The Complementary Nature of Perceived and Actual Time Spent Online in Measuring Digital Well-being

LILLIO MOK, University of Toronto
ASHTON ANDERSON, University of Toronto

As online platforms become ubiquitous, there is growing concern that their use can potentially lead to negative outcomes in users' personal lives, such as disrupted sleep and impacted social relationships. A central question in the literature studying these problematic effects is whether they are associated with the amount of time users spend on online platforms. This is often addressed by either analyzing self-reported measures of time spent online, which are generally inaccurate, or using objective metrics derived from server logs or tracking software. Nonetheless, how the two types of time measures comparatively relate to problematic effects—whether they complement or are redundant with each other in predicting problematicity—remains unknown. Additionally, transparent research into this question is hindered by the literature's focus on closed platforms with inaccessible data, as well as selective analytical decisions that may lead to reproducibility issues.

In this work, we investigate how both self-reported and data-derived metrics of time spent relate to potentially problematic effects arising from the use of an open, non-profit online chess platform. These effects include disruptions to sleep, relationships, school and work performance, and self-control. To this end, we distributed a gamified survey to players and linked their responses with publicly-available game logs. We find problematic effects to be associated with both self-reported and data-derived usage measures to similar degrees. However, analytical models incorporating both self-reported and actual time explain problematic effects significantly more effectively than models with either type of measure alone. Furthermore, these results persist across thousands of possible analytical decisions when using a robust and transparent statistical framework. This suggests that the two methods of measuring time spent measure contain distinct, complementary information about problematic usage outcomes and should be used in conjunction with each other.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in collaborative and social computing**; • **Applied computing** → *Law, social and behavioral sciences*.

Additional Key Words and Phrases: problematic use of online platforms, online well-being, survey methodology, specification curve analysis

ACM Reference Format:
Lillio Mok and Ashton Anderson. 2020. The Complementary Nature of Perceived and Actual Time Spent Online in Measuring Digital Well-being. *J. ACM* 37, 4, Article 111 (August 2020), 27 pages. https://doi.org/10.1145/1122445.1122456

Authors' addresses: Lillio Mok, lillio@cs.toronto.edu, University of Toronto; Ashton Anderson, ashton@cs.toronto.edu, University of Toronto.

0004-5411/2020/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Online platforms support an increasingly broad range of human activities, and are often associated with social [10], economic [29], and psychological [19] benefits. However, there is growing apprehension that people can use these services in ways that might negatively affect their personal lives [28, 31]. A rich body of literature has thus developed around understanding the undesirable effects of online platforms on users' well-being. Many treat unhealthy platform use as pathological, akin to a "technology addiction" [28, 31, 38, 54, 61], while others consider the phenomenon to be more related to impulse control in nature [60]. Some ascribe minimal clinical significance and investigate the more informal notion of "problematic use" instead [12, 14]. Nonetheless, the primary concern across this work is that online activities can become unhealthy when they give rise to effects like sleep interference [11], decreased work or school performance [28], or harm to close relationships [39].

A central question in the literature is the relationship between the quantity and quality of time spent online. In other words, are negative psychosocial effects associated with elevated usage of online platforms? Previous research addresses this with one of two approaches: they either relate well-being to self-reported usage, or relate well-being with usage data derived from server logs instead. By virtue of their survey-based methodology, the majority of existing studies take the first approach of correlating problematic usage effects with self-reported measures of total time spent [38, 51, 57, 63]. Indeed, many correlate psychometric inventories against self-reported aggregate usage to assess criterion validity [28]. However, comparisons of self-reports with usage logs have shown them to be typically inaccurate [9, 24, 25, 50]. Participants have difficulty recalling even the simplest measures of usage, e.g. time spent on the preceding day, thus casting doubt over correlations between problematic effects and self-reported usage in existing work. In response to this, an increasingly popular strategy has been to use objective measures of time spent derived from data traces instead of self-reports [14, 37, 52]. Identifying how negative effects are related to actual usage can potentially provide an understanding of unhealthy technology use without recall limitations, and also allow for predicting problematic episodes from behavior traces.

However, these two bodies of work have remained separate with no comparisons of how self-reports and data are associated with problematic effects. Whether self-reported and actual measures of usage are equivalent, complementary, or mutually redundant in predicting these effects thus remains empirically unknown. There are important functional reasons for elucidating these relationships. For one, if self-reports and data traces together have a stronger relationship with problematic effects than either alone, they could potentially be combined to more effectively carry out digital well-being interventions (cf. [40]). Secondly, if self-reported and actual time spent are related to problematic effects in different ways, then correlating self-reported time with psychometric problematic use scales may not be sufficient for assessing criterion validity (see [28]). It is therefore important to supplement our current understanding of how technology use and offline well-being are associated by investigating self-reports and actual measures in conjunction.

We thus seek to fill the gaps in the existing literature through the following research question:

> **RQ**: *How are self-reported and actual measures of time spent on online platforms comparatively associated with problematic effects of platform use?*

Following existing work, we operationalize these problematic effects as perceived sleep disruption, impacted relationships, and affected academic and workplace performance [14, 27, 28, 34, 38, 42, 54, 61]. We also assess negative affective outcomes like loss of control and regret that are linked with unhealthy technology use [16, 27, 28]. Even at moderate levels that do not clinically qualify as pathological or as an addiction, understanding the way in which these effects are comparatively

connected to perceived and actual platform use will elucidate how spending time online affects users' lives.

Parallel to this unanswered empirical question, concerns have also been raised about a lack of transparency in the data and analyses underpinning existing work [45]. In this paper, we consider two of these issues in conjunction with our research question. Firstly, the prior literature focuses almost exclusively on closed, for-profit online platforms [6, 12, 14]. Research on closed platforms is difficult to reproduce and hold to rigorous scientific standards. Access to the required data and materials is restricted only to affiliated researchers, closing off data-driven approaches to understanding problematic usage effects from the broader research community. Furthermore, for-profit platforms may be designed to incentivize engagement, thus potentially confounding "organic" problematic behavior with the effects of platform design.

Secondly, quantitative analyses that yield large, negative correlations between technology use and well-being have come under scrutiny for potentially making selective analytical decisions. These include construct operationalization using survey response variables, diagnostic boundaries for negative phenomena, covariates involved in regression models, and criteria for excluding responses [45]. Selective decisions could lead to results that are specific to only one set of analytical assumptions and are difficult to reproduce and extend [53].

We address these barriers to transparent and open science with two decisions.

(1) *Platform choice.* We study Lichess.org, an open-source online chess platform that has made their complete game logs publicly available[1]. Its openness allows the broader research community to inspect and extend our results, and also to address other digital well-being questions using the data provided by the platform. This diverges significantly from the majority of existing studies targeting closed platforms, whose data and participant pools are inaccessible for the unaffiliated [12, 14]. Furthermore, Lichess has no advertisements and does not operate for a profit. This enables the study of problematic usage effects in the absence of design patterns that could opaquely affect engagement[2].

(2) *Analytical procedure.* To measure the relationship between using online platform use and negative psychosocial effects, we use Specification Curve Analysis (SCA) [45, 53]. This method combines the results from many reasonable analytical models simultaneously, so that they cannot be selectively chosen for larger and more significant effect sizes, thereby reducing the researcher degrees of freedom. SCA also minimizes biases that would otherwise affect decisions such as including or excluding covariates in a regression. Compared to, for example, a single multivariate regression that has one set of analytical assumptions through its included covariates, SCA analyzes all possible combinations of covariates.

Overview of Results. We administered a gamified survey to 131 Lichess users to elicit several self-reported measures of their time spent on Lichess, as well as the extent to which they perceived negative effects of using Lichess. These effects included physical outcomes like disrupted sleep, relationships, and work, as well as affective outcomes like loss of control and regret. Participants' responses were linked to their complete game logs in the previous month. We then compared the associations between both measures of time spent on the platform and problematic effects, and robustly analyzed the strongest associations within the SCA framework.

Quantitatively, we find that self-reported and actual measures of time spent on Lichess are correlated to problematic effects with similar magnitudes, and are also all statistically significant.

[1]All data is available at http://database.lichess.org.

[2]This is a fundamental value of the platform. Their press kit explains, "With no investors demanding profits, and a commitment to never show advertisements or charge for features, Lichess staff can focus on improving the site as their only goal." [1]

We additionally use SCA to examine the persistence of the strongest correlations under many analytical decisions. We find that the strongest self-reported time measure, maximum time spent in a day, is significantly predictive of problematic effects across all 8,196 possible combinations of included covariates. In comparison, the strongest data-derived measure, total time spent in the month, is only significantly predictive of problematic effects in a fraction of analytical specifications.

We further find that self-reports and trace data actually have complementary associations with problematic usage effects. Models combining both self-reported and actual measures of time are more predictive of these effects than either measure in isolation. For example, using the most predictive self-reported variable and the most predictive actual variable together yields an R^2 of 0.232, whereas using either alone yields only 0.136 and 0.118, respectively. This is also consistent across many analytical specifications under SCA. Our findings suggest that self-reports and data contain distinct information about problematic effects, and can be used together to more comprehensively study how and when online platform usage becomes unhealthy.

2 BACKGROUND

2.1 Problematic Effects of Using Online Technologies

The study of problematic effects arising from technology use has a long history. Prior to the mass adoption of Internet-connected platforms, researchers evaluated the potential for problematic use of offline games [21], televisions [41], mobile phones [8], and computers [15]. Since the development of online technologies, this work has been extended to the problematic use of online analogues, such as online multiplayer games [28, 32], social networking sites [31], smartphones [27, 34], and the Internet itself [13, 22, 62]. As a result, academic and industrial researchers alike have developed a wide range of computer-aided solutions to improve self-control and reduce problematic effects from online activities [40].

This diverse body of existing work has several commonalities. Firstly, a significant portion of empirical studies focus on **detecting negative constructs** such as "pathological use" or some type of technological "addiction" (see, for instance, [6, 28, 32, 35, 37, 38]). The majority of these draw their theoretical foundations from clinical psychology, such as the Diagnostic and Statistical Manual of Mental Disorders [4] and related research on its definition of pathological gambling (cf. 26). For example, one of the seminal empirical examinations of problematic Internet use applies a modified questionnaire for detecting pathological gambling [61, 62]. Although they also measure a single construct, other studies ascribe less clinical significance and instead attribute the issue to impulse control disorders [60], deficient self-regulation [36, 51], or, more informally, to simply be "problematic use" [12, 14, 52]. It is therefore unsurprising that problematic technology use spans a vast range of phenomena across these studies [28]. Indicators of problematicity vary from impacted relationships and sleep [14] to more abstract notions like compromised self-control [54], or even technological analogues of "withdrawal" [34].

A second commonality shared by previous work is the **use of surveys and psychometric scales** to diagnose or identify problematic use. These are typically administered to a specific sample of participants, e.g. adolescents [54], students [55], video-game players [51], and adult employees [63], from which at-risk individuals are then identified. A common analytical strategy is to compare at-risk participants with the remainder of the sample with hypothesis tests. This has led to results that illustrate differences in problematic usage effects between demographics like gender and age groups [12, 14, 62]. The vast majority of studies find clear usage differences between problematic and non-problematic groups; those who are identified as exhibiting problematic use are shown to spend more time on the platform, regardless of whether the measures of time spent are self-reported [38, 54, 57, 62] or data-derived [12, 14, 37, 52]. We discuss this further in Section 2.2.

Note that, similar to the debate over the psychological underpinnings of problematic use, there have been no universally recognized diagnostic criteria for those afflicted. Decision boundaries vary between papers, such as answering positively to a fraction of binary questions [62], scoring 3 out of 5 points on Likert scales [38], or scoring 40 out of 60 total survey points [37].

Thirdly, the majority of existing research on problematic effects is primarily focused on usage of **closed, for-profit platforms** that have inaccessible data and may be designed to incentivize user engagement. For example, previous work on problematic use of social networking sites is often centered on Facebook and similar monetized platforms [6, 14, 31, 48]. In the area of video-games, studies have asked participants about commercial games of the Massively-Multiplayer Online Role-playing Game (MMORPG) subgenre [12]. However, reproducibility of work on closed platforms is limited due to inaccessibility of the participants and data employed. This restricts the scrutiny and extension of existing work only to members of the scientific community affiliated with the platforms. Unaffiliated research analyzing data from closed platforms often has to develop bespoke trackers [24, 37] or analyze screenshots by hand [23], thus impacting their scalability and cost-effectiveness. Similarly, participant pools and recruitment methods are limited outside of these platforms, again increasing the effort and cost required to conduct large-scale research for the unaffiliated (e.g. [3]). Furthermore, it is unclear whether existing results are confounded by e.g. engagement-maximizing mechanisms [20]. Problematicity within a platform could potentially vary more drastically if these mechanisms disproportionately lead certain subgroups to overuse the platform, as opposed to a platform without these incentives. Research on the association between problematic effects and using open, non-profit platforms is therefore not only desirable to avoid these pitfalls; non-profit platforms are also *understudied* in existing work. While some work investigates broad categories of online platforms (e.g. asking about gaming habits in general [38, 51, 57]), none explicitly consider open, non-profit variants.

Recently, related concerns over a lack of transparency in this line of research have accelerated. For example, questions remain over the significance of the results in existing work, with some arguing that the most drastic relationships between technology use and problematic effects are only specific to certain analytical decisions (cf. [45, 56]). In particular, large-scale survey studies with many response variables allow for a vast range of possible construct operationalizations and model choices. This could potentially lead to results being (knowingly or unknowingly) cherry-picked for effect size or statistical significance. Combined with the focus on inaccessible, closed platforms and vast selection of psychometric instruments, these concerns emphasize the need for transparent methods in the study of problematic effects of online platform use.

Relation to this study. Our work diverges from these commonalities in several ways. Firstly, our work investigates Lichess, a non-profit, open platform with publicly-accessible datasets. The platform's openness allows for more transparent research to be conducted, scrutinized, and extended by the wider academic community. This openness not only stems from its publicly-available data[3], but extends even to its open-sourced code that can be freely inspected and re-purposed[4]. The absence of engagement-maximizing mechanisms on Lichess also distinguishes it from platforms that are examined in existing work, many of which often monetize attention through e.g. advertisements (e.g. [6, 14, 16, 31]). Our work thus adds to the existing body of literature on online well-being; it

[3]As stated on https://database.lichess.org: "All games played on lichess.org are in the public domain. These collections of games are in the public domain, with no rights reserved. Use them in any way you like, for data mining, research, commercial purpose, publication, anything. You can download, modify and redistribute them at will, without asking for permission."

[4]See https://lichess.org/source.

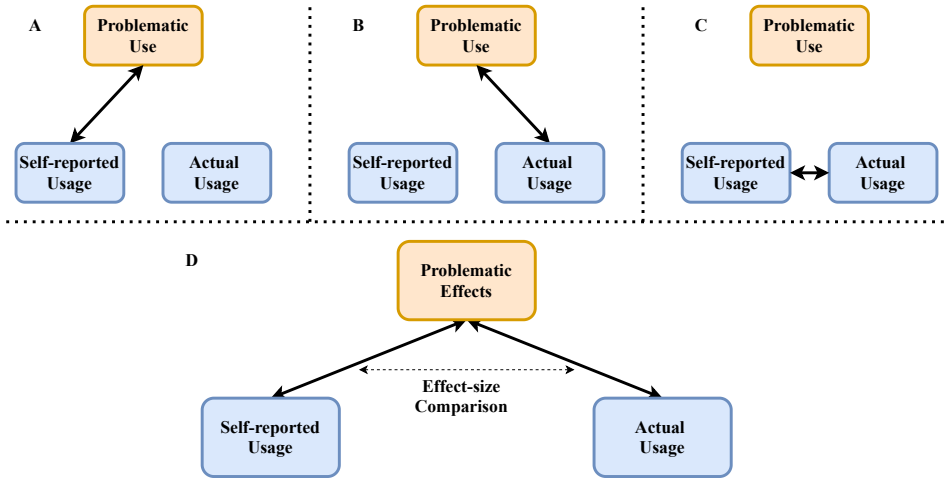

Fig. 1. *A*: majority of existing work relating self-reported use and problematic use (e.g. [17, 28, 38, 51, 57, 63]). *B*: extensions using e.g. server and device logs instead of self-reported use (e.g. [12, 14, 37, 52]). *C*: unrelated work comparing self-reported and actual measures of time spent online (e.g. [9, 18, 24, 25, 59]). *D*: our current work comparing self-reports and problematic effects versus actual measures and problematic effects.

is unknown whether using explicitly prosocial platforms can be correlated to problematic offline effects.

Secondly, we measure negative psychological and social effects like compromised relationships, sleep disruption, and impacted school and work performance that are commonly associated with technology use [14, 27, 28, 34, 38, 42, 54, 61], as well as negative affective outcomes like loss of control and regret [16, 27, 28]. To this end, we find that the survey employed by Cheng et al. [14] for studying problematic Facebook use to be the most relevant as it assesses all of these outcomes. Its brevity also suits online participant groups that may be prone to attrition. We thus use a minimally-adapted version for Lichess in this work (see Section 3).

However, instead of diagnosing a binary label of "addiction" or other underlying psychiatric phenomena, which is pervasive in the aforementioned pathological use literature, we are more interested in the correlation between usage and the effects themselves. We also do not identify more abstract, clinical constructs like salience, tolerance, and withdrawal [28, 31, 51]. On the one hand, it is unclear whether hobbies like chess can rise to the clinical significance of "addiction", at which point these constructs become more relevant. On the other, diagnosis would require administering lengthier surveys containing full psychometric scales. This may not only impact attentiveness but could also disillusion participants who feel strongly positive about chess. Nonetheless, chess has the potential to induce negative effects that are still undesirable, if not formally clinical[5]. We thus focus our work on the negative effects that can arise from using Lichess.

2.2 Time and Problematic Effects

A central question in the problematic use literature is how problematic effects of online platform use are associated with time spent on the platform. Many survey-based studies investigate this by correlating self-reported usage with perceived negative outcomes, with most finding at least a loose relationship between problematicity and self-reported time spent [17, 38, 51, 57, 63]. This approach

[5]We find many anecdotal accounts online, e.g. searching Google with the query *site:reddit.com/r/chess "addicted"*.

is depicted in Figure 1*A*. Note that negative correlations between well-being and time spent are not universal; some have found subtle, potentially positive associations instead [46]. Nonetheless, others postulate a causal relationship between time spent online and problematic effects, such that self-reported total use should be used to provide criterion validity for psychometric problematic-use scales [28]. In other words, for surveys to validly measure self-reported problematic effects, their outcomes are expected to correlate well with self-reported platform use.

However, it is well-known that self-reported measures of time spent correlate only weakly with actual usage metrics derived from server logs. For example, playing time estimates per week from MMORPG gamers are systematically inaccurate and under-reported, particularly amongst female, older, and more educated players [25, 59]. Self-reports from mobile phone users do not reflect even the previous day's log data with high fidelity [9]. Users of Facebook and Twitter are also known to be unable to reliably recall the amount of time they spend on social networks [24]. Indeed, recent recommendations urge researchers to replace self-reported usage with data from server logs and monitoring applications when possible [18]. These studies directly compare self-reported and data-derived measures of time spent, independently of their relationship with problematic use, as depicted in Figure 1*C*. Since the relationship between self-reported and data-derived measures of times spent is weak, it is possible that studies of problematic use that rely on self-reported measures may not accurately uncover how problematic use relates to time spent online.

In response, a growing number of studies replace self-reported measures of time spent with data-derived measures to investigate problematic use. This is depicted in Figure 1*B*. Work on smartphones, for example, has linked assessment survey responses to digital traces of smartphone use [37, 52]. Those with access to Facebook data have linked usage habits and demographic information to problematic use assessment surveys [14]. A similar technique can be used to investigate players of online games [12]. In principle, data-augmented surveys can circumvent obstacles that lead to inaccurate self-reported independent variables – server logs are scalable, granular, and immune to biases and recall issues that can affect participants [49].

Relation to this study. While these three groups of existing work compare self-reported use, actual use, and problematic use separately as depicted by Figure 1*A, B, &* C, none combine all three.

Whether actual, data-derived measures and self-reported measures have similar, different, or even complementary relationships with problematic effects is unclear. It is plausible, for example, that substituting data-derived metrics for self-reported numbers may obscure information that is contained in subjective, perceived usage quantities. Additionally, discrepancies between perceived and actual usage time could, for instance, be indicative of self-regulatory failures [51] and correlate with problematic use. To our knowledge, neither of these questions have been addressed in existing work. As a result, the relationships between problematic outcomes and both self-reported and actual time together merit further investigation, which we undertake in this study. Figure 1*D* visualizes comparative associations between self-reports, actual data, and problematic usage effects.

Beyond addressing empirically unanswered questions, clarifying these comparative associations can provide utility to guide future screen-time research. Firstly, if actual data were much more closely related to problematic effects, then correlating psychometric scales with self-reported usage may not be the most appropriate measure of criterion validity [28]. Instead, correlations between trace data and survey responses would be stronger measures of validity in this case. Secondly, if self-reports and actual use have complementary relationships with negative outcomes, both measures may need to be combined to more accurately predict problematic effects. Eliciting self-reported usage is also significantly less invasive and demanding than administering full, lengthy psychometric scales. Demonstrating that self-reported and actual usage can be combined to proxy problematic effects could lead to, for example, more intelligent and noninvasive interventions for

Statistic	*Description*
Average time played	Mean time (mins) played per playing day.
Playing days	Number of days on which player was active.
Max time	Most time (mins) spent on one calendar day.
Max day	Calendar day on which player spent most time.
Weekly pattern	Mean time (mins) played per day of the week.
Max day of week	Day of week with highest mean time played.
Diurnal pattern	Mean time played each hour of the day, among active days.

Table 1. Summary statistics extracted from game logs for June 2019, per player.

enhancing digital self-control [30, 40]. Another possibility is larger-scale imputation of problematic effects on the general populace to better understand their prevalence. Currently, the majority of existing findings on prevalence is dependent on applying full psychometric scales, which restricts participants to those with substantial incentives for study completion (e.g. course credit for students; see [28, 33, 54]). Being able to approximate these scales with trace data and light-weight self-reports would make study completion much easier for a wider segment of the population.

3 METHODOLOGY

Our research design involves administering a gamified survey to Lichess users to elicit their self-reported time usage and their self-perceptions of the extent to which problematic effects arise from their Lichess usage. We then combine these measures with actual usage metrics derived from Lichess server data. We now present the data analysis and survey design steps we followed to create this survey, which consist of two main parts. First, we analyzed the complete chess game logs to extract playing times and summary statistics of each player's usage in June 2019. Then, we integrated these statistics into a quiz that elicited both the players' self-reports of their time spent and their perceptions of the impact Lichess had on their personal lives during the month.

3.1 Game Log Analysis

To extract objective, data-derived measures of how much time each player spent on Lichess, we obtained the digital traces of all rated games played in June 2019 from publicly available server logs. This data includes how long each player spent thinking about every individual move down to the tenth of a second. For each game, we computed how much time each player spent playing by summing over the time taken per move for both players for all moves in the game. In this dataset, 590,000 users spent 4.9 million in-game hours (563 person-years) playing 34 million games during June 2019.

Using this dataset, we calculated various usage time metrics for every player. Following previous work, we first derived overall measures of time spent on the platform. For example, we measured the total time spent per player by aggregating over all games they played in the period. We also counted the total number of distinct days each player spent some time on the platform.

In addition to these *aggregate* measures, we also calculated more *granular* measures for each player at the day-, week-, and month-level. For example, we measured each player's diurnal playing pattern by calculating, on average, how long they played during each hour of the day. Similarly, by aggregating the data at a daily level, we calculated how much time each player spent on average for each day of the week. We also measured the maximum amount of time players spent on a particular day, and which hour of the day they spent the most time playing (Table 1). In contrast to related work on online habits that only consider total time within a given interval (cf. 12, 14, 25),

Question	*Type*	*Choices*
A1a: Since you joined Lichess, which game type have you played the most games in?	Categorical	*13*
A1b: ... and for that game type, guess your current rating.	Numerical	-
A2: During June, on which day of the week do you think you spent the most game time on Lichess?	Categorical	*7*
A3: On how many days in June do you think you played at least one game on Lichess?	Numerical	-
A4: On the days you played, how many minutes per day do you think you spent in-game on Lichess?	Numerical	-
A5: Of the following three days in June, on which do you think you spent the most time in-game on Lichess?	Categorical	*3*
A6: What do you think was the amount of game time, in minutes, that you spent on that day?	Numerical	-
A7: During which hours of the typical day in June did you spend the most time on Lichess?	Categorical	*8*

Table 2. Questions asked in Phase A of the interactive survey, arranged into three separate pages. Note that A1a displayed all variants provided by the Lichess API; A5 presented the days on which the maximum, median, and least amount of time was spent by the player; A7 grouped hours of the day into eight three-hour blocks. Participants' reported time zones were used to shift and localize the hours shown in A7.

we elected to decompose time spent into finer-grained quantities. For example, a player's total time played in June can be separated into the number of days they were active and the average time they spent per active day. This allows us to perform a more granular analysis of problematic usage effects and habit-awareness, such as whether a player knew how many days they were online.

We note that all of our measures of time spent are derived from the in-game data described here. We are thus not counting the time players spend between games, either searching for the next game, watching other games, or chatting with other users.

3.2 Interactive Survey Design

We now describe our survey design, the main component of our research methodology. At a high level, our goals for the survey are two-fold: first, to elicit user estimates of aggregate and granular measures of their time spent playing on Lichess; and second, to elicit their self-perceptions of the problematicity of their Lichess usage. To these ends, we designed our survey in two phases. The first phase was a gamified survey, where participants guessed various metrics of their time spent on Lichess and then received feedback on how accurate their guesses were. The second phase was a problematic-effects assessment scale, where we asked participants questions about their Lichess usage derived from previous work on problematic use.

The architecture for our interactive survey consisted of a Web application hosted on AWS instances serving the game log statistics in Table 1. The only identifiers in this data were the players' Lichess usernames, which are already publicly displayed on the platform itself. We distributed the survey with Qualtrics, which allowed us to counterbalance the two phases of the survey.

Before users took the survey, we displayed an institutionally-reviewed form eliciting their informed consent. On an introduction page, they were asked to enter their Lichess username, which they were required to verify upon completion by messaging an account we created on Lichess. This enabled us to connect participants' survey responses with our data-derived measures of their time spent on the platform. Participants were also asked to report their local timezone so that we could adjust their diurnal patterns accordingly.

The main body of our survey contained two main phases: a quiz and a scale.

Phase A: Quiz. We incorporated the statistics we extracted from the Lichess game logs into a gamified quiz, in which participants were tasked with estimating their actual patterns of online chess play in June 2019. We assessed the summary statistics described in Table 1 with three numerical and three categorical questions. We also asked participants to guess their most-played chess variant

and their Elo chess rating for that variant prior to these questions[6]. Players' modal variant and ratings were gathered live from the Lichess API to ensure they were up-to-date.

We separated quiz questions into three distinct pages as summarized in Table 2. The second page containing A2-6 also presented participants with an empty calendar of the month to aid recall.

In order to gamify this phase and attract users of the platform, we presented answers to the questions after each page of questions, as well as interactive visualizations of their usage statistics under consideration. We grouped questions into pages such that receiving the answers to the questions on one page didn't affect their ability to answer questions on subsequent pages. Examples of these are illustrated in Figure 2. Upon finishing the quiz, we asked participants an open-ended question asking them to reflect on how they approached the self-assessment questions ("Please explain how you answered the questions that we just asked about your time on Lichess. Did you have any specific thought processes or strategies?"). Finally, a score out of seven summarizing their performance in the quiz was shown to participants at the end of the phase, along with a statement congratulating or consoling them on their results.

Phase B: Problematic Effects Survey. In contrast to the quiz-like formulation of Phase A, Phase B presented questions to participants intended to assess the degree to which they perceive problematic effects stemming from their Lichess use. Specifically, these are negative offline effects like compromised relationships, altered sleep, and impacted school and work performance, all of which have demonstrated correlations with problematic technology use [14, 27, 28, 34, 38, 42, 54, 61]. Similarly, we also measure negative affective outcomes like loss of control and regret [16, 27, 28]. We chose these effects because they are almost always measured in existing psychometric instruments, and are more translatable to chess than, for example, depression and loneliness on social networks [23, 31, 56].

We found that the scale used in a recent study by Cheng et al. on problematic Facebook use assessed the vast majority of these variables, and so we chose to administer a slightly modified version of their scale in this work [14]. Its brief and lightweight format lends itself to easy completion by online participants, who may otherwise be prone to inattention or dropout in longer questionnaires. While its questions are derived from the same foundational psychology literature, it is also more neutrally coded than surveys that are designed to diagnose pathological constructs (see e.g. [6, 35, 38, 54, 61]). Its neutrality prevents participants from being lead and avoids disillusioning users who feel strongly positive about chess.

Importantly, Phase B questions are not delineated into more abstract, psychiatric constructs like tolerance or withdrawal (e.g. [28, 38, 54]). Because it is unclear whether chess play could rise to the clinical status that these constructs demand, we instead measured correlation between platform use and the negative effects themselves. We further observe that Cheng et al. also do not decompose their definition of problematic Facebook use into these clinical constructs. To indicate that the survey was about players' June 2019 habits only, we prefaced the survey with the phrase "During the past month (June 2019)...". All questions were coded on 5-point Likert scales, and are presented in Table 3.

Debrief. After participants completed both phases, they were asked to report their age range (10-year intervals), gender, and country of residence if they were willing. A debrief page was also presented to the participant upon conclusion of the activity, which asked them whether they had used external sources to aid their answers in the quiz. This did not impact their entry into the prize draw. Participants were prompted to provide any thoughts in an open-ended question during the

[6]Chess variants refer to different types of chess games one can play — either with rule changes or different time limits. Lichess maintains separate chess ratings for each variant to measure how skilled players under these various conditions.

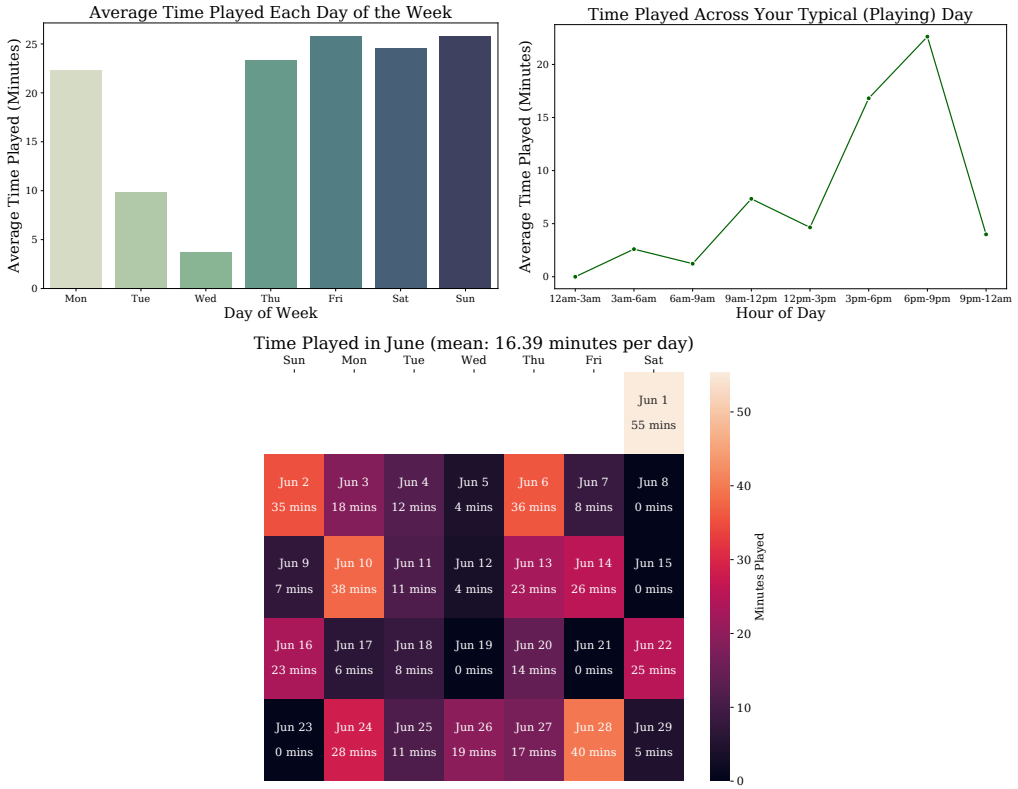

Fig. 2. Example visualizations presented to participants along with answers to preceding questions in the Phase A quiz. *Left*: bar graph of average time on each weekday; *Right*: diurnal time of average play time on active days; *Bottom*: Calendar representing in-game time each day. Hovering over the images while taking the survey provided additional information, such as the percentage of time represented by each point in the graph.

debrief. They were finally required to verify their response by sending a message to our Lichess account.

We counterbalanced the two phases of the survey such that their ordering was randomized between participants. The introductory and debriefing questions were fixed respectively at the beginning and end of the activity.

3.3 Pilot

We initially launched a prototype of our survey with questions adapted for online chess from the Gaming Addiction Scale ("GAS" [38]), which we believed may have some applicability to Lichess as one of the shortest inventories we found for identifying pathological gaming. However, we decided to use the shorter and more neutrally-coded questions from Cheng et al. instead [14]. This was due to several reasons. Firstly, pilot participants ($N = 14$) expressed concerns about the negativity in the GAS questions, nearly all of which alluded to strongly problematic effects (e.g. "Did you think about playing the game all day long?"). These questions not only appeared to be out-of-place for an intellectual hobby like chess, but the constructs they were designed to measure may not be directly applicable (salience, in this case). Secondly, participants also disliked the lack of opportunities to

B1: Overall, do you feel like Lichess has had a positive or negative impact on your life? *Very negative impact, Somewhat negative impact, Neither positive nor negative impact, Somewhat positive impact, Very positive impact*
B2: How often did you get less sleep than you want because you're playing on Lichess? *Never, Rarely, Sometimes, Often, All the time*
B3: Overall, how much did your time on Lichess hurt or improve your relationships with others? *Hurts a lot, Hurts a little, Neither hurts nor improves, Improves a little, Improves a lot*
B4: To what extent did Lichess help or harm your work or school performance? *Helps greatly, Helps somewhat, Neither helps nor harms, Harms somewhat, Harms greatly*
B5: How much control do you feel you have had over the amount of time you spend playing on Lichess? *No control, A little control, Some control, A lot of control, Complete control*
B6: How often do you play on Lichess and then later regret it? *Never, Rarely, Sometimes, Often, All the time*

Table 3. Scale questions asked in Phase B of the interactive survey. All were presented as 5-point Likert scales.

express positive experiences with Lichess due to questions being negatively phrased throughout the survey. We thus found the modified Facebook survey to be more appropriate for our target platform and participants.

Our pilot findings also motivates our focus on degrees of problematic effects rather than diagnosing pathological gaming habits. If Lichess users were generally positive towards the platform but felt lower, non-clinical levels of negative outcomes, pathological surveys with a formal diagnostic cut-off may miss these subtler problematic effects. Additionally, it also reinforces our choice not to delineate survey responses into commonly analyzed psychosocial constructs like salience, tolerance, and withdrawal [28, 33, 51]. This would require administering lengthier surveys containing full psychometric scales, which may not only impact attentiveness but could also disillusion participants who feel strongly positive about chess. We believe our current approach is thus a reasonable trade-off to afford participants space to be engaged and voice positive sentiments.

3.4 Recruitment and Deployment

We distributed our survey with Reddit posts in the *r/chess* community in July 2019. We timed its launch to coincide with the release of data on Lichess at the end of June, and to avoid days on which major chess tournaments were held or when significant chess news was spreading.

For the purpose of encouraging participation from Reddit members, we entered participants who completed the survey into a draw for several $20 USD monetary prizes. Each participant chose between receiving an Amazon gift card or donating their winnings to Lichess on their behalf if they won. We also incentivized attention and engagement with the quiz-like elements in our survey.

We recorded 421 responses from 294 unique reported usernames during the week after our survey was launched. We considered only the first responses of participants who had successfully clicked through to the end of the survey, affirmed that they had answered seriously and without external aid, and had completed all required questions in both phases. This resulted in our final list of $N = 131$ participants. Of the 129 who entered their age and gender, all were below 55 years of age and 115 were below 35. 127 identified as male and 2 preferred not to disclose. On average, participants played 14.6 hours of chess in June 2019 ($SD = 16.4$), with the least totalling 2 minutes and the most having played 120 hours – almost equivalent to a full-time job.

Statistic	*Reported Mean*	*Actual Mean*	*Mean Δ*	*Median Absolute Error %*	*β*
Total time (mins)	942.7	874.8	68.0	48.5	0.721^{***}
Max time (mins)	129.2	144.6	−15.3	39.8	0.577^{***}
Mean time (mins)	48.5	53.8	−5.3	43.1	0.717^{***}
Playing days	17.7	15.1	2.6^{***}	25.0	0.716^{***}
ELO rating	1659.1	1641.3	17.9	0.9	0.954^{***}

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

Table 4. Comparisons of self-reported time measures from Phase A against their actual values, derived from the Lichess dataset. Total time is obtained by multiplying a player's conditioned average by their number of active days; the self-reported value is obtained the same way from A3 and 4. Standardized β values are obtained from univariate linear regressions between self-reported and actual values. p-values for regressions and t-tests are Holm-corrected.

4 RESULTS

We now analyze the survey responses to understand how self-reported and actual measures of time spent are associated with perceived problematic effects of using Lichess.

Preliminary Comparison of Self-Reports and Data. We first compare self-reported measures of time spent, i.e. participants' guesses in the quiz, against the objective, data-derived values. Although not directly addressing our research question (Figure 1*D*), we still conduct this preliminary comparison (Figure 1*C*) to ensure participants' self-reports are not significantly more accurate than those analyzed in studies on social network, games, and smartphones [9, 24, 25]. If they were atypically accurate, it may suggest that e.g. participants referred to their online profiles as recall aids in our study.

As one may expect, participants were very adept at guessing their most-played variant (A1a, accuracy of 0.824) and most recent chess rating (A1b, $\beta = 0.954$, $p < 10^{-4}$). However, the accuracy of their rating estimates is striking; we obtain a correlation coefficient of 0.996 ($p < 10^{-10}$) between self-reported and actual ratings when considering the 108 participants who correctly answered A1a. Only twenty participants knew their ratings exactly, suggesting that the majority of participants likely did not use their online profile for reference. These results provide evidence against several confounding factors: participants were attentive and keenly aware of their ratings, and appeared to self-report honestly without external aids.

In contrast, self-reported time measures were highly inaccurate. Individual answers typically deviated around 40% from their actual values. Furthermore, participants correctly answered the categorical questions A2, A5, and A7 approximately 22%, 40%, and 40% of the time, respectively, and A5 was not statistically distinguishable from random guessing (binomial, $p = 0.138$). Only 52% of the variance in actual total playing time and 33% of the variance in actual maximum time spent in a day were explained by self-reports (see Table 4).

These inaccuracies in self-reported usage measures are in line with existing research on video-games and social networking sites [24, 25, 50]. Compared to the observed under-reporting in that body of work, we find little evidence of over- or under-reporting (binomial tests, $p \geq 0.05$) in any measure except over-reporting in the number of playing days. On aggregate, mean self-reported time measures were not significantly different from mean actual time measures (paired t-tests, $p \geq 0.05$). This indicates that answers both individually over- and under-stated actual values to similar degrees, thus cancelling aggregate errors. One may also expect self-reports to be more accurate when participants were first asked to reflect on their online well-being in the Phase B scale. However, counterbalancing had no effect on the accuracy of any self-reported summary statistic

Statistic	*β (Reported)*	*β (Actual)*	*β (Actual - Reported)*
Total time	-0.275^{*}	0.303^{**}	-0.0159
Max time	-0.346^{***}	-0.265^{*}	-0.0954
Mean time	-0.236^{*}	-0.216^{*}	0.0041
Playing days	-0.259^{*}	-0.271^{*}	-0.0085
ELO rating	-0.130	-0.116	-0.0298

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

Table 5. Correlations between problematic effects of using Lichess and self-reported and actual measures of time spent on the platform. Standardized regression coefficients are obtained from univariate linear regressions; *p*-values are Holm-corrected. Note the significance of actual total and self-reported maximum time.

(unpaired t-tests, $p \geq 0.05$). The consistency of our preliminary results with existing work further reinforces that our Lichess participants can be comparable to those in previous studies eliciting self-reported use.

4.1 Problematic Usage Effects

We now turn to our central question and study how actual and self-reported measures of time spent are associated with problematic effects of using Lichess. Our procedure is as follows. First, we conduct simple regressions to compare the associations between self-reports, server-logs, and problematic effects. We then apply Specification Curve Analysis to our most predictive self-reported and data-derived time measures, as well as a combined model. This allows us to robustly compare effect sizes and goodness-of-fit under many different analytical decisions.

For ease of interpretation, we first recode the responses to Phase B scale questions so that answers expressing positive attitudes correspond to higher values on the 5-point Likert scales. For example, answering "*all the time*" to the question of whether one loses sleep due to online chess (B2) scores 1 point out of 5. We find that responses to the scale yield a Cronbach's α of 0.78, which indicates acceptable internal consistency and is within the range of established psychometric scales (e.g. 0.69-0.97 in gaming inventories [28]). We thus restrict our analysis to one dependent variable representing the sum of a participant's survey score. We operationalize attitudes indicative of feeling problematic effects as being reflected by lower total scale scores. Note that we interpret this score more as an index or measure of problematic effects, rather than a single, pathological construct such as chess "addiction".

On aggregate, participants scored a mean of 21.7 out of 30 ($SD = 3.78$, $min = 9$, $max = 30$) on the scale questions. While this indicates mainly positive sentiments about Lichess, a significant fraction of participants ($N = 59$, 45%) also scored under 3 on at least one scale question. In other words, these participants reported that Lichess had a noticeable non-neutral, potentially negative impact on at least one aspect of their lives. A further 13 participants (10%) scored below 3 on at least 3 out of 6 questions, indicating that they reported net-negative effects arising from using Lichess on our survey. This is particularly surprising given that Lichess does not incentivize user-engagement and is centered only on an intellectual pastime.

Simple Regressions. To investigate how measures of time spent relate to the problematic usage effects outlined above, we conduct univariate linear regressions using the self-reported and actual numerical statistics to predict overall scale scores (Table 5). Time spent is clearly associated with lower scores — all measures of game time that we consider have significant, negative effects, and are consistent across both self-reported and actual measures of time spent. Thus, despite the very

weak correlation between self-reports and actual quantities that we found above, both actual and self-reported measures have some predictive power. For example, the relationship between both measures of total time and problematic effects is consistent with the body of work we described in Section 2.2, albeit reflecting a smaller effect size (compare to correlations of 0.07 to 0.64 [28]).

Furthermore, the differences between self-reported and actual time measures — i.e. how much users under- or over-reported — have no effect on potentially problematic effects. In contrast, one may have expected those who have worse track of time to play more excessively, and therefore to score lower on the Phase B scale. We also checked whether counterbalancing had any impact on participants' scale scores and found no statistically-significant aggregate differences (unpaired t-test, $p \geq 0.05$). This is surprising, as being asked to reflect on especially lengthy playing sessions or days could plausibly have induced feelings of problematic behavior. However, we found no evidence for this effect.

Of the individual measures, self-reported maximum time played in a day and actual total time played during the month are most correlated with attitudes that reflect potentially problematic behavior. We observe that these are very different measures: the former is *granular*, representing usage time in a single day, and *self-reported*, whereas the latter is an *aggregate*, representing usage time across the entire month, and *data-derived* from participants' actual usage histories. Existing literature provides abundant evidence supporting the link between problematic effects and data-derived total usage time [14, 37, 52]; however, we found only one study that describes a similar relationship with maximum usage time [54]. This is especially surprising given the common comparison of scale responses against total time as a means to assess criterion validity in existing problematic use literature [28].

Thus, because it outperformed actual total time as a predictor of problematic effects, self-reported maximum time needs more rigorous investigation to ensure that this is not by chance. For instance, these results are derived from simple regression that could otherwise have included or excluded a number of potential covariates. Additionally, we seek to understand how well the *combination* of actual and self-reported measures predict the problematic effects of using Lichess. To carry out this more involved analysis and ensure that effect sizes are not incidental, i.e. specific to the simple regressions we conducted, we use Specification Curve Analysis. Because one self-reported and one actual measure significantly outperformed other variables, respectively self-reported maximum time and actual total time, we focus our application of SCA below on these two independent variables.

4.2 Specification Curve Analysis

One way of more rigorously checking whether these surprising results are limited to one instance of one analytical model is to use Specification Curve Analysis (SCA). Instead of fixing the covariates that are included or excluded in a single regression, this technique performs a combinatorial search through all reasonable analytical specifications, so that effect sizes are not cherry-picked for size and significance [53]. Using SCA reinforces our decision to study problematic effects of using an open, accessible platform. It transparently presents the results of many analyses at the same time and is therefore resilient to data dredging and manipulation. Recently, SCA has also been used to study the impact of screen-time on youth well-being and life satisfaction [44, 45]. We believe SCA is the most appropriate analysis tool for this work, not only because of the many possible covariates we obtained from our survey and data analysis but also because of its application to related problems.

In the following two SCAs, we identify up front the set of justified, statistically valid, and non-redundant analytic *specifications* one could use — in our case, a specification is the set of covariates we choose to include in a regression. We then test all possible specifications (in our case, $2^{13} = 8192$

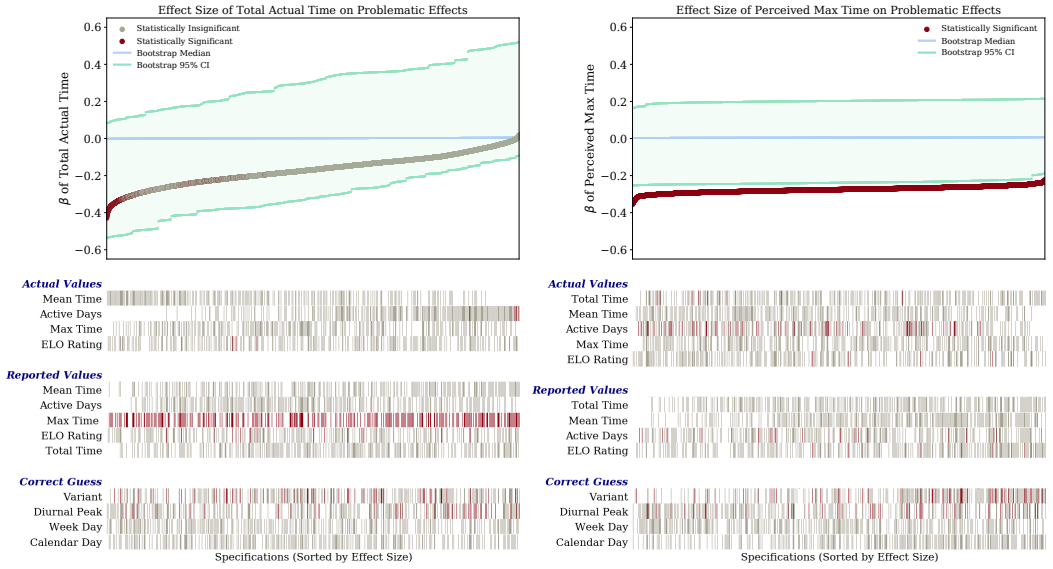

Fig. 3. Figure showing effect sizes of total playing time (left) and self-reported maximum time (right) on Phase B scale score. *Top*: standardized regression coefficients of the target independent variable when predicting scale score, per specified regression. *Bottom*: covariates selected in each of 8192 specifications, aligned with the target independent variable's effect size in top plot. Each tick represents a covariate in a specification; red ticks indicate statistically-significant regression coefficients. *Left*: SCA with actual total time played as target independent variable; median effect −0.169. *Right*: SCA with perceived maximum time played as target independent variable; median effect −0.275. Light blue and green lines depict bootstrapped median and 95% CI SCAs.

regressions), graphically display the results for all specifications simultaneously, and statistically test whether we can reject the null hypothesis with respect to this entire set of specifications, as opposed to making arbitrary decisions to choose one single specification.

Specifications. We measure problematic usage effects using the total scale score from our survey. The models we use are first-order, multivariate linear regressions to predict total scale score, the dependent variable, from the independent variable, which is actual total time played in the first SCA and self-reported maximum time in the second SCA. We are primarily interested in the effect size of the independent variable on scale score under different combinations of covariates. These include:

- *Actual usage measures* from a participant. These include the total time played, mean time on active days, number of active days, maximum time in a day, and rating.
- *Self-reported measures*. These are the self-reported figures corresponding to each actual characteristic.
- *Self-awareness of usage habits*. These include whether the participant answered each categorical question in Table 2 correctly (most-played variant, max day of week, max calendar day, and max 3-hour block of the day). Differences between actual and self-reported numerical characteristics are also implicitly included when both values are selected in a specification.

Any subset of these 14 variables, excluding either actual total time or self-reported maximum time, could be feasibly used as covariates in a multivariate regression to predict scale scores. Our SCAs are thus conducted with $2^{13} = 8192$ specified regressions.

Creating and Interpreting the Curve. The SCAs are illustrated in Figure 3, where all 8192 specifications are plotted along the x-axis and are sorted by their effect size (standardized regression coefficient) of total playing time and self-reported maximum time on scale score. The further above or below zero the thick grey and red curve is, the more positive or negative the independent variable's effect size is on scale score (negative for problematic effects). Flatter curves indicate that the effect size's magnitude and sign are consistent across specifications and cannot be cherry-picked. A red point indicates the corresponding specification's effect size is statistically significant, and a grey point indicates otherwise. Thus, darker curves with more red points indicate that many specifications yield statistically-significant effects, whereas spotty, lighter curves suggest that many effects are not. The green points correspond to a bootstrapped confidence interval, which we discuss below.

Examining the first SCA (Figure 3, left), where we are measuring the relationship between actual total time played and problematic effects scale score, we find that the median effect size across all specifications is -0.169. Out of 8192 possible regressions, 8136 of them have negative correlation coefficients for total time, a further 689 of which are statistically significant. In the second SCA (Figure 3, right), where we are studying the relationship between self-reported maximum time played in a day and problematic effects scale score, we find that the median effect size is -0.275. Furthermore, all 8192 specifications have significant, negative effects. Thus, although actual total time is predictive of perceived problematic outcomes of using Lichess, self-reported maximum time played has a much stronger and more significant effect on lower scale scores in all reasonable linear models.

This is a striking result. While previous work has found connections between problematic use and both perceived [17, 57] and actual [14, 37, 52] total time, there is little existing evidence of a correlation between perceptions of peak episodic usage and problematicity [54]. Furthermore, the SCA for total time in Figure 3 is consistent with arguments that screen-time is only loosely related to well-being under certain analytical assumptions [45]. For example, by picking a specification on the right edge of the plot, e.g. including actual number of active days and excluding actual mean time as covariates, one can craft a regression showing a slightly positive, statistically insignificant relationship between actual use time and problematic outcomes. This could be used to craft a narrative that, for example, use time and problematicity are unrelated. However, our SCA for self-reported maximum time shows that it is significantly correlated regardless of our choice of covariates, and is thus completely resilient to cherry-picking.

Significance Testing. This result is further reinforced when we test whether our results could arise from the null hypothesis that *actual total (or self-reported maximum) time played is uncorrelated with problematicity*. We constructed 10,000 bootstrapped specification curves, in each of which we randomly shuffled scale scores between participants. This process involved over 81 million multivariate regressions for each SCA to test the null hypothesis. We considered three test statistics: the median effect size of each bootstrapped curve, the share of specifications in each curve that are the same sign as the majority of specifications, and the share of these specifications that are also statistically significant [53]. We derive p-values from the number of bootstrapped curves that have as- or more-extreme values than our observed curve, and are tabulated in Table 6.

We find that the effect of actual total time on scale score is significant in the median test ($p < 0.05$), and that the effect of self-reported maximum time is significant under both the median and dominant-significant share tests ($p < 0.0001$). Thus, while it is likely that total time is associated

Test Statistic	*Observed Value*	*Bootstrapped p*
Actual Total Time Played		
Median	−0.169	0.043
Dominant Share	8136/8192	0.081
Dominant & Significant Share	689/8192	0.201
Self-Reported Maximum Time Played		
Median	−0.275	0.0009
Dominant Share	8192/8192	0.146
Dominant & Significant Share	8192/8192	0.0003

Table 6. Bootstrapped test statistics for the specification curves in Figure 3, derived from 10,000 samples.

with perceived problematic usage effects, it is almost certain that self-reported maximum time is strongly predictive of these effects. This is regardless of participant's playing characteristics, how well they knew their playing patterns, and how these are operationalized and specified by regression variables.

Combining Total and Perceived Maximum Time. The previous two SCAs established that actual total time and self-reported maximum time are each independently associated with our scale score. But how informative is combining the two forms of time measurement? Are data-derived and self-reported measures largely capturing the same information, or do they contribute different kinds of information?

To answer these questions, we ran an SCA to investigate the relationship between the problematic effects scale score and both actual total time and self-reported maximum time. Since we now have three effects instead of one (actual total time, self-reported maximum time, and their interaction), we use adjusted R^2 instead of effect size as our measure of informativeness. For reference, we find that in the two SCAs from the previous section, actual total time and self-reported maximum time yield median adjusted R^2 values of 0.118 and 0.136, respectively.

In contrast, this SCA — with total actual time, self-reported maximum time, and their interaction term — yields a median adjusted R^2 of 0.232 (shown in Figure 4). The amount of variance in problematic effects score explained by actual and self-reported time together is almost double the variance explained by either alone. Once again, all 4,096 possible specifications are statistically significant, suggesting that this result is extremely robust to covariate choice. Furthermore, the adjusted R^2 is far outside of the bootstrapped null hypothesis confidence interval in every specification, reiterating the significance of this finding.

The combination of objective and subjective time measures performs markedly better than using either a single self-reported or actual measure of time alone. Clearly, both survey-recorded and data-derived measures of time provide useful and distinct information about the problematic effects that can arise from platform use. If self-reports were simply noisy proxies for actual data that have weaker associations with problematic effects, one might expect the addition of self-reports to a data-driven model to yield little improvement. We thus conclude that both are necessary in predicting problematic effects. While studies relying on self-reports can clearly benefit from having more accurate measures of actual behavior through data, those relying on data should not disregard self-reported usage as imperfect proxies whose true association with dependent variables are already captured by e.g. server logs [18].

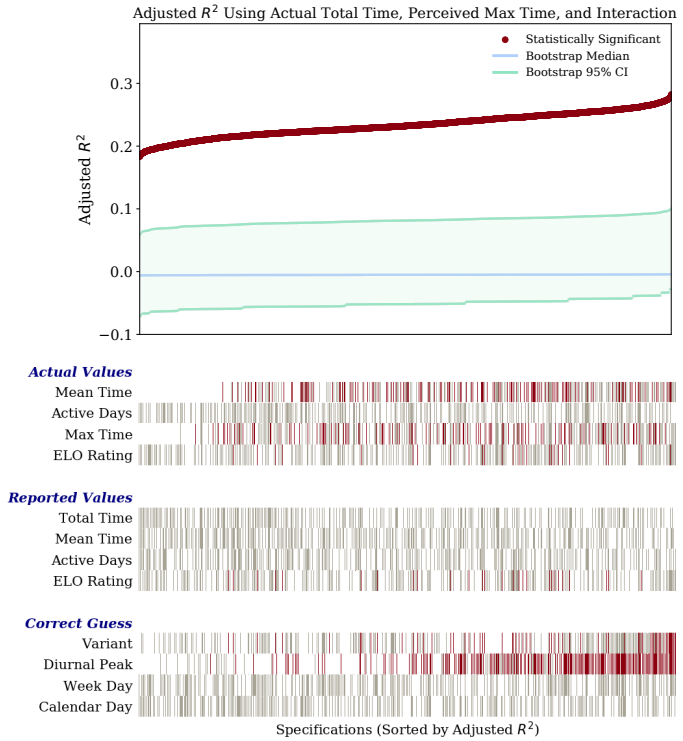

Fig. 4. SCA depicting R^2 of multivariate regressions using actual total time, self-reported maximum time, and their interaction term to predict Phase B scale score. Median $R^2 = 0.232$, compared to 0.118 and 0.136 in Figure 3.

5 RESPONSES TO OPEN-ENDED QUESTIONS

We included two open-ended questions in our survey: one after the quiz asking participants how they estimated their playing habits, and one during the debrief asking participants for thoughts on the activity. Both questions were optional, but nonetheless elicited 85 and 35 responses respectively (62 and 24 with more than 10 words). We performed thematic analysis on these answers using iterative, open coding. Because our survey was not designed for rigorous qualitative analysis, we briefly present these not as formal results (cf. 7) but as a source of context for interpreting our quantitative findings.

Reflections on Usage Habits. As may be expected, a majority of responding participants indicated that they answered Phase 1 with instinctive guesses rather than sophisticated strategies ($N = 44$). For example, one participant noted that they *"took random stabs at it, just went with the first gut number"* (P72), while another said they used no strategy *"beyond instinctual estimates"* (P9). However, several respondents also alluded to using knowledge of their diurnal and weekly playing patterns to answer our questions ($N = 17$ and 21 respectively). One noted: "*I was able to answer some questions (time of the day I usually play, number of days I played at least one game in June) with confidence because I'm conscious of my daily habits*" (P81).

Some also invoked their real-life commitments ($N = 20$), often acknowledging their weekly and diurnal patterns in conjunction. For example, one answered: "*I know I play a fair amount almost*

every single day, and I know I play on my lunch break at work and if I'm off it's right after my kid goes to bed" (P59). Another observed: "*... I know I generally spend an hour or two most days playing chess especially after I've come home from work and had a chance to relax which is what I used to make my general estimate*" (P95). Answers like these suggest that deconstructing the notion of "total time" into more granular statistics may have helped participants contemplate more deeply about playing habits and their real-life implications. Indeed, 12 participants appreciated the insight into their usage habits afforded by our activity, such as P111: "*...it's nice how you can see the time spent playing for each day of the month. I wish I could see that anytime I wanted to.*" It is therefore surprising that counterbalancing the phases of our survey had no impact on the Phase B scale scores. Those tasked with reflecting on their usage first felt similarly problematic to those who immediately completed the scale questions.

Negative Effects of Chess. Although neither of our open-ended questions asked participants about problematic outcomes arising from chess play, several participants also mentioned potentially problematic behaviors in answering the two questions ($N = 2$ and $N = 4$ respectively). For example, one answered "*I'm addicted to bullet chess, and have known that for years*" (P60); another participant echoed this sentiment: "*... I know that I spend a lot of time on playing chess online as a form of procrastination, which is a problem I'm currently working on*" (P81). Upon seeing their visualized playing patterns, P111 remarked "*Also the time I spent is shockingly large.*" One of the more extreme examples we found was: "*Dude lichess actually fucked up my last relationship. I went to the bathroom for 15 minutes during a date to play a game and that's just the tip of the iceberg*" (P89). These answers are particularly valuable because they suggest that how people use even a seemingly innocuous board-game platform can have perceived negative psychosocial consequences.

Despite these explicit expressions of negativity, no participant would classify as a problematic user according to criteria laid out by Cheng et al. (2 points or lower on one question each in Q1-4 and Q5-6 respectively [14]). This further justifies our interpretation of the survey scores as a measure of the degree to which one perceives negative usage effects, rather than a diagnosis of usage patterns that automatically become problematic once the score surpasses a numerical boundary. It also suggests that strict diagnosis criteria may lead to oversight of subtler, more nuanced negative outcomes like the ones we find in this study.

Lichess as Enrichment. In contrast, some responses to the debrief question highlighted enriching aspects of online chess, especially Lichess, and even expressed concerns that Phase B questions were too negatively-skewed ($N = 4$). For instance, P81 also emphasized: "*I didn't like the wording of the question, whether I feel Lichess has negative impact on my life. I think, Lichess is a great one of a kind resource with passionate developers behind*". Others felt similarly positive about online chess: "*I felt like there was a focus on online chess as a sort of escapist crutch, which is really funny for me since playing online (and exposing myself to embarassing failure as I learn) has been part of a process for me to try and overcome my own escapist crutches*" (P68).

These answers reinforce the conventional view of chess as an activity that improves well-being, and thus support our decision to measure the problematic side-effects of platform use as opposed to diagnose pathological usage patterns. Furthermore, they also support the view of Lichess as an open, non-profit platform designed purely for its end-users to enjoy chess. Tellingly, 69 of 131 participants opted to donate potential winnings to Lichess instead of receiving giftcards.

6 DISCUSSION

Self-Reported and Actual Time Measures. Existing research on the problematic effects of using online platforms falls into one of two categories. Studies either relate self-reported measures of

time spent online with perceived negative outcomes [17, 28, 38, 51, 57, 63], or they correlate actual measures from trace data with these outcomes instead [14, 37, 52]. In this work, we show that self-reported and actual time spent have distinct and complementary associations with negative offline outcomes, and thus there are potentially large benefits to incorporating both self-reported and actual time spent when assessing perceived problematic usage effects. We also demonstrate that time measures of varying granularity, such as maximum episodic time, can have additional explanatory power beyond the more commonly used metric of total usage time.

On its own, actual total playing time over the month-long period is loosely related to perceived problematic effects, with an effect size of -0.169 ($p < 0.05$). This is consistent with previous work [12, 14], suggesting that there is some basis for the common use of total time as a means of providing criterion validity in the psychology literature [28].

However, perceived maximum time spent in a day is much more strongly related to problematic outcomes, with an effect size of -0.275 ($p < 0.001$). This is especially surprising because it was the least accurately guessed numerical quantity by our participants (33% variance explained versus 51% for other temporal measures). Indeed, identifying one's maximum-usage day was the only categorical question on which our participants failed to outperform random guessing. This was despite the fact that we gave participants only two other choices: their median playing day, and one of their least playing days — often a day with no activity at all. Nonetheless, perceived maximum time, and not the delta between its self-reported and actual values, is strongly associated with perceived problematic effects. This suggests that the more participants felt their usage led to problematic outcomes, the longer they thought they played on their maximum usage day.

Most strikingly, perceived maximum time and actual total time, one a specific, self-reported quantity and the other an aggregate, data-derived quantity, contain complementary useful information with respect to predicting perceived negative effects. Actual total time and perceived maximum time in a day respectively explain 11.8% and 13.6% of the variance in problematic outcomes alone, but both of them together explain 23.2% of the variance. This suggests that the construct of time spent could be refined in the context of analyzing when online platforms use becomes unhealthy. For example, maximum episodic time might provide a better baseline for validating psychometric scales designed to identify at-risk individuals, most of which currently are evaluated against total time instead [28]. Furthermore, if self-reports were solely noisy proxies for actual data, one would not expect a combined model to significantly outperform models with only one type of time measure. Therefore, self-reported usage should not be regarded simply as imperfect proxies for data that can be replaced by server-logs or monitoring software [18]. Rather, self-reports, actual use, and negative outcomes have a complex relationship that will hopefully be elucidated by future work.

There are many plausible explanations for these findings, which we speculate on at present. For example, it may be that self-reported maximum time reflects *how* time was used by participants instead of their sheer usage amount, a distinction that is often mentioned but less frequently investigated in related work [12, 28]. Similarly, self-reported maximum time may also capture helpful information about participants' subjective experience of their time spent online, which is difficult to glean from data traces alone. Alternatively, self-reported time may be more distorted in those who reported more severe problematic effects, for instance due to confirmation bias in those undergoing negative life events. However, this is unlikely firstly because we found no effects from counterbalancing (administering the problematic effects scale first), and secondly because the differences between self-reports and data-derived metrics were not predictive of our survey score. The underlying reasons for the complementary nature of data and self-reports in predicting problematic effects thus merit further study.

Transparent Research into Problematic Usage Effects. We conducted this study with two considerations for advancing transparent research into online well-being. Firstly, a key difference between our work and previous research is our focus on an open, non-profit online platform. Lichess makes all of its game logs publicly available without reserving any rights, thus encouraging data exploration by anyone for any purpose. Combined with our recruitment of online participants through a public forum, these aspects allow for future work in this space to freely scrutinize, reproduce, or extend our study. Indeed, even Lichess's underlying source code can be inspected as it is completely open-source. We found no existing study into problematic usage effects that is transparent to this extent.

In contrast, prior work on problematic use is almost entirely focused on closed platforms with inaccessible data [6, 12, 14, 16, 24]. While follow-up research could focus on self-reports, studies that investigate actual, data-derived usage are restricted to affiliated researchers. This renders existing work opaque to the wider scientific community and limits how results can be interpreted. Furthermore, unaffiliated researchers studying data-derived usage have to develop bespoke applications (e.g. usage trackers [24, 37]) or manual methods (e.g. parsing screenshots [23]) to collect this data, which hinders the scalability and generalizability of both their methods and results. Similarly, research without access to the platforms may find participant pools to be small or expensive to recruit (e.g. compare [14] and [6]). We hope that our choice of analyzing Lichess users invites more open, data-supplemented studies into how using online platforms can lead both to problematic and to desirable offline outcomes.

Another unique property of studying a non-profit online platform is the absence of engagement-maximizing or monetization mechanisms [20]. The Lichess founders and contributors pride themselves on enriching the chess-playing community, which was tellingly reflected in the majority ($N = 69$) of participants who opted to donate their potential winnings to Lichess instead of receiving it themselves. Chess itself has historically been depicted more as an intellectual pursuit than a guilty pleasure [2, 5, 43]. This allowed us to study potentially problematic use in a setting where any indication of it is relatively likely to be organic and uninfluenced by algorithmic effects. Furthermore, there are few, if any, existing studies that address problematic outcomes arising from using online platforms that are not stereotypically "addictive" like Facebook or MMORPG games [6, 12, 14]. This is consistent with the responses defending Lichess as a prosocial platform that we received in our pilot. It would therefore seem that both experts and users alike would not consider Lichess as a problematic online service in and of itself. Thus, one might have guessed that users would have universally positive or neutral opinions about the platform's impact on their offline lives.

And yet, 45% of participants scored lower than 3 out of 5 on at least one scale question in our survey. This suggests that they felt non-neutral, negative effects of Lichess in at least one aspect of their lives, despite Lichess being a community-driven platform for an intellectual hobby. Furthermore, several open-ended responses indicated that some of our participants were consciously aware of the problematic effects of using Lichess. While the participant conveying the most negative feelings about Lichess also scored lowest on the scale (P89), others who mentioned negative sentiments scored within one standard deviation below the Phase B mean. It is possible that others scoring similarly to or lower than these participants may also feel negatively about the impact of online chess on their lives, but did not divulge this in the open-ended questions. Thus, we believe that platforms like Lichess should be studied further for potentially inducing problematic offline effects. For example, if a future comparative study were to find similar effect sizes in both Lichess and MMORPGs, it would suggest the possibility that explicit incentive mechanisms such as loot boxes [64] are not entirely responsible for negative outcomes.

The second decision we made for enhancing transparency is our use of SCA as an analytic framework [53]. We strove to present the exact way in which we operationalize constructs and made analytical decisions clearly in this work[7]. By using SCA, we also aim to minimize any biases we might otherwise have held in choosing model covariates, and considered the results of many possible specifications at once. This allows effect sizes to be robustly measured and contextualized [45]. For example, we show that the relationship between self-reported maximum time and problematic usage effects is statistically significant across every possible linear regression in our study. Furthermore, SCA ensures that our findings can be more closely scrutinized and extended in future work by following similar analytical procedures. Alongside our choice of an open platform, we hope this enables the broader scientific community to conduct scalable, data-driven work on problematic usage effects that would otherwise be harder to access.

7 LIMITATIONS AND FUTURE WORK

Although we strove to conduct this research at a high standard, our work has several limitations and opportunities for future work, which we outline here.

First, our survey responses were elicited from a non-representative sample. We obtained participants from the *r/chess* community on Reddit, who self-selected into participating in our study. Additionally, many of our survey responses were recorded in the first few hours after posting, suggesting that most respondents saw the advertisement soon after it was posted. This was reflected in the 51 participants who self-identified as living in the USA (followed by 12 for Germany) out of the 128 who disclosed their country. Also, the self-reported demographic information illustrates that practically all of our participants were young males, although since the vast majority of chess players are male we do not know how non-representative our sample is in this respect.

However, our objective was not to impute diagnoses of "addiction" on all Lichess users of different demographics (cf. 14). We instead demonstrate that a measurably diverse range of potential problematic effects are felt even within a specific subset of forum frequenters in the Lichess population. As participants' open-ended responses suggest, even moderate levels of problematic outcomes can have noticeable effects on well-being and are worth understanding in the context of time spent online. Forum recruitment is also a technique that has been used for similar studies [51].

Like most work in the area of problematic technology use, and most work linking surveys with data traces, our study is correlational. While we have found significant associations between both self-reported and actual measures of use time with problematic effects, identifying the causal effects between them remains a challenging open problem. For example, causality has been observed between actual aggregates use and problematicity [3, 23], but not more granular measures like maximum time. Further research is needed to clarify the causal link between self-reports, actual use time, and problematic effects.

One high-level concern about methodologies such as ours is the construct validity and comparability of survey scores that measure problematic usage effects. While we included survey items that might be neutrally and objectively answered, such as how often one has disrupted sleep to play online chess, the same aggregate Phase B scores for two participants may not imply that they feel negative outcomes to the same degree. We note that this limitation applies to almost all existing work on problematic use that proxies actual problematic outcomes via self-reports [14, 28, 34, 45, 51]. In this work, we have attempted to address this limitation by supplementing Phase B scores with a qualitative analysis of open-ended responses. For example, P89 expressed both the strongest negative view of Lichess and also obtained the lowest Phase B score. However, more rigorous

[7]Our analysis code will be open-sourced upon publication.

qualitative research is needed to better understand the subtleties of technology use (cf. 7) and how they are reflected by surveys assessing subjective experiences of problematicity.

Because we focused on the correlations between time measures and problematic usage effects, a natural extension of our present work is to combine self-reported and actual time measures for predicting and identifying problematic usage without psychometric scales. Full diagnostic questionnaires in the existing literature tend to be time-consuming, negatively phrased, and may bias participants in longitudinal or diary studies. Although some have attempted to predict [37, 52] or impute [14] problematic use diagnoses via behavioral data and without surveys, none use self-reported usage measures as features. These could be obtained with, for example, frequent self-reports of maximum episodic usage with a single short question. While not as scalable as the purely data-driven approaches, our work suggests that this approach would be more predictive of problematic effects, and it would be much easier to have participants answer short questions eliciting perceived time spent rather than lengthy psychometric scales.

Better automated predictions of problematic effects also hint at potential new directions in the design of digital self-regulation tools for online platforms [40]. For example, applications that moderate stimuli for at-risk users benefit from predicting when usage patterns may lead to problematic effects, before applying interventions [30]. This identification task may be made easier by eliciting self-reported time measures from users and combining them with data-derived measures, as our results suggest. Furthermore, responses to our open-ended questions indicate that gamified comparisons of self-reported and actual usage are fun activities for helping participants reflect on their playing habits. Incorporating activities like these may improve engagement in time visualization tools, many of which already track actual usage [47, 58]. Finally, while a litany of digital self-regulation aids have now been developed [40], many are from for-profit organizations who benefit from higher user engagement online[8]. Their large-scale empirical effectiveness also remains unknown in the existing literature, with evaluations often taking place in small samples. Therefore, open platforms and transparent methods like ours could be used to support open, reproducible, and transparent research on these aids.

8 CONCLUSION

As online platforms increasingly enable people across the globe to connect, share information, and enjoy socioeconomic benefits, critics have voiced growing concern over unhealthy patterns of technology use. In this paper, we seek to fill gaps in our current understanding of how time spent online relates to perceived problematic offline effects. Firstly, we show that both self-reported and actual usage measures are associated with problematic outcomes, despite the poor correlation between the two types of measures. Secondly, we find that the most predictive self-reported and actual measures contain complementary information about problematic effects. Additionally, we promote transparency in our study by examining an open, accessible online platform and using a robust statistical framework for our analysis. We hope that these findings will serve as a step towards a more complete picture of how and when use of online platforms can lead to problematic effects in users' lives.

REFERENCES

[1] 2019. *Lichess.org Press Kit.* Retrieved September 18, 2019 from https://lichess.cdn.prismic.io/lichess%2F758a8bb6-c1c5-4982-9c6b-6fcebeea424f_2018-03-04-lichess-press-kit.pdf

[2] 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.

[8]See for example: https://www.samsung.com/us/support/answer/ANS00083150, http://support.apple.com/HT208982, https://about.fb.com/news/2018/08/manage-your-time.

[3] Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. The welfare effects of social media. *American Economic Review* 110, 3 (2020), 629–76.

[4] American Psychiatric Association and others. 2013. Diagnostic and statistical manual of mental disorders. *BMC Med* 17 (2013), 133–137.

[5] Ashton Anderson and Etan A Green. 2018. Personal bests as reference points. *Proceedings of the National Academy of Sciences* 115, 8 (2018), 1772–1776.

[6] Cecilie Schou Andreassen, Torbjørn Torsheim, Geir Scott Brunborg, and Ståle Pallesen. 2012. Development of a Facebook addiction scale. *Psychological reports* 110, 2 (2012), 501–517.

[7] Julie H Aranda and Safia Baig. 2018. Toward JOMO: the joy of missing out and the freedom of disconnecting. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 19.

[8] Adriana Bianchi and James G Phillips. 2005. Psychological predictors of problem mobile phone use. *CyberPsychology & Behavior* 8, 1 (2005), 39–51.

[9] Jeffrey Boase and Rich Ling. 2013. Measuring mobile phone use: Self-report versus log data. *Journal of Computer-Mediated Communication* 18, 4 (2013), 508–519.

[10] Moira Burke, Cameron Marlow, and Thomas Lento. 2010. Social network activity and social well-being. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1909–1912.

[11] Neralie Cain and Michael Gradisar. 2010. Electronic media use and sleep in school-aged children and adolescents: A review. *Sleep medicine* 11, 8 (2010), 735–742.

[12] Scott Caplan, Dmitri Williams, and Nick Yee. 2009. Problematic Internet use and psychosocial well-being among MMO players. *Computers in human behavior* 25, 6 (2009), 1312–1319.

[13] Hilarie Cash, Cosette D Rae, Ann H Steel, and Alexander Winkler. 2012. Internet addiction: A brief summary of research and practice. *Current psychiatry reviews* 8, 4 (2012), 292–298.

[14] Justin Cheng, Moira Burke, and Elena Goetz Davis. 2019. Understanding Perceptions of Problematic Facebook Use: When People Experience Negative Life Impact and a Lack of Control. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 199.

[15] Robert S Davidson and Page B Walley. 1985. Computer fear and addiction: Analysis, prevention and possible modification. *Journal of Organizational Behavior Management* 6, 3-4 (1985), 37–52.

[16] Amandeep Dhir, Puneet Kaur, Sufen Chen, and Kirsti Lonka. 2016. Understanding online regret experience in Facebook use–Effects of brand participation, accessibility & problematic use. *Computers in Human Behavior* 59 (2016), 420–430.

[17] Tony Durkee, Michael Kaess, Vladimir Carli, Peter Parzer, Camilla Wasserman, Birgitta Floderus, Alan Apter, Judit Balazs, Shira Barzilay, Julio Bobes, et al. 2012. Prevalence of pathological internet use among adolescents in Europe: demographic and social factors. *Addiction* 107, 12 (2012), 2210–2222.

[18] Sindhu Kiranmai Ernala, Moira Burke, Alex Leavitt, and Nicole B Ellison. 2020. How well do people report time spent on Facebook? An evaluation of established survey questions with recommendations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[19] Isabela Granic, Adam Lobel, and Rutger CME Engels. 2014. The benefits of playing video games. *American psychologist* 69, 1 (2014), 66.

[20] Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. 2018. The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 534.

[21] Mark D Griffiths. 1991. Amusement machine playing in childhood and adolescence: A comparative analysis of video games and fruit machines. *Journal of adolescence* 14, 1 (1991), 53–73.

[22] Mark D Griffiths. 2000. Does Internet and computer" addiction" exist? Some case study evidence. *CyberPsychology and Behavior* 3, 2 (2000), 211–218.

[23] Melissa G Hunt, Rachel Marx, Courtney Lipson, and Jordyn Young. 2018. No more FOMO: Limiting social media decreases loneliness and depression. *Journal of Social and Clinical Psychology* 37, 10 (2018), 751–768.

[24] Reynol Junco. 2013. Comparing actual and self-reported measures of Facebook use. *Computers in Human Behavior* 29, 3 (2013), 626–631.

[25] Adam S Kahn, Rabindra Ratan, and Dmitri Williams. 2014. Why we distort in self-report: Predictors of self-report errors in video game play. *Journal of Computer-Mediated Communication* 19, 4 (2014), 1010–1023.

[26] Ronald C Kessler, Irving Hwang, Richard LaBrie, Maria Petukhova, Nancy A Sampson, Ken C Winters, and Howard J Shaffer. 2008. DSM-IV pathological gambling in the National Comorbidity Survey Replication. *Psychological medicine* 38, 9 (2008), 1351–1360.

[27] Dongil Kim, Yunhee Lee, Juyoung Lee, JeeEun Karin Nam, and Yeoju Chung. 2014. Development of Korean smartphone addiction proneness scale for youth. *PloS one* 9, 5 (2014), e97920.

[28] Daniel L King, Maria C Haagsma, Paul H Delfabbro, Michael Gradisar, and Mark D Griffiths. 2013. Toward a consensus definition of pathological video-gaming: A systematic review of psychometric assessment tools. *Clinical psychology*

review 33, 3 (2013), 331–342.
[29] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 453–456.
[30] Geza Kovacs, Zhengxuan Wu, and Michael S Bernstein. 2018. Rotating Online Behavior Change Interventions Increases Effectiveness But Also Increases Attrition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 95.
[31] Daria J Kuss and Mark D Griffiths. 2011. Online social networking and addiction – a review of the psychological literature. *International journal of environmental research and public health* 8, 9 (2011), 3528–3552.
[32] Daria J Kuss and Mark D Griffiths. 2012. Internet gaming addiction: A systematic review of empirical research. *International Journal of Mental Health and Addiction* 10, 2 (2012), 278–296.
[33] Daria J Kuss, Mark D Griffiths, Laurent Karila, and Jöel Billieux. 2014. Internet addiction: A systematic review of epidemiological research for the last decade. *Current pharmaceutical design* 20, 25 (2014), 4026–4052.
[34] Min Kwon, Dai-Jin Kim, Hyun Cho, and Soo Yang. 2013. The smartphone addiction scale: development and validation of a short version for adolescents. *PloS one* 8, 12 (2013), e83558.
[35] Min Kwon, Joon-Yeop Lee, Wang-Youn Won, Jae-Woo Park, Jung-Ah Min, Changtae Hahn, Xinyu Gu, Ji-Hye Choi, and Dai-Jin Kim. 2013. Development and validation of a smartphone addiction scale (SAS). *PloS one* 8, 2 (2013), e56936.
[36] Robert LaRose, Carolyn A Lin, and Matthew S Eastin. 2003. Unregulated Internet usage: Addiction, habit, or deficient self-regulation? *Media Psychology* 5, 3 (2003), 225–253.
[37] Uichin Lee, Joonwon Lee, Minsam Ko, Changhun Lee, Yuhwan Kim, Subin Yang, Koji Yatani, Gahgene Gweon, Kyong-Mee Chung, and Junehwa Song. 2014. Hooked on smartphones: an exploratory study on smartphone overuse among college students. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2327–2336.
[38] Jeroen S Lemmens, Patti M Valkenburg, and Jochen Peter. 2009. Development and validation of a game addiction scale for adolescents. *Media psychology* 12, 1 (2009), 77–95.
[39] Shao-Kang Lo, Chih-Chien Wang, and Wenchang Fang. 2005. Physical interpersonal relationships and social anxiety among online game players. *Cyberpsychology & behavior* 8, 1 (2005), 15–20.
[40] Ulrik Lyngs, Kai Lukoff, Petr Slovak, Reuben Binns, Adam Slack, Michael Inzlicht, Max Van Kleek, and Nigel Shadbolt. 2019. Self-Control in Cyberspace: Applying Dual Systems Theory to a Review of Digital Self-Control Tools. *arXiv preprint arXiv:1902.00157* (2019).
[41] Robert McIlwraith, Robin Smith Jacobvitz, Robert Kubey, and Alison Alexander. 1991. Television addiction: Theories and data behind the ubiquitous metaphor. *American Behavioral Scientist* 35, 2 (1991), 104–121.
[42] G-J Meerkerk, Regina JJM Van Den Eijnden, Ad A Vermulst, and Henk FL Garretsen. 2009. The compulsive internet use scale (CIUS): some psychometric properties. *Cyberpsychology & behavior* 12, 1 (2009), 1–6.
[43] Monty Newborn. 2012. *Kasparov versus Deep Blue: Computer chess comes of age*. Springer Science & Business Media.
[44] Amy Orben, Tobias Dienlin, and Andrew K Przybylski. 2019. Social media's enduring effect on adolescent life satisfaction. *Proceedings of the National Academy of Sciences* 116, 21 (2019), 10226–10228.
[45] Amy Orben and Andrew K Przybylski. 2019. The association between adolescent well-being and digital technology use. *Nature Human Behaviour* 3, 2 (2019), 173.
[46] Andrew K Przybylski and Netta Weinstein. 2017. A large-scale test of the Goldilocks Hypothesis: Quantifying the relations between digital-screen use and the mental well-being of adolescents. *Psychological Science* 28, 2 (2017), 204–215.
[47] John Rooksby, Parvin Asadzadeh, Mattias Rost, Alistair Morrison, and Matthew Chalmers. 2016. Personal tracking of screen time on digital devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 284–296.
[48] Tracii Ryan, Andrea Chester, John Reece, and Sophia Xenos. 2014. The uses and abuses of Facebook: A review of Facebook addiction.
[49] Matthew J Salganik. 2017. *Bit by bit: social research in the digital age*. Princeton University Press.
[50] Michael Scharkow. 2016. The accuracy of self-reported internet useâĂŤA validation study using client log data. *Communication Methods and Measures* 10, 1 (2016), 13–27.
[51] A Fleming Seay and Robert E Kraut. 2007. Project massive: Self-regulation and problematic use of online gaming. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 829–838.
[52] Choonsung Shin and Anind K Dey. 2013. Automatically detecting problematic use of smartphones. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 335–344.
[53] Uri Simonsohn, Joseph P Simmons, and Leif D Nelson. 2015. Specification curve: Descriptive and inferential statistics on all reasonable specifications. *Available at SSRN 2694998* (2015).
[54] Ricardo A Tejeiro Salguero and Rosa M Bersabé Morán. 2002. Measuring problem video game playing in adolescents. *Addiction* 97, 12 (2002), 1601–1606.

[55] Naomi J Thomas and Frances Heritage Martin. 2010. Video-arcade game, computer game and Internet activities of Australian students: Participation habits and prevalence of addiction. *Australian Journal of Psychology* 62, 2 (2010), 59–66.

[56] Jean M Twenge, Thomas E Joiner, Megan L Rogers, and Gabrielle N Martin. 2018. Increases in depressive symptoms, suicide-related outcomes, and suicide rates among US adolescents after 2010 and links to increased new media screen time. *Clinical Psychological Science* 6, 1 (2018), 3–17.

[57] Antonius J Van Rooij, Tim M Schoenmakers, Ad A Vermulst, Regina JJM Van Den Eijnden, and Dike Van De Mheen. 2011. Online video game addiction: identification of addicted adolescent gamers. *addiction* 106, 1 (2011), 205–212.

[58] Steve Whittaker, Vaiva Kalnikaite, Victoria Hollis, and Andrew Guydish. 2016. 'Don't Waste My Time': Use of Time Information Improves Focus. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1729–1738.

[59] Dmitri Williams, Mia Consalvo, Scott Caplan, and Nick Yee. 2009. Looking for gender: Gender roles and behaviors among online gamers. *Journal of communication* 59, 4 (2009), 700–725.

[60] Peter M Yellowlees and Shayna Marks. 2007. Problematic Internet use or Internet addiction? *Computers in human behavior* 23, 3 (2007), 1447–1453.

[61] Kimberly S Young. 1996. Psychology of computer use: XL. Addictive use of the Internet: a case that breaks the stereotype. *Psychological reports* 79, 3 (1996), 899–902.

[62] Kimberly S Young. 1998. Internet addiction: The emergence of a new clinical disorder. *Cyberpsychology & behavior* 1, 3 (1998), 237–244.

[63] Kimberly S Young and Carl J Case. 2004. Internet abuse in the workplace: new trends in risk management. *CyberPsychology & Behavior* 7, 1 (2004), 105–111.

[64] David Zendle and Paul Cairns. 2018. Video game loot boxes are linked to problem gambling: Results of a large-scale survey. *PloS one* 13, 11 (2018), e0206767.