Byrnes, Alex

Working Paper

Reanalysis of Sanders et al. (2024): An Umbrella Review of the Benefits and Risks Associated with Youths' Interactions with Electronic Screens

I4R Discussion Paper Series, No. 154

Provided in Cooperation with:
The Institute for Replication (I4R)

INSTITUTE for REPLICATION

No. 154
I4R DISCUSSION PAPER SERIES

Reanalysis of Sanders et al. (2024): An Umbrella Review of the Benefits and Risks Associated with Youths' Interactions with Electronic Screens

Alex Byrnes

September 2024

I4R DISCUSSION PAPER SERIES

I4R DP No. 154

Reanalysis of Sanders et al. (2024): An Umbrella Review of the Benefits and Risks Associated with Youths' Interactions with Electronic Screens

Alex Byrnes[1]

[1]*University of Leeds/Great Britain*

SEPTEMBER 2024

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

Reanalysis of Sanders et al. (2024): An umbrella review of the benefits and risks associated with youths' interactions with electronic screens*

Alex Byrnes

May 23, 2024

Abstract

Sanders et al. (2024) made the central claim that effects found in eight meta-analyses are "strong evidence" (P<0.001) for various health and educational outcomes in children associated with use of electronic screens ("screen time"). The eight effects highlighted as strong evidence ranged in size from r = –0.14 to 0.33. Although some of the primary studies were experimental and some observational, and the individual claims of strong evidence were causal, the design of the analysis – the umbrella review – does not pool data from meta-analyses. The authors converted effects in the screen time literature to a common measurement (r) and summarized the meta-analyses that survived exclusion. Therefore, the paper as a whole made a descriptive claim that this evidence exists in the literature and, in some cases, it is strong evidence.

Sanders et al. (2024) excluded meta-analyses when they found significant publication bias using Egger's and excess significance tests. The remaining effect sizes were not corrected for publication bias. This robustness replication calculated the same eight effects using a technique to correct for publication bias (PET-PEESE) and attempted to find coding, mathematical, data, and reporting errors.

This analysis found publication bias reduced the effect size in three of the eight meta-analyses to such a degree that these findings failed to replicate. It also found a pattern of results indicative of p-hacking in the screen time literature, and evidence that a more moderate interpretation of the data could have been presented in Sanders et al. had the authors chosen a different set of eight – or the full set – of high-certainty effects.

*Author: Alex Byrnes: University of Leeds. E-mail: od23a2b@leeds.ac.uk. The author declares no conflicts of interest or financial support. This paper is prepared as part of a collaboration between the Institute for Replication and Nature Human Behaviour (Brodeur et al. (2024))

1 Introduction

Sanders et al. (2024) is a prospectively-registered umbrella review (Papatheodorou and Evangelou (2022)) on the effect of screen time in children. The umbrella review methodology seeks to summarize literature on a topic by choosing a population, set of independent variables, and outcomes that fit within broad categories. In the case of Sanders et al., the population was children, the exposures were types of screen time (e.g. video games, television, and social media) and the outcomes were associated with education and health (e.g. literacy, and body composition). They excluded meta-analyses based on standard criteria: duplication, appropriateness, measures reported, whether or not a larger study was available, among other criteria, and then summarized the final body of evidence. The way the papers are described and the methods that are used to arrive at the final set of effects constitute the central claim in an umbrella review, and in Sanders et al.

The authors used the National Health, Lung and Blood Institute's Quality Assessment of Systematic Reviews and Meta-Analyses tool (NIH (2014)), which recommends assessment of, but does not mention correction for publication bias. This is summarized in the tool as "Reviewers assessed and clearly described the likelihood of publication bias." Sanders et al. state that they "did not assess risk of bias in the individual studies that were included in each meta-analysis." In this robustness replication, and prior to seeing the source code or data in "Data availability" and "Code availability", the authors' statement was interpreted as suggesting that they did not correct for selection bias by the original meta-analysis authors, nor publication bias, other than to exclude meta-analyses from the final set presented[1].

Use of uncorrected effect sizes has led to low rates of robustness replication in the screen time literature. In Hilgard et al. (2017), a reanalysis of video games and violent tendencies, the original effect sizes were within the reanalyzed confidence

[1]Additionally, the authors included the PRISMA 2020 checklist (Page et al. (2021)), which recommends that users "Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases)." The checklist also does not mention correction, a particular method, or otherwise limit the assessment methods.

interval in 77% to 85% of meta-analyses on the association between video games and violent tendencies, depending on the correction method[2]. A later re-analysis[3] of Hilgard et al. (2017) agreed that publication bias moderated effect sizes but disagreed with its conclusion that the effects were inconsequential (Kepes et al. (2017)).)

Inflation of magnitude and statistical certainty is prevalent in published research due to publication bias. In a large study of meta-analyses in Ecology and Evolution (Yang et al. (2023)) 66% of meta-analytic effects would no longer be significant if corrected for publication bias. In psychology, publication bias was found to have inflated magnitude of effects by 50% and statistical certainty by 60% (Bartoš et al. (2023)).

The method used in Sanders et al., that is, eliminating meta-analyses with significant publication bias, assumes publication bias exists in part of the publication system or it does not, or at least that we can judge its salience by the significance threshold of these tests. Publication bias and the file drawer effect (part of "reporting bias") can reasonably be assumed to exist in almost all published literature (Fanelli (2011)). It will likely affect meta-analytical r statistics to some degree, whether significant or not. Although sensitive to heterogeneity, p-hacking, whether or not the effect exists, its size, and other factors (Van Aert et al. (2016), Simonsohn et al. (2014b), Carter et al. (2019)), methods for correcting effects sizes for publication bias such as p-curve, puniform, PET-PEESE, and RoBMA have been around for several years. These methods at least attempt to account for publication bias in meta-analyses.

The results show bias introduced by the study design, the underlying literature, the standards followed, and by the authors themselves. This report found no coding, mathematical, data, or reporting errors, however, and the materials were expertly prepared and transparent.

[2]A mean replication rate of 79% across analyses that include confidence intervals, compared to a 34% replication rate in Yang et al. (2023).

[3]or re-re-analysis of Anderson et al. (2010)

1.1 Threshold for replication

The registration for Sanders et al., as disclosed in its methods section, only included the analysis, that is the search terms, exclusion criteria, and other methodology. There was no hypothesis given in the registration or the paper. It is difficult, therefore, to falsify a particular set of statements (Anvari and Lakens (2021)) unless there are errors in the calculations, exclusions, or search terms.

The interpretation of effect size in psychology in general, and screen time studies in particular further complicates this effort. In the screen time literature, small effect sizes are called both "the indispensable foundation for a cumulative psychological science" (Götz et al. (2021)) and not interpretable as supporting the hypothesis (Ferguson and Heene (2021)). Screen time has been described as nearly equivalent to eating potatoes (Orben and Przybylski (2019)) and "more strongly linked to happiness among girls than selling drugs" (Twenge et al. (2022)).

The Open Science Collaboration 2015 (Open Science Collaboration (2015)) used three methods to measure reproducibility: subjective assessment, p-values, and confidence intervals. Since confidence intervals produced the highest replication rate ($\approx 47\%$) and greater objectivity, this reanalysis will test whether or not the recalculated 95% confidence interval contain the r values in Sanders et al. The meta-analyses highlighted as strong evidence additionally introduce a significance threshold ($\alpha = 0.001$) that is lower than the threshold generally used in replication studies to signify publishability. Therefore, the findings could be unpublishable because they are inaccurate, and yet publishable because they still have a p-value less than 0.05.

Sanders et al. did not pool the eight studies highlighted. These studies contain heterogeneous methods, exposures, and outcomes and so a pooled replication threshold is difficult to justify (Harrer et al. (2021)).

The lack of hypothesis, and the umbrella review design make interpretation of the replication of the whole paper ambiguous. However, the metascientific literature provides ample basis for assessing individual meta-analyses that will, at minimum,

allow comparison with previous replication efforts, and to the degree that this analysis is plausibly preregistered, the threshold should be considered non-arbitrary and not motivated by the results (see Methods for precise definitions).

The selection of effects in the literature to reanalyze may be a source of bias (Simonsohn et al. (2014b)). The selection here is motivated by the reproduction of Sanders et al. and the intention is not to add new point estimates to the screen time debate, but to give Sanders et al. its due replication effort and to test the analytical flexibility in its design.

1.2 Methods

The threshold for reproducibility has been criticized as unstandardized (Devezer et al. (2021), Schauer and Hedges (2021)). Although replication, reproducibility, and replication robustness might necessitate different standards – assuming a standard is possible – similar factors contribute to the difficulty for all of these falsification efforts, for instance, arbitrary thresholds for publication of the original work (often alpha levels) and infrequent use of self-imposed criteria for falsification such as the smallest effect size of interest (SESOI).

The publication bias correction method that has the best performance in a simulation with the expected parameters (high heterogeneity, moderate average meta-analysis size, moderate publication bias, and moderate levels of questionable research practices) is PET-PEESE (Carter et al. (2019)). The technique also has precedent in Hilgard et al. (2017), is relatively well-established, comparable to Egger's test used in Sanders et al. (Harrer et al. (2021)) and it has the advantage of simplicity and ease of interpretation. The implementation comprises less than one hundred lines of R code (Bartoš et al. (2022)).

1.2.1 Hypothesis This robustness replication tested the following hypothesis:

Publication bias exists in almost any field and any literature review. The statistical choice in Sanders et al. to exclude meta-analyses with significantly-high

publication bias likely left meta-analyses with some degree of bias. Some[4] *of the eight highlighted studies (Table 1) will fail to replicate. That is, the corrected 95% confidence interval will not contain the original effect size using PET-PEESE based on the implementation in Bartoš et al. (2022) and available as an R script (https://codeocean.com/capsule/8141708/tree/v1).*

Additionally, this reanalysis sought to uncover:

1. Major coding and mathematical errors.
2. Apparent flaws in the data provided.
3. Undisclosed discrepancies with the preregistered methodology.

Any analyses beyond these, if there are any, will be marked as exploratory and described separately, and any deviations from this protocol reported.

1.3 Planned Analysis

Three of the eight highlighted studies (Table 1) failed to replicate. In all of these cases, the estimate from Sanders et al. had an absolute value higher than the corrected confidence interval, suggesting publication bias remained in the included studies. In two cases, the corrected estimate was larger than the estimate in Sanders et al. However, the mean effect size absolute value was reduced from 0.2 to 0.16. The replication rate (63%) was lower but comparable to the 77% to 85% rate in Hilgard et al. (2017).

1.4 Exploratory Analysis

The unplanned analysis in this replication focused on the studies that met the criteria for high certainty chosen by Sanders et al., of which eight were highlighted in the paper. The three criteria were: non-significant Egger's and excess significance

[4] This is only intended to specify the hypothesis, not as a threshold for the whole paper's replicability.

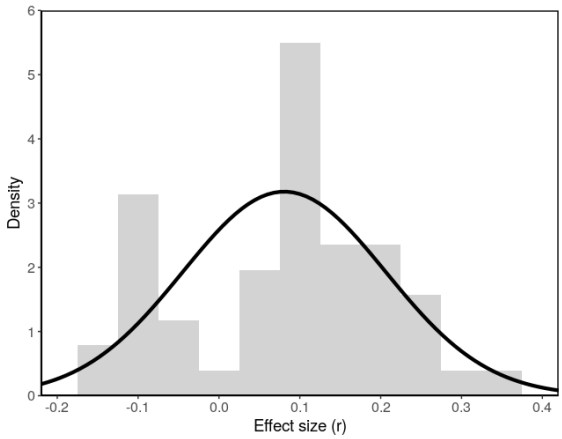

Figure 1: High-certainty effects (r) in Sanders et al. compared to a normal distribution (bin width 0.05).

tests, and sample size greater than 1,000. According to the analysis code, there were 51 of these meta-analyses.[5]

The distribution of effect sizes (Figure 1) in the 51 high-certainty meta-analyses is consistent with a pattern found in p-values in the presence of p-hacking. This pattern is characterized by a gap in the distribution of outcome measurements around a value necessary for publication and a greater density at more desirable values nearby. Usually, the critical value is the traditional p-value threshold 0.05. In this case, the meta-analyses were chosen using thresholds the original meta-analysis authors couldn't precisely predict. However, the effect sizes were known and may be biased away from zero, that is, meta-analysts and the primary researchers may prefer to publish non-zero effects.

The distribution of 51 high-certainty effects shows a clear gap at zero-effect, and an artificially-high density around -0.1 and 0.1. This pattern is thought to be indicative of questionable research practices (QRPs), specifically re-analyzing the data until a publishable result is found (Simonsohn et al. (2014a), Gelman and Loken (2013), Simonsohn et al. (2020)).

Furthermore, the mean effect of these studies is 0.08, which is consistent with the hypothesis that screen time is associated with a small effect and studies of screen

[5]The authors correctly pointed out in their response that there was an error in the original draft describing the high-certainty criteria as including an ordinary p-value. The p-values in the "high certainty" filter were Egger's and excess significance p-values. This is further explained in the addendum to this section.

time have some unavoidable measurement and sampling error.

Random-effects meta-analysis assumes a normal distribution of effect sizes with total variance that is the sum of the variance due to measurement and sampling error (v_i) and tau-squared (τ^2), the true heterogeneity of effects due to differences in study design ($\epsilon_i \sim N(0, v_i)$; $Var(y_i) = v_i + \tau^2$) (Veroniki et al. (2016)). The use of τ^2 is an assumption, and its size must be estimated from the sample (Borenstein et al. (2010)). By choosing eight effects to highlight and stating, for instance,"We recommend that caregivers and policymakers carefully weigh the evidence for potential harms and benefits of specific types of screen use" the authors are assuming between-study heterogeneity is causing true heterogeneity in effect size (tau-squared). It could be that this assumption is at least partially wrong, the true effect heterogeneity is small, and the observed variance is primarily due to measurement and sampling error.

With this assumption, similar to a fix-effect meta-analysis in which tau is zero, these effects are drawn from a single normal distribution, and the specific types of screen time are all different ways of measuring the same small effect.

These conclusions may be more apparent, or attractive, to researchers with a strong prior belief in the prevalence of publication bias and QRPs. However, the omission of this evidence from Sanders et al. is striking. The eight highlighted studies are among the strongest effects in the high-certainty set and the authors portrayed *distinct* results, that some types of screen time are associated with positive outcomes and some negative. This interpretation may be supportable and it errs on the side of caution for childrens' health and education but the evidence in the full 51 meta-analyses is less alarming and – consistent with the planned replication (1.3) – suggests overstatement of effects in the literature (Hilgard et al. (2017), Bartoš et al. (2023), Yang et al. (2023)).

1.4.1 Addendum to Exploratory Analysis The authors' response points out that effects didn't need to have low p-values to be considered "high certainty," as stated in the first version of this report. This is correct and the text has been amended ac-

cordingly in section 1.4. This didn't affect the conclusion that the eight highlighted effects were chosen from low p-values because the Sanders et al. used the p-value filter later in the process. The high certainty effects themselves (Figure 1) were not filtered for low p-value and can be considered *more* representative of the effect of screen time, although with due skepticism covered in section 1.3 and elsewhere.

For verification of the filter calculation, it can be confirmed without code using Sanders' Supplementary file 2 and filtering with the "high certainty" criteria in a spreadsheet program: *reanalysis_n* >= 1000, *reanalysis_eggers_p* > 0.05, and *reanalysis_tes_p* > 0.05.

1.5 Limitations

The number of effect-size-correcting meta-analytic methodologies (PET-PEESE, p-curve, puniform, RoBMA, and others) and limitations and disagreements found in each (Van Aert et al. (2019), Carter et al. (2019)) suggests that the science of publication bias correction may still be imprecise, or difficult to perfectly justify.

PET-PEESE has high bias when the number of primary studies is low ($k < 20$) or variance due to heterogeneity is high ($I^2 > 80\%$) (Stanley (2017)). However, the Type I error rate for the original random effect meta-analysis under the same conditions is 87% or higher with reporting bias of 50% (Stanley (2017)). The eight highlighted meta-analyses in Sanders et al. have lower I^2 or higher k than these thresholds, other than two of the three effects from Vannucci et al. (2020) in which $k = 14$ and $I^2 = 96\%$. These have k only slightly greater than the cutoff for Egger's test (10). Given the effect sizes of $r = 0.19$ and 0.21, the Type I error rate for both the original estimate and Egger's test for publication bias may be unacceptably high already, and so the correction will be as difficult to interpret as the original.

Correcting the two Vannucci et al. (2020) studies with $k < 20$ also introduces some statistical flexibility since these studies could be dropped or included, or declared likely false positives with or without the correction.

There is no known precedent for replicating an umbrella review and there is

flexibility in the choice of analysis: re-running the literature search and exclusions, eliminating studies based on a different test for publication bias, or testing the P<0.001 threshold claims for significance in the same direction. The analysis chosen here doesn't depend on subjectively reevaluating meta-analyses for inclusion or exclusion and it is focused on the calculations in an umbrella design. It is not exhaustive.

PET-PEESE is similar to Egger's test and so the results will depend on the strength of using small-study effects as a proxy for publication bias. The same regression line is used in both and either the slope, or the y-intercept represents the degree of publication bias (Harrer et al. (2021)). Under moderate conditions the two tend to agree. Additionally, high between-study heterogeneity[6] entails large confidence intervals in the reanalysis. Therefore, Sanders et al. was more likely to replicate than it would have with a less restrictive registration. The benefit to this trade-off is greater isolation of the specific statistical choice of uncorrected effect sizes.

Future literature may help decide to what degree these ambiguities can be eliminated with replication methodology. Some meta-analysts have found contradiction in trusted methods and doubt any can correct publication bias post hoc (Van Elk et al. (2015)).

Despite these cautions and limitations to interpretation, Sanders et al. is, as a whole paper, descriptive and so its publication was presumably subject to a inexact threshold. It is appropriate and unavoidable therefore that its reproduction is imperfect and this analysis should be interpreted with caveats similar to the original paper.

1.6 Discussion

1.6.1 Replication of umbrella reviews As of 2018, publication of umbrella reviews was increasing exponentially (Papatheodorou (2019)). Although there is some

[6]The statistic reported (I^2) in Sanders et al. is the *percentage of variation due to heterogeneity*, not the total heterogeneity, which can be derived from τ^2 (Borenstein (2023))

uniformity to the design, authors disagree on how much evidentiary value umbrella reviews have, and on how to carry them out. Relatedly, they disagree on eliminating primary study overlap, and whether or not to assess quality of the individual meta-analyses at all, (Gianfredi et al. (2022)). (Overlap and quality were both addressed in Sanders et al.)

It could be argued that the claims in an umbrella review are solely those of the original meta-analysis authors and summary statistics are *cited*, not presented as original estimates. There is some merit to this argument. However, umbrella reviews are generally presented as evidentiary with the usual considerations of "confounding, reverse causality, selection bias, and information bias" are important to their interpretation (Belbasis et al. (2022)). Caregivers and policymakers in particular may have trouble understanding this distinction or its statistical properties.

In the case of Sanders et al., publication bias was addressed and so it is a fair topic for reanalysis separate from the interpretation concerns of umbrella reviews.

1.6.2 Quality and heterogeneity Only seven percent[7] of meta-analyses had low risk of bias on all criteria evaluated (NIH (2014)) in Sanders et al. and the authors expressed the need for "larger, high-quality studies" in screen time research. Additionally, the heterogeneity and sensitivity to reanalysis in the field is well-documented (Orben (2020), Twenge and Campbell (2019)), even by researchers who have come to very different conclusions about screen use, and despite accounting for the multiplicity of statistical specifications (Orben and Przybylski (2019), Twenge et al. (2022)). New research is free to avoid acknowledging the uncertainty due to analytical flexibility, or offer estimates that claim to correct past mistakes. It is potentially fruitful, therefore, to use preregistration, data and code transparency, and to reduce the impact of publication pressures in lending weight to future estimates (Van Elk et al. (2015), Hilgard et al. (2017), Yang et al. (2023)).

The threshold for noteworthy effect sizes is debatable in research on any topic,

[7] The authors disclosed excluding "Eligibility criteria predefined and specified" from this count. Including the eligibility criterion gives 3%.

and the threshold should be lower for more consequential ones. The effect of screen time on children is certainly one of these topics. However, the results of Sanders et al. underscores the need for preregistration of the smallest effect size of interest (SESOI) (Anvari and Lakens (2021)) or other "minimal important differences" (King (2011)).

In the interest of reducing publication bias, the findings in Sanders et al. should be published whether significant or not, or with an effect smaller than the SESOI. Preregistering the SESOI would remove some ambiguity from the umbrella review design. Are these effects simply the most extreme ones in the literature? Or are they a "positive" finding in the authors' estimation *a priori*? Furthermore, in the case of public health concerns potentially affecting a large portion of the global population, that – as Sanders et al. note – is widely believed to be is harmful, a null result may be as noteworthy as a positive one.

1.6.3 Acknowledgements The author wishes to acknowledge Dr. Manon Ragonnet for help with editing and clarity.

1.6.4 Code and data availability The analysis, exploratory analysis, and data are available on Github (`https://github.com/alexbyrnes/sanders_2024_replication`).

References

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., Rothstein, H. R. and Saleem, M.: 2010, Violent video game effects on aggression, empathy, and prosocial behavior in Eastern and Western countries: A meta-analytic review., *Psychological Bulletin* **136**(2), 151–173.
URL: *https://doi.org/10.1037/a0018251*

Anvari, F. and Lakens, D.: 2021, Using anchor-based methods to determine the smallest effect size of interest, *Journal of Experimental Social Psychology* **96**, 104159.
URL: *https://www.sciencedirect.com/science/article/pii/S0022103121000627*

Bartoš, F., Maier, M., Quintana, D. S. and Wagenmakers, E.-J.: 2022, Adjusting for publication bias in jasp and r: Selection models, pet-peese, and robust bayesian meta-analysis, *Advances in Methods and Practices in Psychological Science* **5**(3), 25152459221109259.

Bartoš, F., Maier, M., Shanks, D. R., Stanley, T., Sladekova, M. and Wagenmakers, E.-J.: 2023, Meta-analyses in psychology often overestimate evidence for and size of effects, *Royal Society Open Science* **10**(7), 230224.

Belbasis, L., Bellou, V. and Ioannidis, J. P.: 2022, Conducting umbrella reviews, *BMJ medicine* **1**(1).

Borenstein, M.: 2023, How to understand and report heterogeneity in a meta-analysis: The difference between i-squared and prediction intervals, *Integrative Medicine Research* p. 101014.

Borenstein, M., Hedges, L. V., Higgins, J. P. and Rothstein, H. R.: 2010, A basic introduction to fixed-effect and random-effects models for meta-analysis, *Research synthesis methods* **1**(2), 97–111.

Brodeur, A., Dreber, A., Hoces de la Guardia, F. and Miguel, E.: 2024, Reproduction and replication at scale, *Nature Human Behaviour* **8**(1), 2–3.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M. and Hilgard, J.: 2019, Correcting for bias in psychology: A comparison of meta-analytic methods, *Advances in Methods and Practices in Psychological Science* **2**(2), 115–144.

Devezer, B., Navarro, D. J., Vandekerckhove, J. and Buzbas, E. O.: 2021, The case for formal methodology in scientific reform, *Royal Society Open Science* **8**(3).
URL: *https://doi.org/10.1098/rsos.200805*

Fanelli, D.: 2011, Negative results are disappearing from most disciplines and countries, *Scientometrics* **90**(3), 891–904.
URL: *https://doi.org/10.1007/s11192-011-0494-7*

Ferguson, C. J. and Heene, M.: 2021, Providing a lower-bound estimate for psychology's "crud factor": The case of aggression., *Professional Psychology: Research and Practice* **52**(6), 620–626.
URL: *https://doi.org/10.1037/pro0000386*

Gelman, A. and Loken, E.: 2013, The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time, *Department of Statistics, Columbia University* **348**(1-17), 3.

Gianfredi, V., Nucci, D., Amerio, A., Signorelli, C., Odone, A. and Dinu, M.: 2022, What can we expect from an umbrella review?, *Advances in Nutrition* **13**(2), 684.

Götz, F. M., Gosling, S. D. and Rentfrow, P. J.: 2021, Small Effects: the indispensable foundation for a cumulative psychological science, *Perspectives on Psychological Science* **17**(1), 205–215.
URL: *https://doi.org/10.1177/1745691620984483*

Harrer, M., Cuijpers, P., A, F. T. and Ebert, D. D.: 2021, *Doing Meta-Analysis With R: A Hands-On Guide*, 1st edn, Chapman & Hall/CRC Press, Boca Raton, FL and London.

Hilgard, J., Engelhardt, C. R. and Rouder, J. N.: 2017, Overstated evidence for short-term effects of violent games on affect and behavior: A reanalysis of Anderson et al. (2010)., *Psychological Bulletin* **143**(7), 757–774.
URL: *https://doi.org/10.1037/bul0000074*

Kepes, S., Bushman, B. J. and Anderson, C. A.: 2017, Violent video game effects remain a societal concern: Reply to hilgard, engelhardt, and rouder (2017).

King, M. T.: 2011, A point of minimal important difference (mid): a critique of terminology and methods, *Expert review of pharmacoeconomics & outcomes research* **11**(2), 171–184.

NIH: 2014, Study quality assessment tools| nhlbi, *NIH. nd https://www. nhlbi. nih. gov/health-topics/study-quality-assessment-tools.[Accessed 2 July 2021]. accessed* .

Open Science Collaboration, T.: 2015, Estimating the reproducibility of psychological science, *Science* **349**(6251).
URL: *https://doi.org/10.1126/science.aac4716*

Orben, A.: 2020, Teenagers, screens and social media: a narrative review of reviews and key studies, *Social psychiatry and psychiatric epidemiology* **55**(4), 407–414.

Orben, A. and Przybylski, A. K.: 2019, The association between adolescent well-being and digital technology use, *Nature Human Behaviour* **3**(2), 173–182.
URL: *https://doi.org/10.1038/s41562-018-0506-1*

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T., Mulrow, C. D., Shamseer, L., Tetzlaff, J., Akl, E. A., Brennan, S., Chou, R., Glanville, J., Grimshaw, J., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E., Mayo-Wilson,

E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V., Whiting, P. and Moher, D.: 2021, The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, *The BMJ* p. n71.
URL: *https://doi.org/10.1136/bmj.n71*

Papatheodorou, S.: 2019, Umbrella reviews: what they are and why we need them, *European journal of epidemiology* **34**, 543–546.

Papatheodorou, S. I. and Evangelou, E.: 2022, Umbrella reviews: what they are and why we need them, *Meta-Research: Methods and Protocols* pp. 135–146.

Sanders, T., Noetel, M., Parker, P., Del Pozo Cruz, B., Biddle, S., Ronto, R., Hulteen, R., Parker, R., Thomas, G., De Cocker, K. et al.: 2024, An umbrella review of the benefits and risks associated with youths' interactions with electronic screens, *Nature Human Behaviour* **8**(1), 82–99.

Schauer, J. M. and Hedges, L. V.: 2021, Reconsidering statistical methods for assessing replication., *Psychological Methods* **26**(1), 127–139.
URL: *https://doi.org/10.1037/met0000302*

Simonsohn, U., Nelson, L. D. and Simmons, J. P.: 2014a, P-curve: a key to the file-drawer., *Journal of experimental psychology: General* **143**(2), 534.

Simonsohn, U., Nelson, L. and Simmons, J. P.: 2014b, P-Curve and effect size, *Perspectives on Psychological Science* **9**(6), 666–681.
URL: *https://doi.org/10.1177/1745691614553988*

Simonsohn, U., Simmons, J. P. and Nelson, L. D.: 2020, Specification curve analysis, *Nature Human Behaviour* **4**(11), 1208–1214.

Stanley, T. D.: 2017, Limitations of pet-peese and other meta-analysis methods, *Social Psychological and Personality Science* **8**(5), 581–591.

Twenge, J. M. and Campbell, W. K.: 2019, Media use is linked to lower psychological well-being: Evidence from three datasets, *Psychiatric Quarterly* **90**, 311–331.

Twenge, J. M., Haidt, J., Lozano, J. and Cummins, K.: 2022, Specification curve analysis shows that social media use is linked to poor mental health, especially among girls, *Acta Psychologica* **224**, 103512.
URL: *https://www.sciencedirect.com/science/article/pii/S0001691822000270*

Van Aert, R. C. M., Wicherts, J. M. and Van Assen, M. A.: 2016, Conducting Meta-Analyses Based on p Values, *Perspectives on Psychological Science* **11**(5), 713–729.
URL: *https://doi.org/10.1177/1745691616650874*

Van Aert, R. C., Wicherts, J. M. and Van Assen, M. A.: 2019, Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis, *PloS one* **14**(4), e0215052.

Van Elk, M., Matzke, D., Gronau, Q. F., Guan, M., Vandekerckhove, J. and Wagenmakers, E.-J.: 2015, Meta-analyses are no substitute for registered replications: A skeptical perspective on religious priming, *Frontiers in psychology* **6**, 1365.

Vannucci, A., Simpson, E. G., Gagnon, S. and Ohannessian, C. M.: 2020, Social media use and risky behaviors in adolescents: A meta-analysis, *Journal of Adolescence* **79**(1), 258–274.
URL: *https://pubmed.ncbi.nlm.nih.gov/32018149/*

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D. and Salanti, G.: 2016, Methods to estimate the between-study variance and its uncertainty in meta-analysis, *Research synthesis methods* **7**(1), 55–79.

Yang, Y., Sánchez-Tójar, A., O'Dea, R. E., Noble, D. W., Koricheva, J., Jennions, M. D., Parker, T. H., Lagisz, M. and Nakagawa, S.: 2023, Publication bias impacts on effect size, statistical power, and magnitude (type m) and sign (type s) errors in ecology and evolutionary biology, *BMC biology* **21**(1), 71.

2 Tables

Table 1: Effects supported by strong evidence (P<0.001)

Outcome	**Specific outcome**	**Exposure**	**Lead author, date**	r **with corrected 95% CI**	**Replicated**
Literacy	General	Screen use: general	Madigan, 2020	–0.14 [-0.09, 0.03]	No
Body composition	Body composition	TV programs and movies: general	Marshall, 2004	0.06 [-0.06, 0.07]	Yes
Psychological health	Depression	Internet use: general	Shin, 2022	0.25 [0.15, 0.20]	No
Risky behavior	Risk taking (general)	Social media: general	Vannucci, 2020	0.21 [0.14, 0.26]	Yes
Risky behavior	Risky sexual behavior	Social media: general	Vannucci, 2020	0.21 [-0.08, 0.16]	No
Risky behavior	Substance abuse	Social media: general	Vannucci, 2020	0.19 [0.15, 0.33]	Yes
Learning	General	Education (touch screen)	Xie, 2018	0.21 [0.16, 0.31]	Yes
Learning	General	Augmented reality	Tekedere,2016	0.33 [0.24, 0.52]	Yes

Table 2: Papers highlighted as "strong evidence," outcome, exposure, effect sizes, and confidence interval. Partially reproduced from Sanders et al. (2024).