Referee Report on *Digital Addiction* by Allcott, Gentzkow, and Song (2022)

Summary

The main research question addressed by Allcott et al. (2022) is how habit formation and self-control problems contribute to digital addiction, particularly in smartphone social media use. By developing an economic model of digital addiction and employing a randomized experiment, the authors aim to quantify these behavioural factors and distinguish their effects. Specifically, they test whether temporary incentives for reduced screen time create lasting behavioural changes - evidence of habit formation - and whether self-imposed app limits effectively reduce usage, reflecting self-control problems. Ultimately, their research seeks to clarify the nature and scale of digital addiction to better inform users, tech companies, and regulators.

This question is critically important because smartphone and social media use have become pervasive behaviours, widely perceived as harmful and potentially addictive, akin to traditional addictions such as smoking or gambling. These technologies significantly affect subjective well-being, influencing individual and parental decisions about managing screen time. Insights into habit formation and self-control issues can inform the development of effective digital self-control tools, enable technology companies to better align their products with consumer welfare, and guide policymakers evaluating regulations like the Social Media Addiction Reduction Technology (SMART) Act, ultimately addressing substantial economic welfare concerns arising from digital overuse.

The paper's primary empirical method is a randomized controlled trial (RCT) involving approximately 2,000 Android smartphone users who installed the Phone Dashboard app. Participants were independently randomized into two treatments and corresponding control groups: a "bonus" treatment offering temporary financial incentives ($50 per reduced daily hour) to limit smartphone use on FITSBY apps (Facebook, Instagram, Twitter, Snapchat, browsers, and YouTube), and a "limit" treatment providing functionality to set daily app usage limits. Usage data were continuously tracked, supplemented by four surveys at three-week intervals. The analysis utilized a balanced panel (1,933 participants), directly estimating treatment effects on app usage, self-reported addiction, ideal screen-time changes, and subjective well-being, followed by a structural model estimation. This report will primarily focus on the RCT, rather than the structural model.

Main Results

The authors provide empirical evidence supporting their main conclusion that habit formation and self-control problems significantly drive digital addiction. Specifically, Figure 4 illustrates that during the incentive period (Period 3), the bonus treatment substantially reduced participants' daily FITSBY use by approximately 56 minutes per day, representing a 39% decrease compared to the control group. Importantly, this reduction persisted into periods 4 and 5, even after removing financial incentives, with sustained reductions of about 19 and 12 minutes per day, respectively. This persistence demonstrates habit formation rather than temporary behavioural change. Similarly, the limit treatment consistently reduced FITSBY usage by roughly 22 minutes per day (16%) across periods 2 to 5, with minimal attenuation (still 19 minutes per day reduction at the end), reflecting ongoing self-control challenges among users rather than confusion or novelty effects. Thus, Figure 4 highlights how both temporary incentives and self-control tools effectively address key behavioural mechanisms underlying digital addiction. Given the substantial and persistent reductions, the authors' interpretation of meaningful behavioural change is appropriate.

Figure 8 and Table A6 also provide empirical support for the authors' conclusions. The bonus treatment, intended to disrupt habit formation by temporarily incentivising reduced screen time, significantly decreased participants' desire for further screen-time reductions ("ideal use change") by 0.41 SD, compared to 0.23 SD for the limit treatment. Both treatments also significantly reduced self-reported addiction scores (main addiction scale and SMS addiction scale), consistent with disrupted habitual usage patterns observed earlier in Figure 4. Meanwhile, the limit treatment, designed to address self-control problems by enabling participants to impose restrictions, uniquely increased participants' perceptions that smartphone use enhanced their lives. This likely arose because participants felt greater control and alignment between actual and ideal smartphone usage. Unlike the externally-driven financial incentive of the bonus, the limit functionality promoted intrinsic motivation and agency, leading users to view their smartphone use more positively.

Regarding subjective well-being (SWB), both interventions showed positive effects, though only the bonus treatment yielded a statistically significant improvement (0.09 SD, $p \approx 0.026$). This improvement primarily resulted from enhanced concentration and reduced distraction rather than reductions in anxiety, which individually did not significantly change in either treatment group. However, Anxiety was measured as part of the SWB index, consistent with

psychological literature suggesting a possible link between anxiety-related personality traits (such as neuroticism) and addictive behaviours on platforms like Facebook (Andreassen et al., 2012). These SWB results closely align with Allcott et al. (2020), who found similar improvements (0.09 SD) after Facebook deactivation. Nonetheless, the modest magnitude and partial insignificance of SWB outcomes suggest some caution in generalising the interventions' broader welfare implications.

When evaluating potential cherry-picking or selective reporting, the authors' transparency is particularly notable. They report deviations from their Pre-Analysis Plan (PAP), clearly stating why changes were made and demonstrating that their results remain robust across alternative specifications. Moreover, their discussion of the limitations regarding their instrumental variables (IV) analysis provides additional evidence against selective reporting. Specifically, the authors acknowledge that their IV approach, used to estimate Local Average Treatment Effects (LATE), might not fully satisfy the exclusion restriction - particularly for the limit treatment, where effects could operate through alternative channels like increased feelings of personal control. By openly presenting limitations, despite them not necessarily strengthening their conclusions, the authors demonstrate methodological rigor and transparency.

Nevertheless, several limitations could affect the generalisability of the authors' conclusions. For instance, the sample's limited representativeness - consisting predominantly of younger, educated, female Android users - raises caution about generalising the large and sustained reductions in app usage seen in Figure 4 and Table A6 to broader populations. While the authors partially address this by using sample weights in their structural model estimation (Online Appendix Table A13) to better approximate national demographics (income, education, gender, race, and age), this reweighting notably increases the modelled effect of self-control problems, suggesting that the original findings may be conservative relative to the general population. Additionally, the study was conducted during the early COVID-19 pandemic, a context of increased smartphone use and altered daily routines, possibly amplifying participants' responsiveness to the interventions, and raising concerns about whether the substantial behavioural changes observed would persist under more typical circumstances. These limitations affect external validity and imply careful interpretation and caution in broadly generalising the results reported in the figures and tables to typical, more diverse contexts.

Identifying Assumptions

The primary identifying assumption underlying the RCT conducted by Allcott et al. (2022) is that random assignment successfully created statistically equivalent treatment and control groups at baseline. This assumption is fundamental because it enables any observed differences in outcomes - such as reductions in smartphone usage (FITSBY apps), subjective well-being improvements, or reductions in self-reported addiction - to be causally attributed to the interventions themselves (the screen-time limits or monetary bonus incentives).

The authors assessed the effectiveness of randomization through extensive balance checks presented in Online Appendix Table A2. The checks confirm strong baseline equivalence across treatment and control groups, with no statistically significant differences in key observable covariates - including income, educational attainment, gender, race, age, and baseline FITSBY usage (p-values ranging from 0.13 to 0.76). Joint tests for overall covariate balance further reinforce these findings (p-values of 0.65 for Limit and 0.94 for Bonus). Thus, the robust statistical checks support internal validity, confirming successful random assignment into comparable groups.

However, it's important to recognise that participants installed the Phone Dashboard app before the baseline period, potentially affecting their awareness of smartphone usage and consequently influencing their baseline FITSBY use. Despite this, because such awareness uniformly affected both treatment and control groups, relative comparisons and estimated treatment effects remain valid. The authors also convincingly argue that sustained use of limit-setting tools throughout the 12-week study period, coupled with consistent supporting evidence from independent bonus valuations and strong correlations between limit-setting behaviour and other self-control measures, reflects genuine self-control motivations rather than mere reactivity.

Another essential component supporting the internal validity of the study is the impressively low attrition rate (approximately 5.6%) over the 12-week intervention period. Attrition was carefully monitored and balanced across treatment conditions, as detailed in Online Appendix Table A1. The authors' careful experimental design - frequent reminders, structured payment schedules, and explicit communication stressing participants' importance in completing the study - effectively mitigated attrition, thereby reducing the risk of selection bias stemming from differential dropout rates.

Although randomisation successfully created balanced groups at baseline, frequent use of the 'snooze' functionality allowed participants to temporarily bypass limits, reducing the actual usage reductions observed (22 minutes/day) relative to intended reductions (53 minutes/day). The authors transparently report this issue, and sustained participant engagement with limit-setting (78% through the final period) indicates genuine, albeit imperfect, compliance. Thus, while noncompliance slightly complicates interpretation, it does not substantially undermine the authors' main causal claims.

The possibility of spillover effects is another critical consideration affecting internal validity. The authors addressed these by examining substitution patterns. Notably, the bonus treatment did not induce significant substitution toward other smartphone apps, with confidence intervals ruling out large substitution relative to the observed reductions (main paper Figure 5). Conversely, the limit treatment led to measurable substitution - about 12 minutes/day - to apps not targeted by limits. Additional analyses of device substitution (Online Appendix Figure A14), based on self-reported data, showed small but clearly differentiated spillovers: limit participants slightly increased FITSBY usage on other devices, whereas bonus participants slightly reduced their usage. While substitution effects introduce complexity, their explicit documentation enhances transparency and enables cautious interpretation of the observed treatment effectiveness.

Referee Decision

Overall, I find the empirical results presented by the authors convincing and credible. They have transparently acknowledged limitations and conducted extensive robustness checks, thoroughly addressing potential internal validity concerns, as previously discussed. This methodological rigor significantly enhances confidence in their findings.

Given the extensive existing robustness analysis, additional checks are challenging. However, a useful extension could involve examining participants' smartphone usage data from before the Phone Dashboard app was installed. Most modern smartphones automatically track screen-time data independently of additional apps. Extracting this pre-intervention data would provide a more accurate baseline measure of habitual usage, further enhancing the balance of observables across treatment and control groups. Incorporating such data as a covariate in baseline equivalence tests would clarify whether app installation influenced participants' behaviour at the start of the study. Additionally, analysing this more precise

baseline could offer nuanced insights into the true magnitude of the treatment effects by accurately capturing participants' typical smartphone usage prior to any intervention.

Future empirical work could beneficially address some limitations directly. Firstly, replicating the experiment in a non-pandemic context would clarify whether the observed substantial and sustained behavioural responses hold under more typical circumstances. Secondly, extending the analysis to include non-Android (particularly iOS) users would further test external validity. Evidence from Shaw et al. (2016) suggests significant psychological differences between Android and iOS users, particularly in emotionality which can involve traits like anxiety, which are directly relevant to subjective well-being measures and smartphone addiction behaviours examined by Allcott et al. (2022). Investigating the differences would provide additional insights into the generalizability of the interventions and their broader applicability across user populations.

Building on the earlier discussion of substitution effects, a valuable avenue for future research would be systematically examining why the bonus and limit treatments induced distinct substitution patterns. The differing responses suggest that financial incentives (bonus) might directly weaken the underlying desire or habit associated with specific digital content, whereas self-imposed limits could primarily redirect digital engagement toward unrestricted apps or devices. Further investigation could clarify whether monetary rewards broadly diminish digital engagement or if limits merely shift usage patterns without addressing the root of problematic behaviours. Understanding these mechanisms would significantly inform the design of more effective and targeted interventions.

In conclusion, given the comprehensive robustness checks, transparency, and openness about limitations, I recommend this paper for publication, noting that future research addressing these suggestions would further enhance the robustness and generalisability of these findings.

References

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The Welfare Effects of Social Media. *American Economic Review*, *110*(3), 629–676. https://doi.org/10.1257/aer.20190658

Allcott, H., Gentzkow, M., & Song, L. (2022). Digital Addiction. *American Economic Review*, *112*(7), 2424–2463. https://doi.org/10.1257/aer.20210867

Andreassen, C. S., Torsheim, T., Brunborg, G. S., & Pallesen, S. (2012). Development of a Facebook Addiction Scale. *Psychological Reports*, *110*(2), 501–517. https://doi.org/10.2466/02.09.18.pr0.110.2.501-517

Shaw, H., Ellis, D. A., Kendrick, L.-R., Ziegler, F., & Wiseman, R. (2016). Predicting Smartphone Operating System from Personality and Individual Differences. *Cyberpsychology, Behavior, and Social Networking*, *19*(12), 727–732. https://doi.org/10.1089/cyber.2016.0324