participants gain experience in the game.

For effect size estimates, we used the method by Nakagawa and Schielzeth (2013), which provides marginal (variance explained by fixed factors) and conditional (variance explained by both fixed and random factors) r^2-values for LMMs. For significance estimates, we used the lmerTest package (Kuznetsova, Brockhoff, Christensen et al., 2017), which applies Satterthwaite's method for approximating the degrees of freedom and calculating p-values for LMMs.

We also performed a custom sliding-window analysis for H2. First, for each participant, deviation score was calculated for runs 1 through 10, then for runs 2 through 11, and so on, until runs 31 through 40. This yielded 31 separate deviation scores for each participant for a specific sliding window of 10 runs. The $F{\sim}d$ effect was then calculated for each sliding-window segment, that is, 31 linear models were fit for each

[5] Because the relative contribution of each of the proposed components of Flow to the overall experience of Flow in specific contexts is unknown, the usual custom of assuming they are all equal a priori appears unjustified (Abuhamdeh, 2020; Jackson & Marsh, 1996). Thus, in one set of analyses, we allowed the weighting of each of the items to vary based on their loadings; we did this for both the 8 and 10 item versions of the scale. Because the results did not differ significantly from those in which all components were fixed to be equal, we used the latter approach, as the associated findings are easier to interpret.

Table 2
Descriptive game statistics.

Variable	Novel dataset (N = 9)		Combined dataset (N = 18)	
	Mean (SD)	Range	Mean (SD)	Range
Average run duration (in seconds)	185.8 (20)	160–300	186.3 (21)	160–413
Average number of collisions	17.3 (6.8)	4–40	17.6 (6.6)	4–43
Average self-reported Flow (scale: 1–7)	4.89 (0.82)	2.36–6.75	4.93 (0.81)	2.38–7

participant with self-reported Flow as the DV and deviation scores as the IV. Next, the slope estimates (B-values) of these models were obtained to determine the strength of the $F{\sim}d$ effect across all 31 sliding windows for each participant. This resulted in time series data on the progression of the $F{\sim}d$ effect within participants, which allow for observing the trend of the $F{\sim}d$ effect over time in detail. Finally, we employed the Minimum Width Envelope method (MWE) (Korpela, Puolamäki, & Gionis, 2014) to visualize the group-wise results and provide statistical confidence estimates. MWEs generalize univariate confidence intervals (CIs) to multivariate time series data. MWE bands tend to be wider than CIs because they account for the non-independent nature of time series data, yet they allow a similar visual interpretation of the data because the true average of the distribution traverses inside the lower and upper bounds with probability of $1 - \alpha$ (where α is the desired level of control of Type I error).

Reported p-values (for analyses from same datasets) were corrected for multiple comparisons using the Bonferroni-Holm method.

3. Results

All participants completed the task. Table 2 summarizes average run duration and number of collisions for the novel dataset and both datasets combined. Fig. 3 shows the distributions of participant- and session-wise self-reported Flow scores across trials. See Figs. A.3, A.4, and A.5 in the Appendix for further detailed descriptive statistics on game performance measures and participants' background.

3.1. Hypothesis 1: replication of the results of Cowley et al. (2019)

Our results clearly replicated findings from Cowley et al. (2019), which focused on **(a)** how performance in CogCarSim changes over time (effect of learning on performance), and **(b)** how Flow is associated with performance.

They found that **(a)** learning to play the game is described by a power-law of practice (linear association of run durations as a function of cumulative number of runs in log-log-space). Similarly, our data showed that log-transformed cumulative runs were strongly negatively associated with log-transformed run duration (B = −0.08, $t = -7.1$, $p < .001$; marginal $r^2 = 0.36$, conditional $r^2 = 0.77$; controlling for age, gender, driving experience and gaming experience: B = −0.08, $t = -7.1$, $p < .001$; marginal $r^2 = 0.44$, conditional $r^2 = 0.75$), meaning that runs were faster as experience in the game increased, and the pattern of performance improvement followed a power-law.

Cowley et al. (2019) also found that **(b)** Flow was not associated with experience (i.e. cumulative number of runs), but was strongly associated with 'deviation-from-anticipated' performance. Our results show that deviation score (model-predicted run duration subtracted from observed run duration) was negatively associated with self-reported Flow (B = −5.68, $t = -3.47$, $p = .01$; marginal $r^2 = 0.12$, conditional $r^2 = .46$[6]; controlling for age, gender, driving and gaming experience: B = −5.69, $t = -3.48$, $p = .01$; marginal $r^2 = 0.1$, conditional $r^2 = 0.57$), but self-reported Flow was not associated with experience (cumulative runs; B = −0.09, $t = -0.68$, $p = .51$). See Fig. 4.

3.2. Hypothesis 2: the moderating effect of learning on Flow

Our results, based on pooled data, show strong support for the hypothesis that the $F{\sim}d$ effect is moderated by task experience. The interaction between deviation scores and cumulative runs was statistically highly significant: B = −0.33, $t = -5.18$, $p < .001$ (LMM with random slope for deviation, cumulative run, as well as their interaction; model marginal $r^2 = 0.22$, model conditional $r^2 = 0.58$), and B = −0.32,

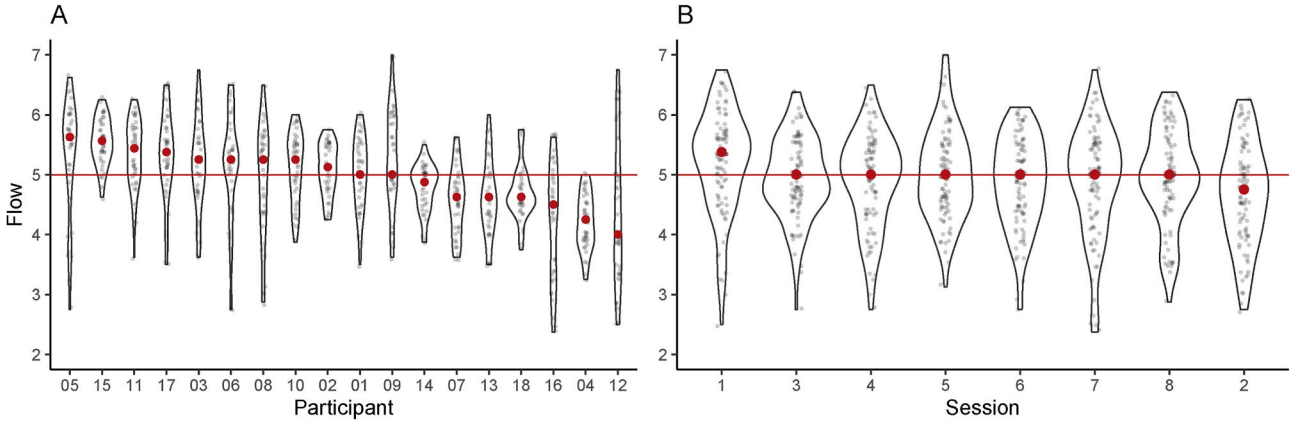

Fig. 3. (A) Participant- and (B) Session-wise violin plots with jittered data points of self-reported Flow scores across trials, organized from highest to lowest median value. The red horizontal lines depict the median value (5) across participants and sessions, while the red dots depict median values within Participant (A) and Session (B). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

[6] This model failed to converge, so we included random intercept only, after which the model converged without issues: B = −4.89, $t = -7.37$, $p < .001$.

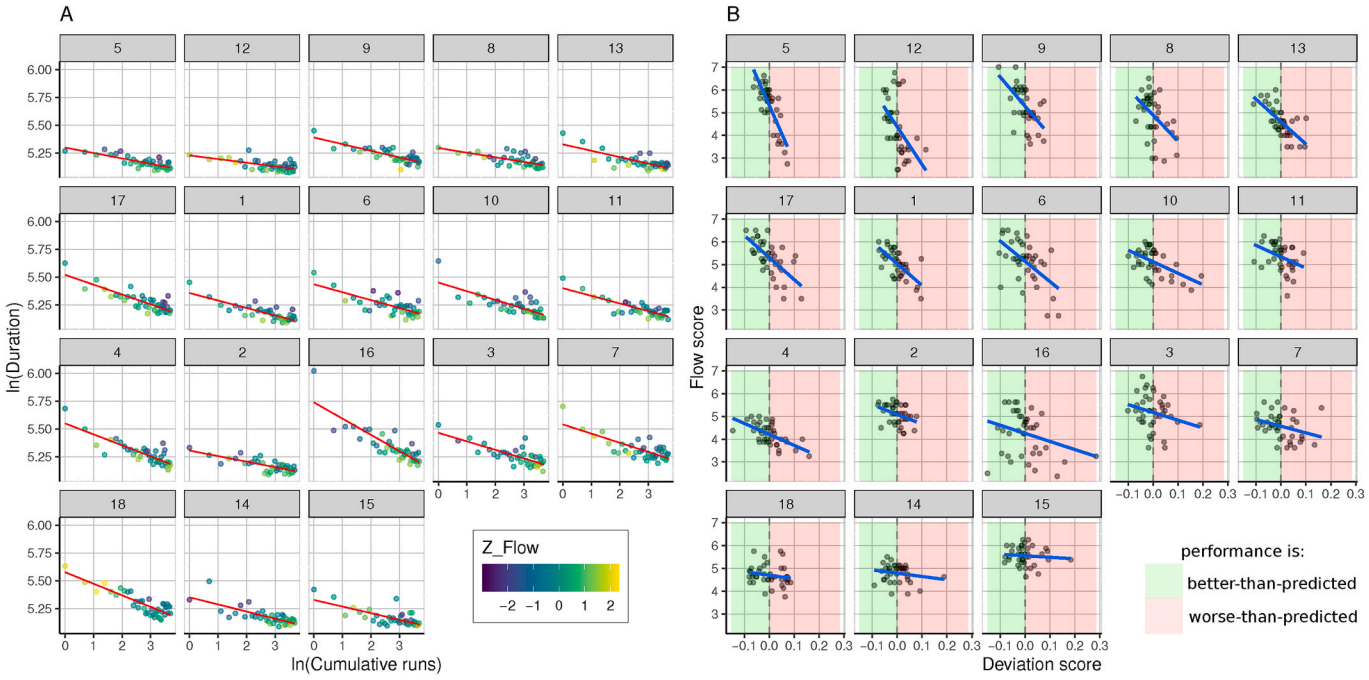

Fig. 4. Participant-wise visualisation of performance and *F~d* effect. *Panel A*: CogCarSim performance across runs (both log-transformed), colored by standardised Flow self-report (FSS) scores. The y- and x-axes show log-duration of runs and log-cumulative run count, respectively. Red lines are power-law learning curves, which transform to linear in log-log space. *Panel B*: Deviation scores (observed run duration minus predicted run duration; negative values indicate better than predicted performance) plotted against Flow scores for each participant, and fitted by linear models (blue lines). The green area indicates better than predicted performance, while the red area indicates worse than predicted performance. *In panel A and B*, subplots are organized in decreasing order of the steepness of the strength of the *F~d* effect (steepness of the slope in panel B). Data for participants 1–9 were originally reported in Cowley et al. (2019); data for participants 10–18 are from the replication experiment. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

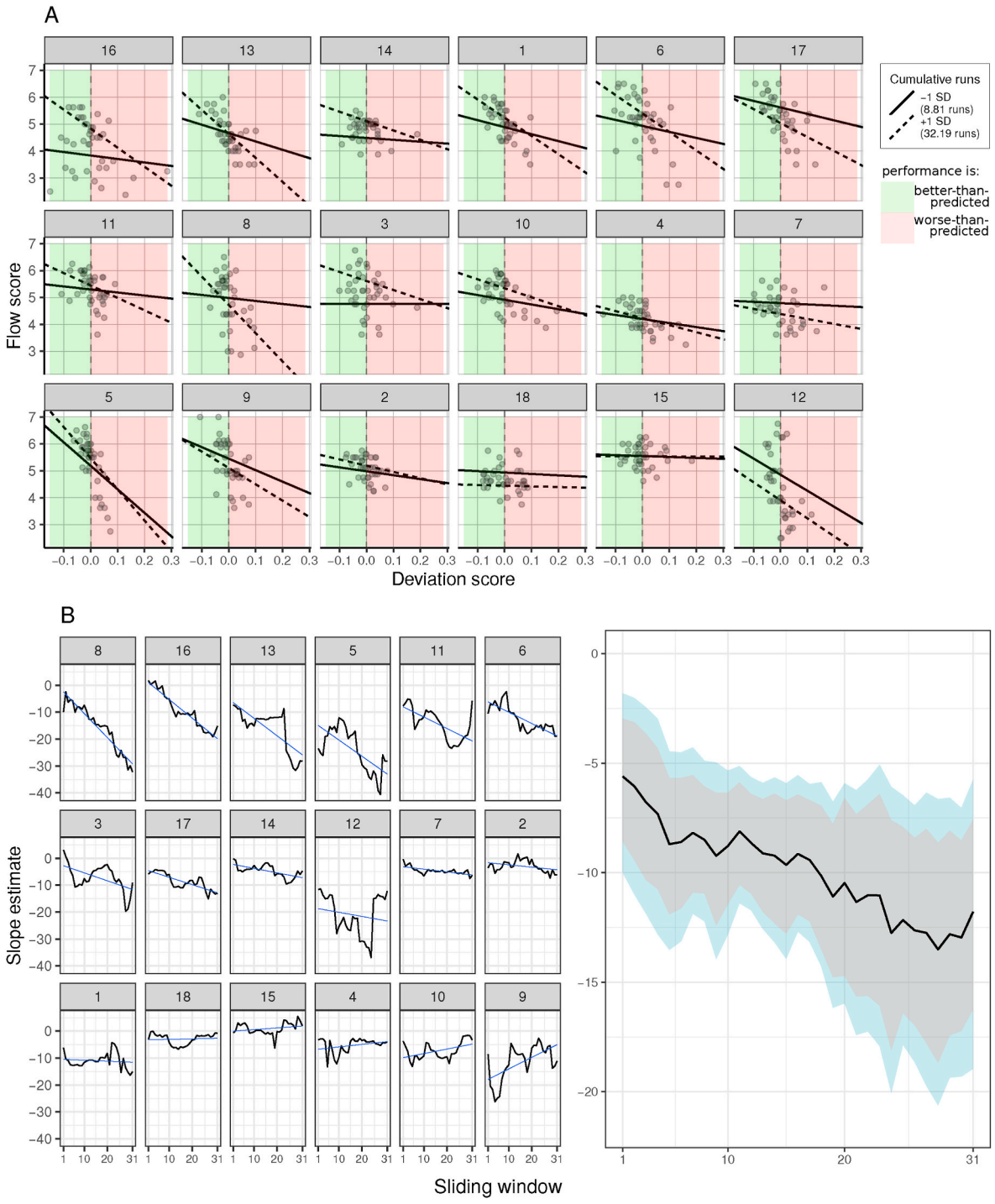

Fig. 5. Analyses of the effect of learning. **Panel A**: Interaction model. Participant-wise linear models depict the effect of deviation score on self-reported Flow at one standard deviation above and below the mean of cumulative runs (moderating variable). Panels are organized from left to right in descending order based on the strength of the interaction. Notably, the slope for deviation score is visibly steeper when cumulative runs is at +1 SD (32.2 runs) than at −1 SD (8.8 runs) for every participant except participants #15, #12 and #18 (for whom the slopes are effectively the same). The green area indicates better than predicted performance, while the red area indicates worse than predicted performance. **Panel B**: Numerical analyses. *Left:* Progressive sliding-window estimation of participant-wise linear models with self-reported Flow as the DV and deviation score as the IV. The model is fit 31 times per participant, using width-10 sliding windows of cumulative run numbers 1) 1–10, 2) 2–11, …, 31) 31–40. These 31 slope estimates (black lines) are then fitted with a linear model (blue lines). Panels are organized from left to right in order of decreasing steepness of the slope of this latter model. Negative trend of the slope estimates implies that as cumulative runs increase (as the sliding window 'shifts right' and participants become more experienced in the game), the *F*~*d* effect becomes stronger (the slope of the model becomes steeper). This is true for all participants except #18, #15, #4, #10 and #9 (5/18 or 27.7% of the participants). *Right:* The group mean of the slope estimates over 31 sliding windows. The wider ribbon shows a 95% confidence band for the overall effect, computed using the Minimum Width Envelope (Korpela et al., 2014) method. The inner ribbon is the naïve 95% confidence *intervals*, i.e. computed per time point without accounting for autocorrelation. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

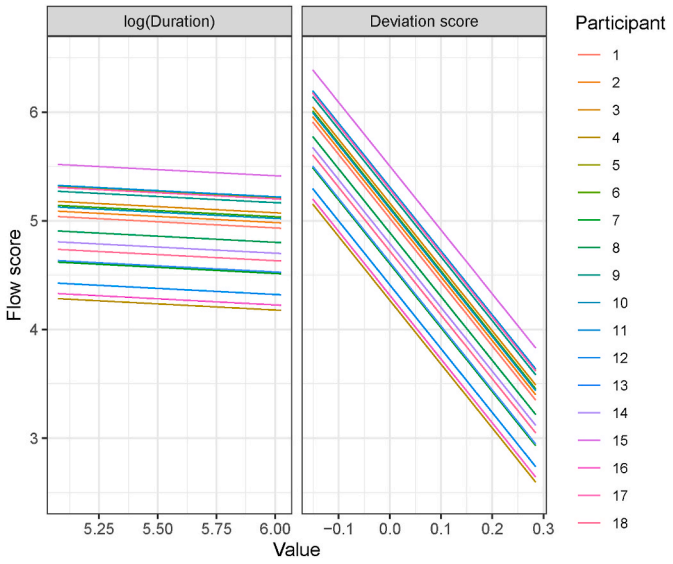

Fig. 6. The effect of log-transformed run duration (left panel; essentially a measure of performance/skill with lower run duration indicating higher skill) and deviation score (right panel; negative deviation scores indicating better than predicted performance) on self-reported Flow, while holding either variable constant at its mean value. The effects are from a linear mixed model where self-reported Flow is the DV, and log-transformed run duration and deviation score are the IVs. A participant identifier (1–18) was used as a random factor, allowing intercepts to vary (but slopes were uniform to allow model convergence). It can be clearly seen that deviation score is a much stronger predictor of self-reported Flow than run duration (i.e. performance), when each is controlled for the other.

$t = -7.84$, $p < .001$ (simpler LMM with random intercept only, and controlling for age, gender, driving experience and gaming experience; model marginal $r^2 = 0.19$, model conditional $r^2 = 0.44$). Simple slopes analysis of this interaction (on the simpler model with random intercepts and control variables) showed that the slopes of deviation scores at the −1 SD (8.94), mean (20.47), and +1 SD (32.01) values of cumulative runs were −3.81 ($t = -6.86$, $p < .001$), −7.54 ($t = -14.62$, $p < .001$), and −11.26 ($t = -13.71$, $p < .001$), respectively, demonstrating the overall decreasing trend. Fig. 5, panel A depicts this interaction on an individual level for each participant, by contrasting the slope of the interacting IV (cumulative runs) at the arbitrary cut-points ±1 SD.

Further, Fig. 5, panel B (right) depicts the steady decrease of the group-mean of slopes from 31 linear models (in 10-run-wide sliding windows) of deviation score predicting self-reported Flow. Bands in this figure show the naive 95% CI, and the wider MWE band which statistically demonstrates that even in the upper limit of 95% confidence, the distribution of model slopes never becomes positive. Fig. 5, panel B (left) illustrates the participant-wise variability in this result. There is (a) clear indication that learning takes place for some who had no change in the interaction model (e.g. #12 whose 'slope estimate' begins at −10, drops to −35 by window 24, then ends back at −10); and (b) strong suggestion that learning follows a linear progression for some, and stepwise for others (e.g. #8, #16 vs. #13, #5).

Overall, these results indicate that the group-wise $F{\sim}d$ effect grows consistently stronger as participants learn to play the game.

3.3. Additional analyses on performance and Flow

We also evaluated whether performance (i.e. log-transformed run duration; shorter run duration = better performance) or deviation score was the better predictor of self-reported Flow. We fit LMMs with log-transformed run duration and/or deviation scores as the IV/predictor variable(s) (together with the control variables of age, gender, driving experience and gaming experience), and self-reported Flow scores as the DV. A participant identifier (1–18) was used as a random effect allowing variability in the intercepts (but not slopes due to model convergence issues) of log-transformed run duration and deviation scores. When log-transformed duration was entered in the model as the sole predictor (together with demographic controls), it was negatively associated with Flow (B = −2.17, $t = -6.87$, $p < .001$), suggesting that self-reported Flow increased with decreasing run duration (i.e. better performance). However, when run duration and deviation score were both entered as predictors, we found that deviation score was linked with Flow (B = −5.94, $t = -9.6$, $p < .001$), but log-transformed run duration was not (B = −0.04, $t = -0.11$, $p = .91$). This suggests that deviation score is a much more significant predictor of Flow than run duration (i.e. performance). See Fig. 6, and Fig. A.9 (in the Appendix) for further details.

3.4. Model assumptions and robustness

All fitted models reported here by and large satisfied the assumptions of linearity, and the residuals were near-normally distributed and homoscedastic. The residuals were also near-normally distributed across the levels of all predictor variables. Further, Q-Q plots indicated that the random effects were near-normally distributed for the models. We nonetheless re-ran our analyses using robust linear mixed modeling (Koller, 2016), but found no significant changes in the pattern of results across all fitted models. All analyses were performed with and without controlling for participants' age, gender, driving experience and gaming experience, but this had no effect on the pattern of the main results. Finally, the analyses were also controlled for a dichotomous variable indexing whether there were physiological measures taken during a session (0 = no physiological measures [sessions 2–4], 1 = physiological measures [sessions 1 and 5–8]), which did not affect the pattern of the main results.

4. Discussion

We sought to replicate and extend the results of Cowley et al. (2019), which utilized a bespoke high-speed steering game, CogCarSim, to reliably elicit Flow. In our replication dataset, learning to play CogCarSim was well-fit by a power-law curve model, but self-reported Flow was not directly associated with experience. Instead, Flow was linked to deviation from performance expectancy ($F{\sim}d$ effect), i.e. better-than-anticipated performance went with increased Flow, and vice-versa. Thus, we successfully replicated the results of Cowley et al. (2019), most notably validating the $F{\sim}d$ effect. Flow was also better predicted by deviation from performance expectancy than by task performance. Moreover, our novel results for the pooled dataset indicate that the $F{\sim}d$ effect grows steadily stronger as participants gain experience in the game, revealing a prominent effect of learning – putatively via increased accuracy in performance anticipation – on self-reported Flow.

Our results also show that learning to play CogCarSim happens reliably, and despite widely different starting times and learning rates (the participant-wise slopes of log-log models in Fig. 4, panel A), the end

points of performance seem to be converging across participants (see Fig. A.3). Specifically, all participants somewhat quickly reached a level of proficiency where the rate of further improvements slowed down progressively.

CogCarSim reliably elicited medium-to-high self-reported Flow scores, despite a wide range across participants in the scores for 'perceived importance' (Fig. A.4, left panel, in the Appendix). Behavioural results further show that the game, by design, tightly links together performance-related variables with the dependent variable (run duration), and also that each performance variable follows its own power-law curve across the extent of cumulative runs (Fig. A.5 in the Appendix). Thus, our data provides clear evidence that all participants experienced Flow and performed well, which supports the generality of our results since they are so consistent.

4.1. Flow, learning, and motivation

Flow theory, supported by several empirical studies, suggests that Flow is directly linked to performance – with either Flow leading to good performance, or good performance leading to Flow (Engeser & Rheinberg, 2008; Jackson et al., 2001; Schüler, 2007; Stavrou et al., 2007; Sumaya & Darling, 2018). However, evidence also shows that the link between Flow and performance might be moderated (Cowley et al., 2019; Garcia et al., 2019; Schüler & Brunner, 2009). Our current results replicate Cowley et al. (2019)'s $F{\sim}d$ effect: deviation from performance *expectancy* is a better predictor of Flow than performance itself. We now *also* show that *task experience* moderates the link between Flow and performance expectancy: the $F{\sim}d$ effect became stronger as participants became more experienced at CogCarSim. These results imply that future research on Flow should account for participants' task experience and learning as a moderating factor on Flow experiences.

Although Flow theory has had a significant impact on motivation theory, it is only one of several contemporary theories of intrinsic motivation (Reeve, 2012). Perhaps even more influential within the field is self-determination theory (Deci, Ryan et al., 1985; Ryan & Deci, 2000), according to which intrinsic motivation is a product of the satisfaction of the "fundamental needs" of perceived competence, perceived autonomy, and relatedness.

Our results have important implications for one of self-determination theory's fundamental tenets: increases in perceived competence in an intrinsically-motivating (i.e. fun) task will result in corresponding increases in intrinsic motivation (Deci et al., 1985). In the current study – if we consider self-reported Flow as a proxy for enjoyment – this was not the case. Although perceived competence increased steadily across the 8 study sessions (see Fig. A.8 in the Appendix), Flow did so only negligibly. We believe this result can be explained by the fact that as one gains experience in an activity, his/her expectations rise accordingly (based on the power-law curve). It is deviations from the predicted performance delineated by this curve, we assert, that will most strongly determine the (dis)satisfaction one derives from the task, rather than one's absolute level of perceived competence at the task. It is interesting to contrast our results with a recent experimental manipulation of self-efficacy (Peifer, Schönfeld, Wolters, Aust, & Margraf, 2020). Peifer et al. (2020) found that while false-normative positive feedback (saying performance was better than average) after one block of a mental arithmetic task did not predict performance and Flow in a second block, there was nonetheless a mediation effect via self-efficacy. In other words, positive feedback increased self-efficacy, which, in turn, had a positive effect on performance and Flow. If this can be interpreted such that individuals with high mental-arithmetic self-efficacy (i.e. "I can do this!"-mindset), will be equipped to perform better than expected and so experience more flow, then this study may be complementary to our $F{\sim}d$ result. On the other hand, Peifer et al. (2020) had only one block to set expectations, and as we have shown, the effect of performance expectancy on Flow strengthens with task learning. Thus the effect of self-efficacy on Flow (purported by Peifer et al. (2020)) might not survive over the long run. Future work should test whether the link between self-efficacy and Flow is explained specifically by exceeding one's own performance expectations while learning a task to proficient level.

4.2. Limitations & implications

Sample size was a limitation in Cowley et al. (2019), which we mitigated here by increasing the sample size to 18 while also getting a more representative gender distribution. While 18 participants may still seem few, in our longitudinal setting it corresponded to 144 measurement sessions and 720 trials across all participants. Moreover, our results were highly consistent across participants, with relatively large statistical effect sizes. This suggests the sample size was quite sufficient to demonstrate the $F{\sim}d$ and learning results.

One issue is that our data can only support the $F{\sim}d$ assertion for early-to-intermediate stages of task mastery, given that our participants each trained about 6 h of 'seat time' and achieved maximum performance levels well below that demonstrated by those of us who developed the game and devoted tens of hours to practice. Certainly, one cannot extrapolate the results to true expert performance, development of which is commonly accepted to require thousands of hours of relevant practice.

Our measures of Flow were based on self-reports (the FSS). Thus, our results and their interpretation pertain only to self-reported Flow-related experiences, and we cannot make conclusions about the actual phenomenological experiences our participants may have had while playing. Psychometric tools such as self-report scales are, however, a gold standard in psychology; and there are currently no validated methods to reliably *objectively* measure the Flow experience. A more practical issue is that FSS *operationalizes* Flow as a continuous phenomenon, whereas Flow is typically *conceptualized* as a discrete phenomenon (i.e. one is either 'in Flow' or not). This discrepancy in operationalizing Flow is also an important topic for future research Abuhamdeh (2020).

After each run, participants saw their current score alongside their previous high scores (up to ten of). Having performed well on the current run compared with previous runs can be seen as a form of positive feedback, which may consequently have affected participants' self-reports of Flow. So, it is possible that self-reports of Flow on a given trial were partly driven by this knowledge of relative performance level. However, we performed two supplementary analyses that speak against this assumption (see Appendix A.1). Firstly, the effect of 'positive feedback' (i.e. having performed better on a given run than the previous one and seeing the score) on Flow was completely mediated, or 'explained away', by the $F{\sim}d$ effect. Secondly, we also found that self-reported Flow on a previous trial was positively associated with Flow on subsequent trials, suggesting a carryover effect of experienced Flow across trials (see Figs. A.1 and A.2 in the Appendix). This effect cannot be accounted for by assuming that Flow is only linked to 'performance feedback'.

Finally, given the observational nature of our data, we cannot draw firm conclusions on directions of causality between our measures – that is, whether higher Flow is caused by better than expected performance, or vice versa. It is ultimately unclear whether this question is answerable

under any circumstances, given the drawbacks of self-report described above. On the other hand, taking our results at face value provides valuable insight for applications in, e.g. design of human-computer interaction (HCI) systems, where Flow is prized (e.g. Huang, 2003; Jin, 2012; Kiili & Lainema, 2008). The $F{\sim}d$ effect (supporting H1b) provides a clear mechanism to understand when and why users of HCI systems might experience Flow in the context of performing a particular skill the system requires of them. Results supporting H2, that show $F{\sim}d$ is refined by task experience, allow further insight into how users will respond over a particular learning curve design. Since applied design of Flow experiences might often stop after considering the foundational tenets of Csíkszentmihályi's original writings, which do not delve deeply into how Flow evolves with skill learning, our results provide a substantial advance in empirical insight.

4.3. Conclusion

We have shown that despite some previous evidence linking Flow and performance (e.g. Engeser & Rheinberg, 2008; Jackson et al., 2001; Schüler, 2007; Stavrou et al., 2007; Sumaya & Darling, 2018), self-reported Flow in our study was linked not to absolute performance *level*, but to *better than expected* performance, and this link was moderated by *task experience*.

Indeed this finding might not contradict the prior work but only clarify it, because without a longitudinal design and/or learning curve modelling, it can be possible that expectation of performance effects were not visible or not prominent. Looking closer at Engeser and Rheinberg (2008), for example, their study of Pac-Man had a similar structure of repeated play-trials followed by FSS response, but with manipulated difficulty level. We can see from their reported results something that was not discussed in their paper: that Flow scores respond (in line with our results) *both* to the current difficulty level and to the relative change in difficulty, which would presumably modulate the expected performance. Thus, it is probable that many empirical studies of Flow contain additional information to help clarify the genesis of self-reported Flow experiences, if re-examined in light of our results. Future work should study this in a systematic review.

Finally, Nakamura and Csikszentmihalyi (2002) foreshadowed our results when they wrote *"As people master challenges in an activity… to continue experiencing Flow, they must identify and engage progressively more complex challenges."* Cowley et al. (2019) conceived of this process as '*introducing complexity*', which can be interpreted in light of our results as meaning that individuals engage with a more complex internalisation of the task as they learn it, and thus demand more of themselves in order to evaluate their own performance as 'exceeding expectations'. More work is needed to examine these possibilities.

Acknowledgment

BC was supported by the Academy of Finland (grant #PROFI4 318913). OL was supported by the Academy of Finland (grants #325694 and #334192, the latter of which also supported JP). ML was supported by Jane and Aatos Erkko Foundation, and the Academy of Finland (grant #323207). We also thank Adriano Pomarè, Anton Berg, and Eetu Nisula for their invaluable help in collecting the data.

A. Appendix

A.1. Knowledge of success in previous runs

In Discussion we raised the issue that in the current study, after each run, participants saw their current score as well as their previous top ten high scores. To obtain an index of 'positive feedback' based on behavioral data, we first subtracted the current run duration from the previous run duration. Positive values indicate that the current run was faster than the previous run, and vice versa for negative values.

We found that this measure of positive feedback was linked with self-reported Flow: if participants performed better on run N than run N-1, they reported significantly more Flow. However, this effect was completely mediated by deviation scores (see Fig. A.1 for statistics). To diagnose multilevel mediation we used the method by Vuorre and Bolger (2018). We first averaged the grand-mean-centered run-level feedback, deviation and Flow scores. These mean values were then subtracted from the corresponding variables' raw values to create within-subject run-by-run deviations from the subject-means. The resulting values are a "within-person version" of feedback, deviation, and Flow, from which between-subjects variability has been removed.

Moreover, Fig. A.2 demonstrates a carryover effect of experienced Flow across runs: Flow on the current run (dependent variable) is predicted by lagged Flow scores, that is, Flow scores on N runs before the current run. This effect is difficult to explain by assuming that Flow is only linked to having seen the scores of one's previous runs.

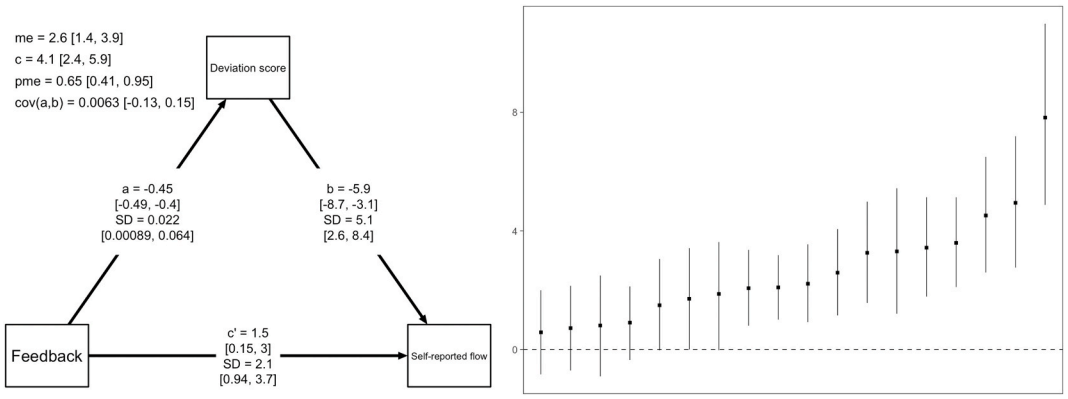

Fig. A.1. Multilevel mediation model; *Panel A*: Deviation mediates the effect of feedback on self-reported Flow. me = mediated effect; c = direct effect; c' = direct effect controlling for Deviation; pme = percentage mediated effect; SD = standard deviation. *Panel B*: The size of the mediation effect for each participant with 90% credible intervals.

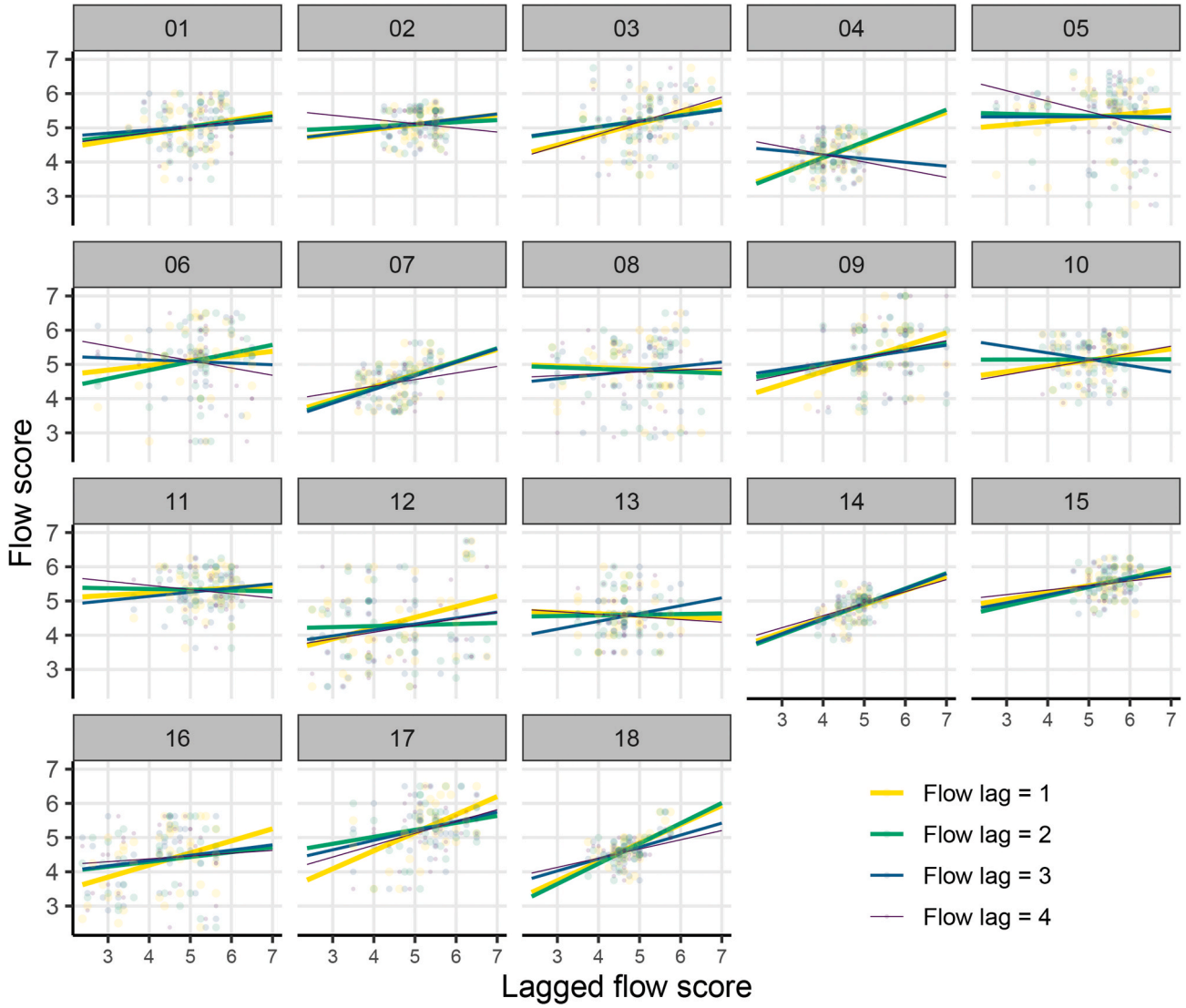

Fig. A.2. Effect of lagged Flow on Flow. Self-reported Flow on a previous run is positively associated with Flow on subsequent runs, suggesting a carryover effect of experienced Flow across runs. Flow lag = 1–4 refers to a Flow score 1–4 runs prior to the current run.

A.2. Demographics and behavioural game variables

This section presents additional results on demographics and behavioural (and self-reported) game variables.

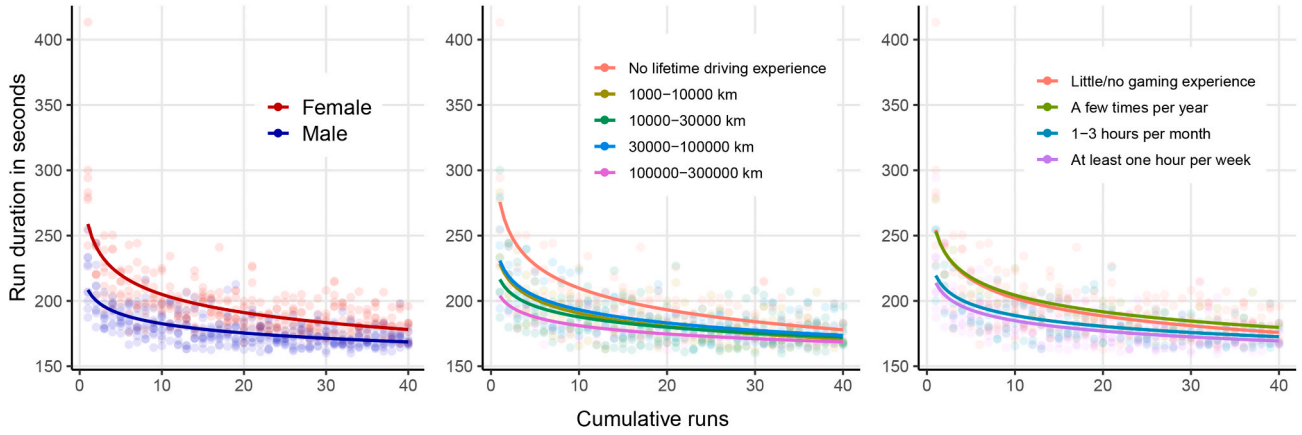

Fig. A.3. Effects on run duration of Gender (left), Driving experience (middle), and Gaming experience (right).

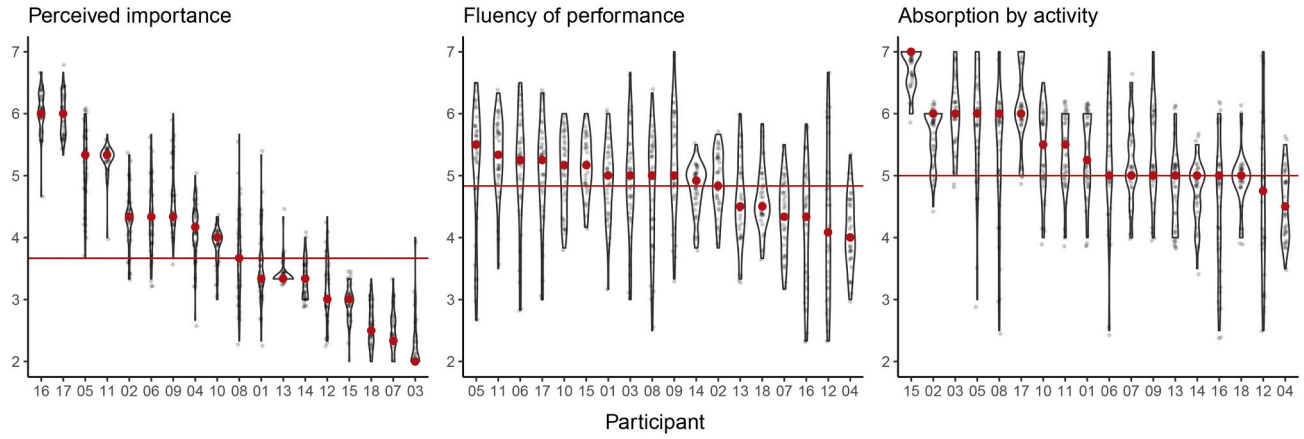

Fig. A.4. Participant-wise violin plots with jittered data points of self-reported perceived importance, fluency of performance, and absorption by activity, organized from highest to lowest median value. The red horizontal lines depict the median values, while the red dots depict median values within Participant.

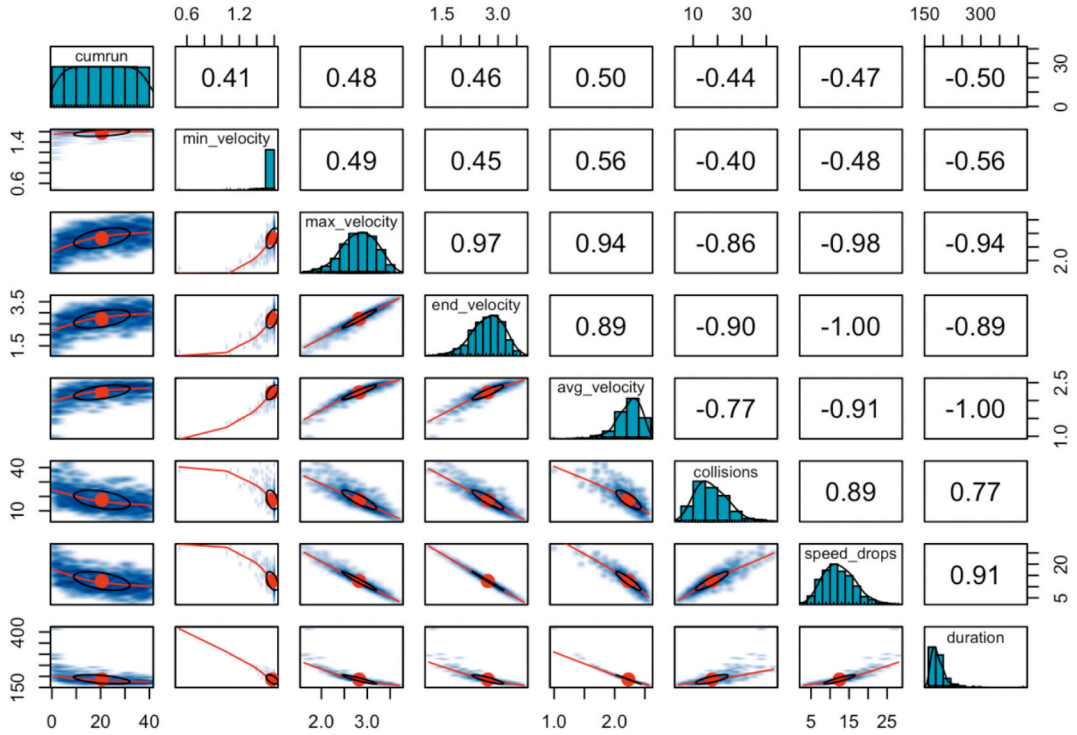

Fig. A.5. Scatterplot matrix of behavioural game variables, showing that the game (by design) tightly links together the performance-related variables (*min_velocity* to *speed_drops*) with the outcome variable *duration*, and also that all performance variables follow a power-law curve across the extent of cumulative runs (*cumrun*).

A.3. FSS reliability

As detailed in Methods, Cronbach's alphas were calculated separately for each run across participants. Histograms of these alpha values are depicted in Fig. A.6.

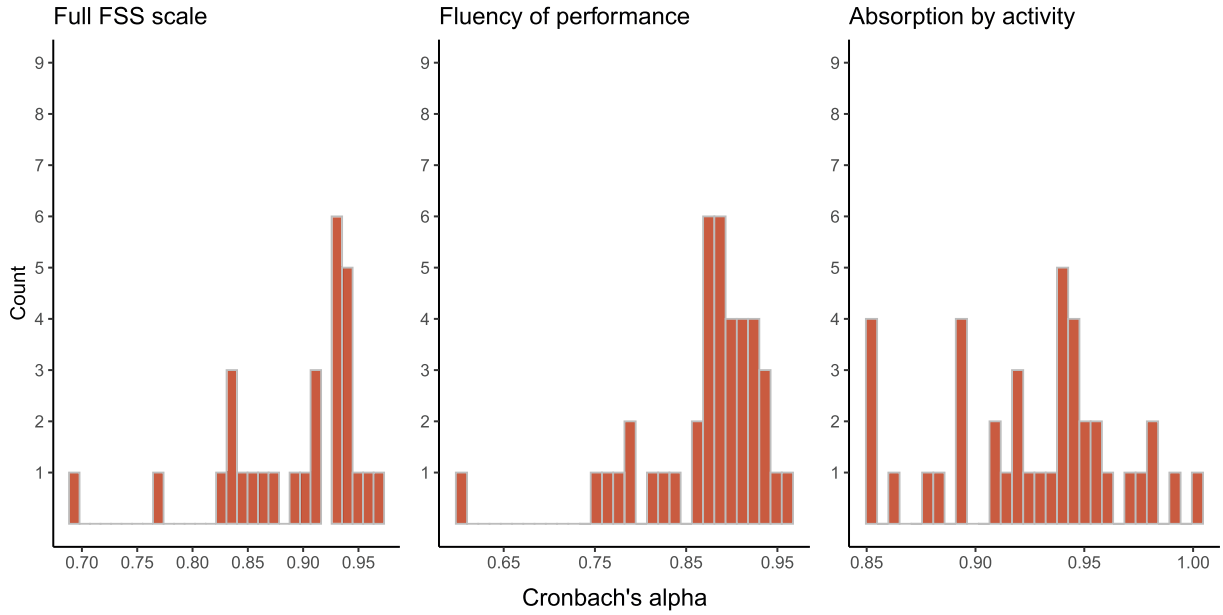

Fig. A.6. Histograms of trial-wise Cronbach's alpha values for the full FSS scale, and the subfactors fluency of performance and absorption by activity.

A.4. Participants' measurement dates

As detailed in Design, participants were measured on 8 different days. Fig. A.7 depicts all measurement session dates across participants.

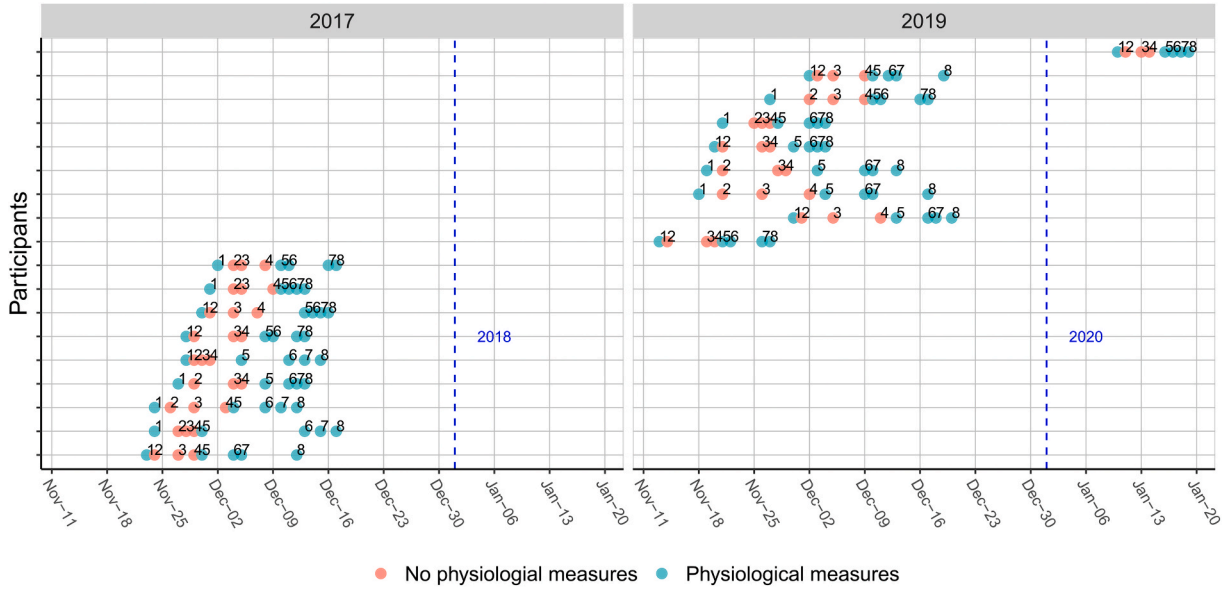

Fig. A.7. Dates of participants' measurement sessions (1–8), separately for data collected in 2017 (originally reported in Cowley et al. (2019)) and 2019 (novel data). Physiology was measured during sessions 1 and 5–8, including pupillometry, electrodermal activity, blood volume pulse, and electrooculography. Participant IDs are blinded from the y-axis for anonymity.

A.5. Self-reported perceived competence across sessions

Fig. A.8 supplements the Discussion on the implications of our results for self-determination theory.

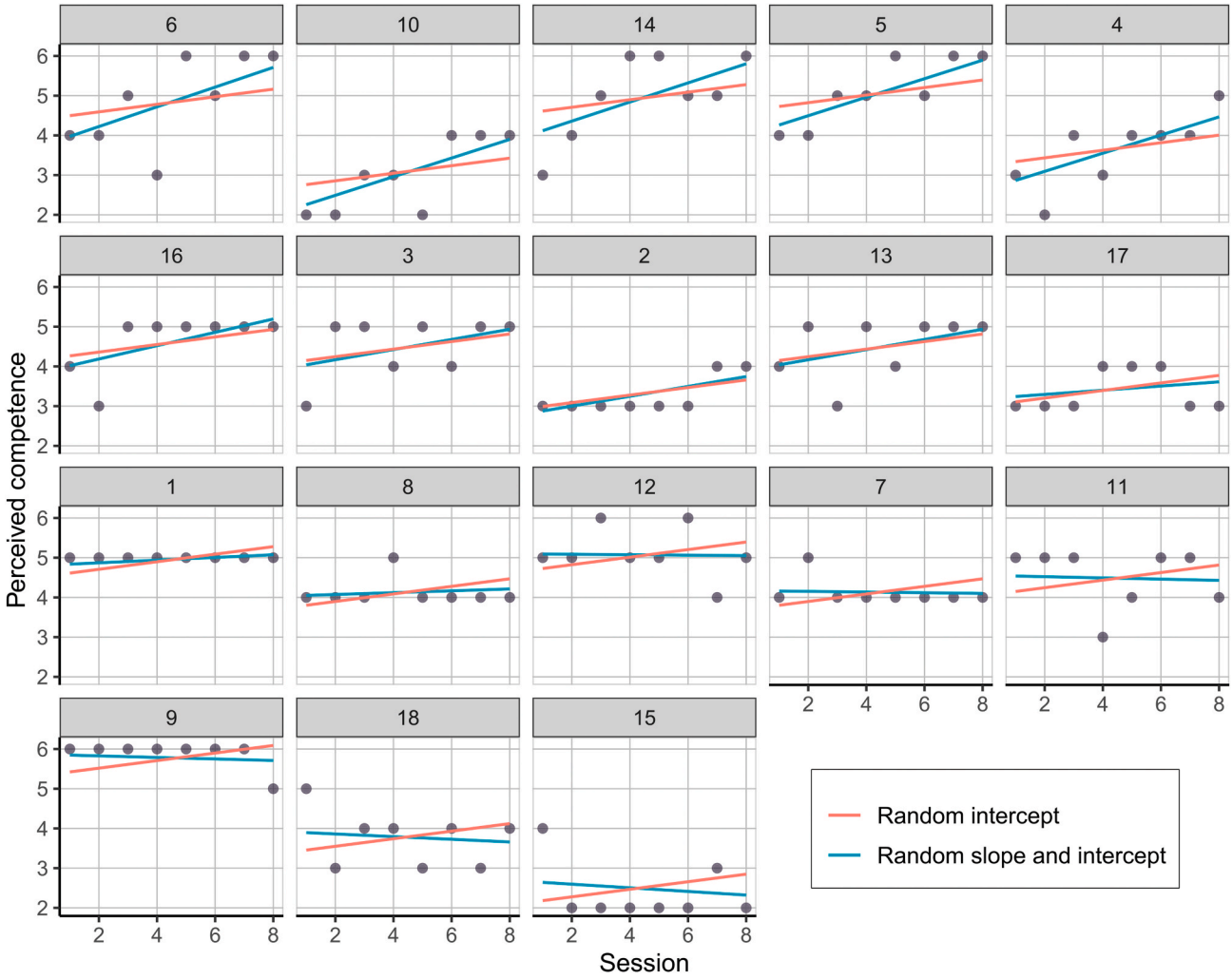

Fig. A.8. Participant-wise data showing self-reported perceived competence ratings ("My level of competence is …", with responses ranging from low (1) to high (6)) across sessions. Perceived competence was measured only once at the end of each session. Model fits are from two separate LMM models: Perceived competence was used as the dependent variable, a "session" variable (ranging from 1 to 8) as the independent variable, and a participant identifier (ranging from 1 to 18) as a random effect, allowing for variability i) only in the intercepts (red slope), or ii) both intercepts and slopes of session (green slop). In both models, the effect of session on perceived competence was statistically significant: B = 0.09, $t = 3.82$, $p < .001$ (random intercept only); and B = 0.09, $t = 2.48$, $p = .024$ (random intercept and slope), with self-reported perceived competence generally increasing as participants gained experience in the game.

A.6. Run duration and deviation scores as predictors of Flow

In Results, we highlighted the difference between run duration (i.e. performance/skill) and deviation scores when predicting self-reported flow. These results are supplemented here by Figs. A.9 and 6.

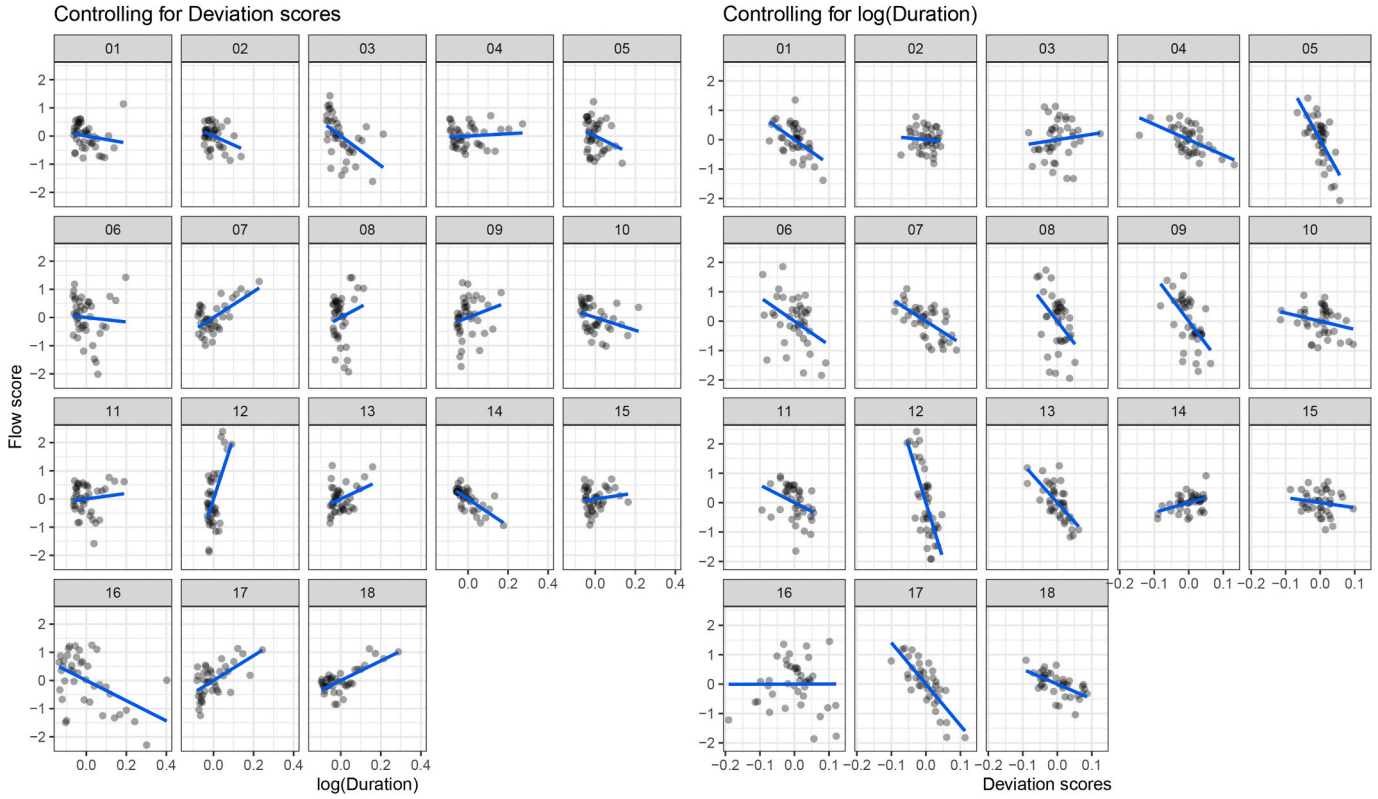

Fig. A.9. Participant-wise partial regression plots depicting the effect of log-transformed run duration (left panel) and deviation scores (right panel) on self-reported Flow scores. The data points are residuals from participant-wise linear models that partial out the effect of either deviation scores (left panel) or log-transformed run duration (right panel). The slopes for deviation scores are negative (except for participants 3 and 14), while the slopes for log-transformed run duration are essentially random (8 negative slopes and 10 positive slopes). This demonstrates that compared with run duration (i.e. performance in the task), deviation score is a much stronger predictor of self-reported Flow.

A.7. FSS item translations

The FSS items as used in the current study were first translated from English to Finnish, but modified slightly to reflect the CogCarSim game. Thus, to clarify for readers the exact meaning of the Finnish phrases answered by participants, Table A.1 shows the items in their original form (right column), the Finnish version (left column) translated from the original and used in the study (all participants were native Finnish speakers), and the English version translated from the Finnish (middle column).

Note that in the original item 2, the words "fluidly" and "smoothly" are almost synonymous, and in a gaming context, they are aptly captured by the single word "sujuvasti" (which could also mean "fluently").

Table A.1
Translated and back-translated FSS items (modified for the current study).

Item	Finnish Translation	Back-translation	Original English
1	Peli tuntui juuri sopivan haastavalta	Playing the game, I felt just the right amount of challenge	I feel just the right amount of challenge
2	Pelasin sujuvasti	I played fluently	My thoughts/activities run fluidly and smoothly
3	En huomannut ajankulkua	I did not notice time passing	I do not notice time passing
4	Pystyin hyvin keskittymään	I found it easy to concentrate	I have no difficulty concentrating
5	Mieleni oli selkeä	My mind was clear	My mind is completely clear
6	Uppouduin täysin pelaamiseen	I immersed (myself) fully in playing	I am totally absorbed in what I am doing
7	Löysin oikeat liikkeet kuin itsestään	I found the right moves spontaneously	The right thoughts/movements occur of their own accord
8	Olin koko ajan tilanteen tasalla	I was able to cope with the task all the time	I know what I have to do each step of the way
9	Tunsin hallitsevani tilannetta	I felt in control of the situation/I felt I had everything in control	I feel that I have everything under control
10	Syvennyin peliin täysin	I delved into the game fully	I am completely lost in thought
11	Koin pelissä onnistumisen tärkeäksi	It was important to me to succeed in the game	Something important to me is at stake here
12	Minusta tuntui siltä, etten saisi tehdä yhtäkään virhettä	I felt like I shouldn't make any mistakes	I must not make any mistakes here
13	Pelkäsin epäonnistuvani	I was worried about failing	I am worried about failing

Author contributions

Jussi Palomäki: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing. Tuisku Tammi: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing. Noora Lehtonen: Investigation, Writing. Niina Peltonen: Investigation, Writing. Michael Laakasuo: Formal analysis, Methodology, Writing. Sami Abuhamdeh: Formal analysis, Methodology, Writing. Otto Lappi: Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Writing. Benjamin Ultan Cowley: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Visualization, Writing.

References

Abuhamdeh, S. (2020). Investigating the "flow" experience: Key conceptual and operational issues. *Frontiers in Psychology, 11*, 10.3389/fpsyg.2020.00158 https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00158/full.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. arXiv preprint arXiv:1406.5823.

Chen, L. X., & Sun, C. T. (2016). Self-regulation influence on game play flow state. *Computers in Human Behavior, 54*, 341–350.

Cowley, B. U., Dehais, F., Fairclough, S., Karran, A. J., Palomäki, J., & Lappi, O. (2020). Editorial: High performance cognition: Information-processing in complex skills, expert performance, and flow. *Frontiers in Psychology, 11*, 10.3389/fpsyg.2020.579950 https://www.frontiersin.org/articles/10.3389/fpsyg.2020.579950/full.

Cowley, B. U., Palomäki, J., Tammi, T., Frantsi, R., Inkilä, V. P., Lehtonen, N., et al. (2019). Flow experiences during visuomotor skill acquisition reflect deviation from a power-law learning curve, but not overall level of skill. *Frontiers in Psychology, 10*, 1126.

Csikszentmihalyi, M. (1975). Play and intrinsic rewards. *Journal of Humanistic Psychology, 15*, 41–63.

Deci, E. L., Ryan, R. M., et al. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality, 19*, 109–134.

Delrue, J., Mouratidis, A., Haerens, L., De Muynck, G. J., Aelterman, N., & Vansteenkiste, M. (2016). Intrapersonal achievement goals and underlying reasons among long distance runners: Their relation with race experience, self-talk, and running time. *Psychologica Belgica, 56*, 288.

Emerson, H. (1998). Flow and occupation: A review of the literature. *Canadian Journal of Occupational Therapy, 65*, 37–44. https://doi.org/10.1177/000841749806500105, 10.1177/000841749806500105.

Engeser, S., & Rheinberg, F. (2008). Flow, performance and moderators of challenge-skill balance. *Motivation and Emotion, 32*, 158–172.

Garcia, W. F., Codonhato, R., Mizoguchi, M. V., Nascimento Junior, J. R. A., Vissoci, J. R. N., Aizava, P. V. S., et al. (2019). Dispositional flow and performance in brazilian triathletes. *Frontiers in Psychology, 10*, 2136.

Harris, D., Allen, K., Vine, S., & Wilson, M. (2020). *A systematic review and meta-analysis of the relationship between flow states and performance*. https://doi.org/10.31234/osf.io/qg852

Huang, M. H. (2003). Designing website attributes to induce experiential encounters. *Computers in Human Behavior, 19*, 425–442.

Jackson, S. A., & Marsh, H. W. (1996). Development and validation of a scale to measure optimal experience: The flow state scale. *Journal of Sport and Exercise Psychology, 18*, 17–35.

Jackson, S. A., Thomas, P. R., Marsh, H. W., & Smethurst, C. J. (2001). Relationships between flow, self-concept, psychological skills, and performance. *Journal of Applied Sport Psychology, 13*, 129–153.

Jin, S. A. (2012). "toward integrative models of flow": Effects of performance, skill, challenge, playfulness, and presence on flow in video games. *Journal of Broadcasting Electronic Media - J BROADCAST ELECTRON MEDIA, 56*, 169–186. https://doi.org/10.1080/08838151.2012.678516

Keller, J., & Bless, H. (2008). Flow and regulatory compatibility: An experimental approach to the flow model of intrinsic motivation. *Personality and Social Psychology Bulletin, 34*, 196–209.

Keller, J., & Blomann, F. (2008). Locus of control and the flow experience: An experimental analysis. *European Journal of Personality: Published for the European Association of Personality Psychology, 22*, 589–607.

Kiili, K., & Lainema, T. (2008). Foundation for measuring engagement in educational games. *Journal of Interactive Learning Research, 19*, 469–488.

Koller, M. (2016). robustlmm: an r package for robust estimation of linear mixed-effects models. *Journal of Statistical Software, 75*, 1–24.

Korpela, J., Puolamäki, K., & Gionis, A. (2014). Confidence bands for time series data. *Data Mining and Knowledge Discovery, 28*, 1530–1553.

Kosunen, I., Palomäki, J., Laakasuo, M., Kuikkaniemi, K., Ravaja, N., & Jacucci, G. (2018). Heart-rate sonification biofeedback for poker. *International Journal of Human-Computer Studies, 120*, 14–21.

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H., et al. (2017). Lmertest package: tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1–26.

Linden, D.v.d., Tops, M., & Bakker, A. B. (2020). Go with the flow: A neuroscientific view on being fully engaged. *European Journal of Neuroscience*. https://doi.org/10.1111/ejn.15014. n/a https://onlinelibrary.wiley.com/doi/abs/10.1111/ejn.15014.

Moneta, G. B. (2012). On the measurement and conceptualization of flow. In *Advances in flow research* (pp. 23–50). Springer.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*, 133–142.

Nakamura, J., & Csikszentmihalyi, M. (2002). The concept of flow. In C. R. Snyder, & S. J. Lopez (Eds.), *Handbook of positive psychology* (pp. 89–105). New York (NY): Oxford University Press.

Newell, A., & Rosenbloom, P. S. (1982). *Mechanisms of skill acquisition and the law of practice*. Design Research Center. Pittsburgh, Pa: Carnegie-Mellon University.

Peifer, C., Schönfeld, P., Wolters, G., Aust, F., & Margraf, J. (2020). Well done! Effects of positive feedback on perceived self-efficacy, flow and performance in a mental arithmetic task. *Frontiers in Psychology, 11*, 10.3389/fpsyg.2020.01008 https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01008/full.

Ravaja, N., Saari, T., Salminen, M., Laarni, J., & Kallinen, K. (2006). Phasic emotional reactions to video game events: A psychophysiological investigation. *Media Psychology, 8*, 343–367.

RCoreTeam. (2013). *R: A language and environment for statistical computing*.

Reeve, J. (2012). A self-determination theory perspective on student engagement. In *Handbook of research on student engagement* (pp. 149–172). Springer.

Rheinberg, F., Vollmeyer, R., & Engeser, S. (2007). *Die erfassung des flow-erlebens. Diagnostik von Motivation und Selbstkonzept/hrsg. von Joachim Stiensmeier-Pelster und Falko Rheinberg. - Göttingen [u.a.] : Hogrefe, 2003. - (Tests und Trends ; N.F., 2*. ISBN 3-8017-1674-0. - S. 261 - 279.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*, 68.

Schüler, J. (2007). Arousal of flow experience in a learning setting and its effects on exam performance and affect. *Zeitschrift für Padagogische Psychologie, 21*, 217–227.

Schüler, J., & Brunner, S. (2009). The rewarding effect of flow experience on performance in a marathon race. *Psychology of Sport and Exercise, 10*, 168–174.

Stavrou, N. A., Jackson, S. A., Zervas, Y., & Karteroliotis, K. (2007). Flow experience and athletes' performance with reference to the orthogonal model of flow. *The Sport Psychologist, 21*, 438–457.

Sumaya, I. C., & Darling, E. (2018). Procrastination, flow, and academic performance in real time using the experience sampling method. *The Journal of Genetic Psychology, 179*, 123–131.

Swann, C., Keegan, R. J., Piggott, D., & Crust, L. (2012). A systematic review of the experience, occurrence, and controllability of flow states in elite sport. *Psychology of Sport and Exercise, 13*, 807–819, 10.1016/j.psychsport.2012.05.006 http://www.sciencedirect.com/science/article/pii/S1469029212000660.

Swann, C., Piggott, D., Schweickle, M., & Vella, S. A. (2018). A review of scientific progress in flow in sport and exercise: Normal science, crisis, and a progressive shift. *Journal of Applied Sport Psychology, 30*, 249–271.

Vuorre, M., & Bolger, N. (2018). Within-subject mediation analysis for experimental data in cognitive psychology and neuroscience. *Behavior Research Methods, 50*, 2125–2143.