Figure #	Figure title One sentence only	Filename This should be the name the file is saved as when it is uploaded to our system. Please include the file extension. i.e.: *Smith_ED_Fig1.jpg*	Figure Legend If you are citing a reference for the first time in these legends, please include all new references in the main text Methods References section, and carry on the numbering from the main References section of the paper. If your paper does not have a Methods section, include all new references at the end of the main Reference list.
Extended Data Fig. 1	Reporting accuracy post hoc moderator and subgroup analyses	Extended Data Figure 1.tiff	*Note:* k: number of separate effect sizes included for the moderator level; *R* = response ratio; *Exp(β)* = exponential transformation of metaregression coefficient from a model in which a categorical moderator with two levels was entered as a predictor. *F* values correspond to the Approximate Hotelling-Zhang with small sample correction omnibus tests for moderators with more than two levels; 95% CI corresponds to the *r* values for individual moderator levels; *p* corresponds to the *F* value for moderators or the subgroup analysis for individual moderator levels. * This analysis did not include the adolescent population category, the general population category and the unknown population category as only two, one, and three effect sizes were available, respectively.
Extended Data Fig. 2	Descriptive statistics for additional post hoc moderator analyses	Extended Data Figure 2.tiff	*Note:* k: number of included effect sizes. *: One study used both a built-in tool and a third-party tool
Extended	Digital media	Extended Data Figure	*Note:* k: number of separate effect sizes included for the moderator level; *r* = Pearson correlation coefficient; *F* values correspond to the

Data Fig. 3	usage post hoc moderator and subgroup analyses	3.tiff	Approximate Hotelling-Zhang with small sample correction omnibus tests for moderators with more than two levels; 95% CI corresponds to the *r* values for individual moderator levels; *p* corresponds to the *F* value for moderators or the subgroup analysis for individual moderator levels. * This analysis did not include the adolescent population group as only two effect sizes were available. † This analysis did not include the other category as only a single effect size was available.

1

Item	Present?	Filename This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. The extension must be .pdf	A brief, numerical description of file contents. i.e.: *Supplementary Figures 1-4, Supplementary Discussion, and Supplementary Tables 1-4.*
Supplementary Information	Yes	Supplementary_Information.pdf	Supplementary tables 1 & 2, Supplementary analyses.
Reporting Summary	Yes	nr-reporting-summary.pdf	

A Systematic Review and Meta-Analysis of Discrepancies Between Logged and Self-Reported Digital Media Use

Douglas A. Parry[1*], Brittany I. Davidson[2, 3], Craig J. R. Sewall[4], Jacob T. Fisher[5], Hannah Mieczkowski[6], Daniel S. Quintana[7, 8, 9]

[1]Department of Information Science, Stellenbosch University, South Africa
[2]School of Management, University of Bath, UK
[3]Faculty of Engineering, University of Bristol, UK
[4]School of Social Work, University of Pittsburgh, USA
[5]College of Media, University of Illinois Urbana-Champaign, USA
[6]Department of Communication, Stanford University, USA
[7]NORMENT, Center for Psychosis Research, Oslo University Hospital and University of Oslo, Norway
[8]KG Jebsen Centre for Neurodevelopmental Disorders, University of Oslo, Oslo, Norway
[9]Department of Psychology, University of Oslo, Oslo, Norway

* Corresponding author: Douglas A. Parry (dougaparry@sun.ac.za; 0000-0002-6443-3425)

Abstract

There is widespread public and academic interest in understanding the uses and effects of digital media. Scholars primarily use self-report measures of the quantity or duration of media use as proxies for more objective measures, but the validity of these self-reports remains unclear. Advancements in data collection techniques have produced a collection of studies indexing both self-reported and log-based measures. To assess the alignment between these measures, we conducted a preregistered meta- analysis of this research. Based on 106 effect sizes, we found that self-reported media use only moderately correlates with logged measurements, that self-reports were rarely an accurate reflection of logged media use, and that measures of problematic media use show an even smaller association with usage logs. These findings raise concerns about the validity of findings relying solely on self-reported measures of media use.

The widespread adoption of digital media technologies has generated substantial public and academic interest in understanding the diverse uses and effects that these media enable. Across almost all areas of social science research, whether researchers are studying digital media use in the context of persuasion, personal well-being, productivity, anxiety, aggression, or other physical, psychosocial or political phenomena, technology (or media) use is frequently adopted as a key predictor or outcome variable. A particularly vivid example is the debate around the impacts of digital media use on psychosocial well-being[1]. Some scholars conclude that media use has "destroyed a generation"[2], while others decry these claims, suggesting that current concern is merely this generation's manifestation of a "Sisyphean Cycle of Technology Panics"[3].

Progress towards resolving these debates and developing a deeper understanding of the role of media use in human behaviour requires "transparent and robust analytical practices" [4], but also confidence that the measures that are adopted to assess use of digital media are valid indicators of actual usage patterns [5, 6]. Before conclusions can be made about media use and the effects thereof, we must first trust not only the theoretical models posed in studies, but perhaps more importantly, the measures used to produce data to test these models. The validity of media use measures is central to the validity of empirical research on media uses and effects [5]. While media use is inherently an observable behaviour, despite longstanding criticisms of the accuracy and validity of media use self-report measures [7—12], the majority of research treats media use as a latent variable, with scholars typically relying on retrospective self-report measures to quantify various forms of media use [13—15].

These self-report measures typically index either the time spent using all media (i.e., 'screen-time'), the time spent using specific media, or the frequency or volume of total or specific media use [16]. In many cases, rather than focusing on use of a particular medium (e.g., a specific social networking service), measures concern the use of metamedia (e.g., a smartphone or the Internet) that themselves contain a "multitude of constituent media" (e.g., various social networking services or instant messaging applications) [17]. Responses are typically collected in the form of single-point estimates or Likert-type scales. In addition, despite concerns about construct validity and measurement validation procedures [18—20], researchers frequently use self-report measures of problematic media use (including excessive usage among other conceptualisations) to make claims about the drivers and outcomes of media use itself [19, 21—23].

A substantial body of psychometric research demonstrates that self-reported measurement of behaviour can be highly unreliable, with participant responses being prone to cognitive, social, and communicative biases [24—27]. Schwarz and Oyserman [26] argue that "even apparently simple behavioural questions pose complex cognitive tasks" for participants. In addition to question comprehension—which has been shown to impact response accuracy with changes in item-wording, formatting, or order impacting outcomes [26, 28, 29]—accurate recall of behaviour is also affected by various cognitive limitations in autobiographical memory [26, 30]. These limitations are particularly apparent for behaviours that are frequent and that are highly integrated into respondents' lives [24, 26, 30]. This makes them difficult to accurately distinguish and retrieve. Self-reports of behaviour are, consequently, an index of what respondents believe that they do—their perceptions of their own behaviour—and not necessarily what they actually do [5, 31].

Accurate estimation of media use is affected not only by these well-established factors that affect survey-response behaviour [24, 26, 27], but also by the fact that use of media is likely to be especially difficult to report accurately. Typically, people use multiple media simultaneously (e.g., using Facebook while listening to music or checking emails) and embed media use alongside other non-media activities (e.g., sports, face-to-face socialising), which creates a difficulty disentangling specific behaviors. Furthermore, media use frequently consists of numerous micro-interactions [32] further blurring the distinction between media and non-media activities [33]. Therefore, given known difficulties estimating frequent behaviours that are highly integrated into respondents' lives [24], media use is likely to be particularly difficult to recall and to accurately estimate without suitable measures that can help guide unbiased responses. Consequently, the validity of self-report measures of media use is likely biased not only by well-known factors that impact the accuracy of self-reports of behaviour, but also by the difficulty of the estimation task itself.

Over the preceding decade, adoption of "data-intensive" approaches for measuring media use has accelerated. In parallel with general developments in personal analytics have come tools that enable researchers to directly measure complete device use, network or call traffic, or even the use of specific applications and services [13, 34, 35]. These developments have led to a number of investigations considering associations between self-reported and logged media use. Early research showed that, for calling and texting on mobile phones, self-reports correlate only moderately with network provider logs [36, 37]. Comparisons between digital trace data of Internet use and self-reported use have indicated similarly moderate correlations[5]. Recently, Ellis et al. [21] compared responses for ten scales and three single estimates for either general or problematic use of smartphones with relevant tracking data. While all self-report measures positively correlated with device use, effect sizes were small—a pattern that seems to hold across a number of studies [5, 32, 36, 37].

These data suggest that self-reported and logged measures, rather than simply serving as different ways to measure media use, may in fact capture distinct constructs [31, 38]. Log-based techniques, although they are not without their own biases and shortcomings [5, 35, 39, 40], provide a more direct, and likely more accurate measure of media use than self-report [5, 21, 32, 41]. As such, there exists a need to systematically assess whether self-reported media use is an accurate indicator of actual usage patterns. To address this knowledge gap, we conducted a pre-registered systematic review and meta-analysis of research wherein both self-reported and logged media use were assessed. Additionally, we assessed whether individuals tend to under- or over-report their media use, and whether these outcomes depend on various media, methodological, or participant-related characteristics.

Results

After describing the included studies, we consider correlations between self-reported and logged measures of digital media use. This is followed by an analysis of potential moderating factors in this analysis. In the next section, we investigate correlations between logged usage and self-reports of problematic use. Finally, we consider the degree to which self-reports are either under- or over-reported relative to logged data. Unless otherwise indicated, all analyses were pre-registered [42]. All materials needed to reproduce the results are available through the Open Science Framework (https://osf.io/dhx48/).

Included effect sizes

The initial search produced 12,132 results. After screening for eligibility (see Figure 1), 47 records were included in the final sample, with 45 either published or available as preprints [5, 21, 31, 32, 36—39, 41, 43—78] and two included on the basis of unpublished raw data received directly from the authors (Burnell et al., unpublished manuscript; Geyer et al. unpublished manuscript). From these records, 106 effect sizes were included in the analyses. Supplementary Table 1 provides a summary of the included effect sizes for measures concerning digital media use and Supplementary Table 2 provides a summary for measures of problematic use.

To evaluate the association between self-reported and logged media use, 66 effect sizes from 44 studies were considered. Across these comparisons the total sample size is 52,007. On average, a comparison involved 787.99 participants (SD = 1,621.27, median = 166, min = 20, max = 6,598). In a second, separate meta-analysis, we investigated associations between self-reported problematic use and logged measures of use. This analysis included 40 effect sizes from 19 studies, with a total sample size of N = 5,552. On average, a comparison involved 138.8 participants (SD = 92.79, median = 139.5, min = 14, max = 294). Finally, to assess whether individuals tend to systematically under- or over-report their media use, we included 49 comparisons from 30 studies and a total sample size of N = 17,523, with an average sample size of 357.61 participants (SD = 955.62, median = 159, min = 20, max = 6,598).

Acknowledging general shortcomings of study quality assessment in systematic reviews [79—81], using the quality of survey studies in psychology (Q-SSP) checklist [82], we classified a majority of included papers as acceptable in quality (55.56%), with the remainder considered lower in quality. The mean quality score (out of 100) is 66.60 (SD = 10.78). Notably, while the Q-SSP includes 20 items, scores for five items (sample size justification; measurement description; information about the person(s) collecting the data; information about the context of data collection; and the relation between the discussion and the population of interest) primarily accounted for lower quality ratings. Overall, given the exploratory nature of many studies in our sample, while there is room for improvement, we consider the quality of evidence to be acceptable for our syntheses.

Correlations between self-reported and logged media use

The correlation between self-reported and logged measures of digital media use was calculated with robust variance estimation (RVE), revealing a relationship that was positive, but only medium in magnitude (r = 0.38, 95% CI [0.33, 0.42], p < 0.001) given conventional effect size interpretations. Figure 2 depicts a forest plot of the effect sizes included in this analysis. Egger's regression test (incorporating

RVE per the Egger Sandwich test)[83], indicated no evidence of small study bias in this sample (β = 0.55, p = 0.136); see Panel A in Figure 3 for a contour-enhanced funnel plot.

Influence diagnostics, performed with the metafor package[84], indicated a single outlier in this sample[51] (n = 45, r = 0.87). A sensitivity analysis excluding this outlier produced a summary effect size that was almost the same as the original analysis (r = 0.37, 95% CI [0.33, 0.42], p < 0.001). Similarly, a sensitivity analysis excluding the only effect size that was extracted using the web plot digitiser tool[53] showed a comparable effect size to the original analysis (r = 0.38, 95% CI [0.34, 0.42], p < 0.001). In a final sensitivity analysis, we considered whether the results presented in peer-reviewed studies differed from non-peer reviewed studies. Of the 66 included effect sizes, 10 (15.15%) were non-peer-reviewed at the time of inclusion (see Supplementary Table 1). While the effect size is larger in peer-reviewed (r = 0.39, 95% CI [0.34, 0.44], p < 0.001, k = 56) than in non-peer-reviewed (r = 0.31, 95% CI [0.21, 0.41], p < 0.001, k = 10) effects, the difference is not statistically significant (β = -0.08, 95% CI [-0.21, 0.04], p = 0.164).

The impact of moderators on the correlational effect size

There was a high level of heterogeneity in the included effect sizes ($Q(63) = 734.89$, p < 0.001; with RVE: $T^2 = 0.012$, $I^2 = 92.18\%$) for the correlation between self-reported and logged media use. Therefore, following our protocol, three moderator analyses were conducted to attempt to identify possible sources of heterogeneity. While sufficient data were available for self-report form (Scale: k = 6; Estimate: k = 60) and self-report category (Duration: k = 47; Volume: k = 19), only two levels for medium (Phone: k = 49; Social media: k = 13) met our requirements, with the three remaining levels holding insufficient observations (Internet: k = 2; Games: k = 1; Computer: k = 1). Therefore, deviating from our analysis plan, we only considered effect sizes for studies investigating use of phones or social media in the moderator analysis for medium.

Table 1 summarises the results of the three moderator analyses as well as the subgroup analyses for each moderator level considered. For medium type, because we only included a sub-sample of effect sizes, we first calculated a summary effect size for studies targeting use of a phone or social media and found it to be comparable to the overall correlation (r = 0.37, 95% CI [0.32, 0.42], p < 0.001). As is evident in Table 1, while the correlation is smaller for social media than for phones, this difference was not statistically significant. Similarly, for self-report form, while the small number of studies using scales (k = 6) impacts interpretability, we found that the difference in the magnitude of the association between scales and single estimates was not statistically significant. Finally, we found no evidence that the association between self-reported and logged measures of media use differs between measures concerning either the duration or the volume of use.

Four additional post hoc moderator analyses (described in full in the Method section) were conducted to further explore possible sources of heterogeneity. Given currently available data, no evidence was found that the association between self-reported and logged measures of media use differs by population ($F(3, 6.57) = 0.42$, p = 0.745), data collection design ($F(2, 21.2) = 0.90$, p = 0.423), nor the logging method adopted ($F(3, 16.9) = 1.4$, p = 0.279). Extended Data Figure 1 provides a summary of the subgroup analyses for each moderator level included in these analyses. Finally, a single post hoc, multiple-moderator model was produced to account for potential confounds among the three original, pre-specified moderators (medium, measure type, and self-report form). An omnibus test using the Approximate Hotelling-Zhang test provided no evidence for a moderating effect ($F(5, 10.1) = 0.457$, $p = 0.718$), with comparable results for medium (β = -0.03, 95% CI [-0.16, 0.10], $p = 0.663$), measure type (β = -0.01, 95% CI [-0.15, 0.12], $p = 0.842$) and self-report form (β = 0.15, 95% CI [-0.17, 0.44], $p = 0.278$). Additionally, heterogeneity remained high ($T^2 = 0.015$, $I^2 = 89.78\%$).

Correlations between problematic and logged media usage

The correlation between self-reported problematic use and logged use (calculated with RVE) was positive, but small ($r = 0.25$, 95% CI [0.20, 0.29], $p < 0.001$), with a low level of heterogeneity ($Q(41) = 60.21$, $p = 0.016$; with RVE: $T^2 = 0.004$, $I^2 = 29.41\%$). Figure 4 presents a forest plot for this analysis. Egger's regression test (incorporating RVE) [83], indicated no evidence of small study bias (β = 0.34, $p = 0.246$; see Panel B in Figure 3 for a contour-enhanced funnel plot). Influence diagnostics did not reveal any outliers. However, because five included effects were reported in non-peer-reviewed studies, we considered whether this influenced the outcome. For peer-reviewed studies the correlation was estimated with RVE while, for non-peer-reviewed studies, there were insufficient observations so a random-effects intercept-only model was calculated. No meaningful difference was observed between peer-reviewed ($r = 0.25$, 95% CI [0.19, 0.31], $p < 0.001$, $k = 35$) and non-peer-reviewed ($r = 0.25$, 95% CI [0.15, 0.34], $p < 0.001$, $k = 5$) effects ($Q_b(1) = 0.01$, $p = 0.973$).

Accuracy of self-report measures

Of the 49 included comparisons, only three (6.12%) mean self-reported media use estimates fell within 5% of the logged mean. Despite this, similar proportions of studies reported mean self-reports of media use that were either over- ($k = 23$, 46.94%) or under- ($k = 23$, 46.94%) reported relative to the logged measure. To produce a summary effect size, we calculated the weighted ratio of means (incorporating RVE after log transformation) between self-reported and logged measures of media use and found that, across studies, participants over-reported their media use (R = 1.21, 95% CI [0.94, 1.54], $p = 0.129$). However, given that the confidence interval for this result includes indicator values for under-reported and accurately reported media use, the evidence is insufficient to conclude whether estimates are typically under- or over-reported compared to logs of media use. Figure 5 provides a forest plot for the effects included in this analysis.

Egger's regression test (incorporating RVE) [83] showed no evidence of small study bias (β = 0.62, $p = 0.41$; see Panel C in Figure 3 for a contour-enhanced funnel plot). Influence diagnostics indicated a single outlier [51] ($n = 45$, $r = 0.87$, self-report mean = 73 minutes, self-report SD = 59, logged mean = 4 minutes, SD = 6; R = 18.25, 5% CI [14.05, 23.71]). A sensitivity analysis excluding this outlier produced a summary effect size that was similar to the original analysis (R = 1.18, 95% CI [0.95, 1.48], $p = 0.136$). Of the 49 effects, nine (18.37%) were non-peer-reviewed at the time of inclusion (see Supplementary Table 1). A sensitivity analysis excluding these studies found no statistically significant difference between peer-reviewed (R = 1.30, 95% CI [0.97, 1.75], $p = 0.075$) and non-peer-reviewed (R = 0.89, 95% CI [0.57, 1.40],

p = 0.543) effects (β = -0.367, *Exp*(β) = 0.69, 95% CI [0.41, 1.16], p = 0.133). A second sensitivity analysis excluding two effects that were included after using the web plot digitiser [49, 55] showed comparable results to the overall analysis (R = 1.21, 95% CI [0.94, 1.56], p = 0.141).

Moderators of reporting accuracy

There was a high-level of heterogeneity in the sample ($Q(48) = 7254.71$, $p < 0.001$; with RVE: $T^2 = 0.32$, $I^2 = 99.50\%$). Two moderator analyses were planned a priori to investigate possible sources of heterogeneity. For medium, only two levels (Phone: k = 41; Social Media: k = 5) held sufficient data, with too few observations reported for the remaining levels (Internet: k = 1; Games: k = 1; Computer: k = 1). For the self-report category, there was sufficient data for measures of duration (k = 35) and volume (k = 14). For the type of medium, as is evident in Table 2, the summary effect size for studies including both self-report and logged measures of phone use was comparable to the overall analysis. For social media, while the effect size indicates a higher degree of over-reporting, the Satterthwaite degrees of freedom for the model were less than 4, indicating a high probability of a Type I error. Consequently, for medium type, no moderator analysis was conducted. For self-report category, while measures of duration showed a larger degree of over-reporting compared to measures of volume which indicated under-reporting, the difference was not statistically significant (β = -0.44, *Exp*(β), = 0.64, 95% CI [0.41, 1.02], p = 0.056).

Four additional post hoc moderator analyses (described in full in the Method section) were conducted to further explore possible sources of heterogeneity. Extended Data Figure 2 reports detailed results for each moderator level. Overall, while differences were observed for various subgroups, we found no indication of a moderating effect of the study population (β = 0.01, *Exp*(β), = 1.01, 95% CI [0.51, 2.00], p = 0.969), data collection design ($F(2, 12.7) = 3.4$, p = 0.066), nor the logging method ($F(3, 14.5) = 2.85$, p = 0.074). Finally, a post hoc, multiple-moderator model was produced to account for potential confounds among the two original moderators (medium and measure type). The Approximate Hotelling-Zhang test provided no evidence for a moderating effect ($F(3, 16.5) = 0.103$, p = 0.903), with comparable results for measure type (β = 0.00, *Exp*(β), = 1.00, 95% CI [0.87, 1.15], p = 0.992) and no statistically significant effect for medium (β = -0.03, *Exp*(β), = 0.97, 95% CI [0.86, 1.09], p = 0.646). While reduced in magnitude, heterogeneity remained high ($T^2 = 0.015$, $I^2 = 91.22\%$).

Discussion

Given the widespread reliance on self-report measures of media use across many areas of social science research [13—15], the validity of these measures is a fundamental concern. Before we can make conclusions about media uses and the effects thereof, we must be confident that the measures we use accurately reflect the behaviour that they are designed to assess [5, 20]. Our findings, however, indicate only a modest association between self-reports and usage logs, leading us to conclude that self-report measures of media use may not be a valid stand-in for more objective measures. Notwithstanding the potential biases affecting log-data[5, 35, 39, 40], if these measures are taken to be a valid reflection of actual usage [5, 21, 32, 41, 85], our findings raise important concerns about the validity of findings and conclusions across many areas of the social sciences in which self-reported media use is a central outcome or explanatory variable.

Although there is no widely accepted threshold for convergent validity [86, 87], given the magnitude of the associations found in this meta-analysis, the available evidence suggests that self-reported measures should not automatically be considered suitable substitutes for logs of media use. Our observation of an even smaller association between problematic use scales and device logs suggests even more caution when adopting measures of problematic use to make claims about media usage itself. Moreover, while the results show that similar proportions of studies indicate either under- or over-reporting, less than 10% of self-reports are within 5% of the equivalent logged value, indicating that, when asked to estimate their usage, participants are rarely accurate.

Given the predominance of self-report measures in much of communication and media or psychology research [5, 22, 50], the implications of the non-correspondence between self-reported and logged media use measures observed in this study are considerable. An important unanswered question is whether the discrepancy is indicative of random or systematic measurement error. Some studies provide support for the argument that self-reports have attenuated effect sizes and increased the likelihood of false negatives [50], a larger number of studies, however, suggest that the (in)accuracy of self-reported media use measures may indeed be systematic. For instance, multiple studies have found that the accuracy of self-reported media use depends, in part, on how much the respondent uses media[5, 31, 37, 44]. Furthermore, a recent study [31] found that the degree of inaccuracy was directly related to the respondent's level of well-being. Although our meta-analysis has shown that, across studies, the association between logged and reported media use is generally insufficient to conclude that the measures are appropriate substitutes, given the information reported in primary studies, further investigation is needed to investigate the likely systematic nature of this discrepancy.

While more research is needed to understand the effects of the discrepancy between self-reported and logged measures of media use on the validity of extant findings, given that study conclusions regarding purported negative effects of media use are often far-reaching and disconnected from the methods of their production, our findings have implications beyond knowledge generation and methodological practices. Because findings regarding media use and well-being have the potential to foment societal or policy changes [88], concerns about the quality of evidence extend to any claims or recommendations made on their basis. The results presented herein suggest pause in drawing wide-reaching conclusions—whether these relate to knowledge claims or policy recommendations—from studies relying solely on self-report measures of media use.

Although our findings are indicative of poor convergent validity, there remains a high-level of heterogeneity in effect sizes for correlations involving self-reported usage as well as for the ratio of means between logged and self-reported media use. Taken together, this indicates that the observed association and degree of over-reporting may not be consistent. Various methodological, contextual, participant, or medium-specific factors may impact the degree of alignment between self-reports and logged measures of media use. To investigate this heterogeneity, we considered whether the findings were influenced by relevant methodological factors. The results, however, indicate that both the reporting accuracy and the pooled correlation were not moderated by the category of use, the population involved, the sampling approach, nor the log collection method. Additionally, the form of self-report measure did not affect the correlation between logged and self-reported media use measures. Our investigation of the moderating effect of different media was, however, hampered by the absence of a sufficient number of studies measuring both logged and self-reported use within each category. For this reason, the results cannot confidently speak to the moderating effect of the medium on the relationship between self-reported and logged measures. The remaining unexplained heterogeneity in associations between logged and self-reported media use, and the degree to which participants accurately estimate their usage, are important avenues for future research. Addressing this gap would bring us closer to being able to incorporate knowledge of reporting inaccuracies to recalibrate models derived on the basis of self-report measures of media use. In contrast to these two assessments, only a low level of heterogeneity was observed for correlations involving self-reported problematic use. This suggests, firstly, that the weak relationship with logged measures of usage is relatively stable across comparisons and, secondly, given the differences in observed correlations and heterogeneity between general usage self-reports and problematic usage self-reports, that measures of problematic use, not unexpectedly, capture constructs distinct from those reflected in general media use self-reports.

Notwithstanding that evidence of poor convergent validity is indicative of weak construct validity, it is not sufficient to claim that a measure is necessarily invalid —just that one or both of the measures of interest may not effectively capture the intended construct [87]. While, at face-value, tracking methods provide more accurate and valid measures of media use than self-reports [5, 21, 41, 46, 85], the possibility of biases and inaccuracies in these tracking measures cannot be ignored [5, 35, 39, 40, 50]. In addition to technical incompatibilities (device or system restrictions and errors), gaps in coverage, possible mismatches between the digital traces measured and the constructs targeted [89, 90], variation in accuracy due to system settings, participant biases (reactivity), and increased resource demands (time, cost, and participant burden), there are substantial ethical, security and privacy related challenges associated with tracking media use[5, 40]. A particular concern with such methods is the possibility that some forms of usage tracking may inadvertently log background activities as instances of active usage, thereby overestimating active usage[5, 39]. Moreover, while the recording accuracy of some tracking tools has been validated against external timers, prospective loggers, or manual recordings [46, 85], more research is needed to understand the accuracy of these tools, especially those developed by third parties for general usage.

Despite these potential biases and concerns with logging techniques, we share the belief that, while "client logs may not be perfect, they should be more reliable and less biased than self-reports"[5]. Therefore, while our findings represent at their core a substantial discrepancy between the two measurement forms, they are also a strong signal for the poor validity of self-reports of media use. If subsequent research, building on existing validation results [46, 85], provides further evidence for the accuracy of media use logs, our conclusion that self-reports of media use are biased and inaccurate will be further supported. Therefore, just as calls for higher standards of evidence have prompted examination of the validity of self-report measures of media use, there is a need to further understand the validity of logged measures [89, 90] and continually develop improved tools for quantifying media use.

In addition to concerns around the validity of logged data, there are other limitations to our review. First, although a number of analyses were conducted to assess potential biases, there remains the possibility that various publication biases may have had an impact on the targeted literature base potentially influencing our study outcomes. Second, the quality of our synthesis is only as good as the quality of evidence in the included studies. While a majority of included studies were rated as acceptable in quality, given the Q-SSP checklist, a small number of studies were considered to be of lower quality. These quality concerns related primarily to the sample size and sampling method used in the included studies. Although small convenience samples are common in the social sciences [91], there is a risk that the observed effect sizes could be unstable or inflated. An additional concern is the non-normality inherent in both self-reported and logged media use measures [31, 37, 52]. While the majority of included studies did not report the distribution of these variables (see the supplementary information for a description of those that did), this likely non-normality may introduce a small positive bias in the included correlation coefficients [92]. A final limitation concerns the heterogeneity of the effect sizes present in our sample. Although moderator analyses were conducted to investigate this heterogeneity, they were largely inconclusive—likely owing to the small number of studies present within each moderator level. As the literature in this domain expands, future work should return to this issue, seeking to understand how the accuracy of self-reported media use is contingent on various respondent attributes and media characteristics.

Overall, the findings presented herein highlight the substantial discrepancy between self-reports of media use and equivalent measures produced through usage logging techniques. Given our conclusion that this discrepancy is also a strong signal for the limited construct validity of self-report measures of media use, researchers interested in measuring media use are faced with the question of how to proceed. To this end, we offer the following recommendations. First, as others have suggested, it is time for researchers to stop pretending that self-reports are accurate indicators of actual behaviour[5]. When reporting findings derived on the basis of self-report measures, variables representing media usage should be clearly indicated as self-reported and scholars should adjust their inferences and conclusions accordingly. Second, researchers should endeavour to use a measure that most closely approximates the behaviour that they are targeting. In almost all cases, therefore, researchers should use tracking or logging services to measure media usage. Third, while statistical approaches cannot resolve all biases and sources of error, if research can identify factors that systematically account for discrepancies, they can be modelled and used to account for the misalignment between self-reported and logged measures of digital media use [93—95].

Finally, the current findings signal a need for us to reflect on our current literature and the measures that underlie its production and, on this basis, reconsider our confidence in extant findings. The conceptual tension brought about by our validity concerns should stimulate a drive for theories that have a higher degree of verisimilitude and greater utility for addressing important questions facing society today. In addition to the need for research on media uses and effects to move on from "the repetitive development of self-report assessments" [21], as Kaye et al. [96], Meier and Reinecke [97], Ernala et al. [47], and Büchi [98] discuss, there is a need for a paradigm shift in which specific affordances, behaviours, and digital practices receive central focus, rather than simply the overall duration or volume of usage. Coupled with more valid measures and transparent and robust analytical practices, such developments will bring us closer to understanding the uses and effects that digital media enable.

References

References marked with an asterisk indicate papers included in the meta-analysis.

1. Dickson, K. et al. Screen-based activities and children and young people's mental health and psychosocial wellbeing: a systematic map of reviews. *London: EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London. (2019).*
2. Twenge, J. M. Have smartphones destroyed a generation. *The Atlantic*. Retrieved at: https://www.theatlantic.com/magazine/archive/2017/09/has-the-smartphone-destroyed-a-generation/534198/ (Retrieved May 19, 2020). (2017).
3. Orben, A. The Sisyphean Cycle of Technology Panics. *Perspectives on Psychological Science*, ***15*****(5)**, 1143–1157 (2020).
4. Orben, A., & Przybylski, A. K. The association between adolescent well-being and digital technology use. *Nature Human Behaviour* **3(2)**, 173 (2019).
5. *Scharkow, M. The accuracy of self-reported internet use—a validation study using client log data. *Communication Methods and Measures*, **10(1)**, 13–27 (2016).
6. House of Commons Science and Technology Committee 14th Report - Impact of social media and screen-use on young people's health. Available at: https://tinyurl.com/y5omat3s (2019).
7. Allen, R. L. The reliability and stability of television exposure. *Communication Research* **8(2)**, 233–256 (1981).
8. Bartels, L. M. Messages received: The political impact of media exposure. *American Political Science Review* **87**, 267–285 (1993).
9. Price, V. The impact of varying reference periods in survey questions about media use. *Journalism Quarterly* ***70*****(3)**, 615-627 (1993).
10. Collopy, F. Biases in retrospective self-reports of time use: An empirical study of computer users. *Management Science* ***42*****(5)**, 758-767 (1996).
11. Greenberg, B. et al. Comparing Survey and Diary Measures of Internet and Traditional Media Use, Communication Reports **18**, 1-8 (2005).
12. Coromina, L., & Saris, W. E. Quality of media use measurement. *International Journal of Public Opinion Research* **21(4)**, 424–450 (2009).
13. Gerpott, T. J., & Thomas, S. Empirical research on mobile internet usage: A meta-analysis of the literature. *Telecommunications Policy,* **38(3)**, 291–310 (2014).

14. Howard, M. C., & Jayne, B. S. An analysis of more than 1,400 articles, 900 scales, and 17 years of research: The state of scales in cyberpsychology, behavior, and social networking. *Cyberpsychology, Behavior, and Social Networking*, **18(3)**, 181–187 (2015).
15. Griffioen, N., Rooij, M. van, Lichtwarck-Aschoff, A., & Granic, I. Toward Improved Methods in Social Media Research. *Technology, Mind, and Behavior,* **1(1)**. (2020).
16. Jungselius, B., & Weilenmann, A. Conceptualizing 'use' in social media studies. In *Proceedings of the 9th international conference on social media and society* (pp. 325–329). (2018).
17. Humphreys, L., Karnowski, V., & von Pape, T. Smartphones as metamedia: A framework for identifying the niches structuring smartphone use. *International Journal of Communication* ***12,*** (2018).
18. Abendroth, A., Parry, D. A., le Roux, D. B., & Gundlach, J. An analysis of problematic media use and technology use addiction scales–what are they actually assessing? In *Conference on e-Business, e-Services and e-Society* (pp. 211-222). (Springer, Cham, 2020).
19. Davidson, B. I., Shaw, H., & Ellis, D. A. Fuzzy constructs in assessment: The overlap between mental health and technology 'use'. Preprint at PsyArXiv 10.31234/osf.io/6durk (2020).
20. Flake, J. K., & Fried, E. I. Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychology* (2020).
21. *Ellis, D. A., Davidson, B. I., Shaw, H., & Geyer, K. Do smartphone usage scales predict behavior? *International Journal of Human-Computer Studies* **130**, 86–92 (2019).
22. Ellis, D. A. Are smartphones really that bad? improving the psychological measurement of technology-related behaviors. *Computers in Human Behavior* **97**, 60–66 (2019).
23. Lemmens, J. S., Valkenburg, P. M., & Peter, J. Development and validation of a game addiction scale for adolescents. *Media psychology* **12(1)**, 77-95 (2009).
24. Jobe, J. B. Cognitive psychology and self-reports: models and methods. *Quality of Life Research* **12(3)**, 219–227 (2003).
25. Sudman, S., Bradburn, N. M., & Schwarz, N. *Thinking about answers: The application of cognitive processes to survey methodology*. (Jossey-Bass, 1996).
26. Schwarz, N., & Oyserman, D. Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation* 22(2), 127–160 (2001).
27. Tourangeau, R., et al. C*ognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines: Report of the Advanced Research Seminar on Cognitive Aspects of Survey Methodology* (National Academic Press, 1984).

28. Belson, W. A. *The design and understanding of survey questions* (Aldershot: Gower, 1981).
29. Mieczkowski, H., Lee, A. Y., & Hancock, J. T. Priming Effects of Social Media Use Scales on Well-Being Outcomes: The Influence of Intensity and Addiction Scales on Self-Reported Depression. *Social Media+ Society* ***6*(4),** (2020).
30. Neisser, U. Nested structure in autobiographical memory. In D. C. Rubin (Ed.), *Autobiographical memory* (pp. 71–88). (Cambridge University Press, 1986).
31. *Sewall, C. J., Bear, T. M., Merranko, J., & Rosen, D. How psychosocial well-being and usage amount predict inaccuracies in retrospective estimates of digital technology use. *Mobile Media & Communication*. (2020).
32. *Andrews, S., Ellis, D. A., Shaw, H., & Piwek, L. Beyond self-report: Tools to compare estimated and real-world smartphone use. *PloS one* **10(10),** (2015).
33. Vorderer, P., Krömer, N., & Schneider, F. M. Permanently online–permanently connected: Explorations into university students' use of social media and mobile smart devices. *Computers in Human Behavior* **63**, 694–703 (2016).
34. Piwek, L., Ellis, D. A., & Sally, A. Can programming frameworks bring smartphones into the mainstream of psychological science? *Frontiers in psychology* **7** (2016).
35. Ryding, F. C., & Kuss, D. J. Passive objective measures in the assessment of problematic smartphone use: A systematic review. *Addictive Behaviors Reports* **11**, (2020).
36. *Boase, J., & Ling, R. Measuring mobile phone use: Self-report versus log data. *Journal of Computer-Mediated Communication* **18(4)**, 508–519 (2013).
37. *Vanden Abeele, M., Beullens, K., & Roe, K. Measuring mobile phone use: Gender, age and real usage level in relation to the accuracy and validity of self-reported mobile phone use. *Mobile Media & Communication* **1(2)**, 213–236 (2013).
38. *Kobayashi, T., & Boase, J. No such effect? The implications of measurement error in self-report measures of mobile communication use. *Communication Methods and Measures* ***6*(2)**, 126-143 (2012).
39. *Deng, T. et al. Measuring smartphone usage and task switching with log tracking and self-reports. *Mobile Media & Communication* ***7*(1)**, 3-23 (2019).
40. Jürgens, P., Stark, B., & Magin, M. Two half-truths make a whole? on bias in self-reports and tracking data. *Social Science Computer Review* **38(5)**, 600-615, (2019).

41. *Ohme, J., Araujo, T., de Vreese, C. H., & Piotrowski, J. T. Mobile data donations: Assessing self-report accuracy and sample biases with the iOS Screen Time function. *Mobile Media & Communication*. (2020).
42. Parry, D., Davidson, B. I., Sewall, C., Fisher, J. T., Mieczkowsk*i,* H., & Quintana, D. S. Associations Between Log-Based and Self-Reported Digital Media Use: A Systematic Review and Meta-analysis Protocol. Preprint at OSF Preprints https://osf.io/xrg5c (2020).
43. *Ahn, H., Nguyen, G. T., Lee, H., Jo, S. J., Kim, E. J., & Yim, H. W. Data collection and analysis of smartphone use and sleep of secondary school children. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 410-413). (2017).
44. *Araujo, T., Wonneberger, A., Neijens, P., & de Vreese, C. How much time do you spend online? Understanding and improving the accuracy of self-reported measures of Internet use. *Communication Methods and Measures*, ***11*****(3)**, 173-190 (2017).
45. *Burke, M., Marlow, C., & Lento, T. Social network activity and social well-being. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1909-1912). . (2010).
46. *Elhai, J. et al. Depression and emotion regulation predict objective smartphone use measured over one week. *Personality and Individual Differences* ***133***, 21-28 (2018).
47. *Ernala, S. K., Burke, M., Leavitt, A., & Ellison, N. B. 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* (pp. 1-14). (2020).
48. *Felisoni, D. D., & Godoi, A. S. Cell phone usage and academic performance: An experiment. *Computers & Education* ***117***, 175-187, (2018).
49. *Inyang, I., Benke, G., Morrissey, J., McKenzie, R., & Abramson, M. How well do adolescents recall use of mobile telephones? Results of a validation study. *BMC Medical Research Methodology* ***9*****(1)**, (2009).
50. *Jones-Jang, S. M., Heo, Y. J., McKeever, R., Kim, J. H., Moscowitz, L., & Moscowitz, D). Good News! Communication Findings May be Underestimated: Comparing Effect Sizes with Self-Reported and Logged Smartphone Use Data. *Journal of Computer-Mediated Communication*. (2020).
51. *Junco, R. Comparing actual and self-reported measures of Facebook use. *Computers in Human Behavior,* **29(3)**, 626–631 (2013).

52. *Kahn, A. S., Ratan, R., & Williams, D. Why we distort in self-report: Predictors of self-report errors in video game play. *Journal of Computer-Mediated Communication* **19(4)**, 1010–1023, (2014).
53. *Katapally, T. R., & Chu, L. M. Methodology to derive objective screen-state from smartphones: A SMART platform study. *International journal of environmental research and public health*, ***16(13),*** (2019).
54. *Lee, H., Ahn, H., Choi, S., & Choi, W. The SAMS: Smartphone addiction management system and verification. *Journal of medical systems* ***38*****(1),** (2014).
55. *Lee, H., Ahn, H., Nguyen, T. G., Choi, S. W., & Kim, D. J. Comparing the self-report and measured smartphone usage of college students: a pilot study. *Psychiatry investigation* ***14*****(2)** (2017).
56. *Lin, Y. et al. Time distortion associated with smartphone addiction: Identifying smartphone addiction via a mobile application (App). *Journal of psychiatric research* **65**, 139-145. (2015).
57. *Loid, K., Täht, K., & Rozgonjuk, D. Do pop-up notifications regarding smartphone use decrease screen time, phone checking behavior, and self-reported problematic smartphone use? Evidence from a two-month experimental study. *Computers in Human Behavior* ***102***, 22-30 (2020).
58. *Mikkelsen, S., Vilstrup, I., Lassen, C. F., Kryger, A. I., Thomsen, J. F., & Andersen, J. H. Validity of questionnaire self-reports on computer, mouse and keyboard usage during a four-week period. *Occupational and Environmental Medicine* ***64(8)***, 541-547. (2007).
59. *Montag, C. et al.. Recorded behavior as a valuable resource for diagnostics in mobile phone addiction: evidence from psychoinformatics. *Behavioral Sciences* ***5*****(4)**, 434-442. (2015).
60. *Newell, C. A. *Smartphone and Social Media use and its Health Associations* (Masters dissertation). The University of Texas at Arlington. (2018).
61. *Noë, B., Turner, L. D., Linden, D. E., Allen, S. M., Winkens, B., & Whitaker, R. M. Identifying indicators of smartphone addiction through user-app interaction. *Computers in human behavior* ***99***, 56-65. (2019).
62. *Oeldorf-Hirsch, A., & Chen, Y. Who Cares about Screen Time? Predicting the use of Mobile Phone Tracking Features. *Paper presented at The 70th annual conference of the International Communication Association (ICA)*, Gold Coast, Australia (Virtual Conference). (2020)

63. *Pan, Y. C., Lin, H. H., Chiu, Y. C., Lin, S. H., & Lin, Y. H. Temporal stability of smartphone use data: Determining fundamental time unit and independent cycle. *JMIR mHealth and uHealth*, ***7*(3).** (2019).

64. *Prasad, S. et al. A study of magnitude and psychological correlates of smartphone use in medical students: A pilot study with a novel telemetric approach. *Indian journal of psychological medicine* ***40*(5)**, 468-475 (2018).

65. *Rosen, L. D., Carrier, L. M., Pedroza, J. A., Elias, S., O'Brien, K. M., Lozano, J., ... & Ruiz, A. The role of executive functioning and technological anxiety (FOMO) in college course performance as mediated by technology usage and multitasking habits. *Educational Psychology* **24(1)**, 14-25 (2017).

66. *Rozgonjuk, D., Levine, J. C., Hall, B. J., & Elhai, J. D. The association between problematic smartphone use, depression and anxiety symptom severity, and objectively measured smartphone use over one week. *Computers in Human Behavior* ***87***, 10-17 (2018).

67. *Rozgonjuk, D., Pruunsild, P., Jürimäe, K., Schwarz, R. J., & Aru, J. Instagram use frequency is associated with problematic smartphone use, but not with depression and anxiety symptom severity. *Mobile Media & Communication* (2020).

68. *Sela, Y., Zach, M., Amichay-Hamburger, Y., Mishali, M., & Omer, H. Family environment and problematic internet use among adolescents: the mediating roles of depression and fear of missing out. *Computers in Human Behavior* ***106***, (2020).

69. *Shaw, H., Ellis, D., Geyer, K., Davidson, B., Ziegler, F., & Smith, A. Quantifying smartphone "use": Choice of measurement impacts relationships between "usage" and health. *Technology, Mind, and Behavior* **1(2)**, (2020).

70. *Shin, C., & Dey, A. K. Automatically detecting problematic use of smartphones. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (pp. 335-344). (2013).

71. *Shin, M., & Lee, K. Measuring smartphone usage time is not sufficient to predict smartphone addiction. *Journal of Theoretical and Applied Information Technology*, ***95*(19)**, (2017).

72. *Shum, M., Kelsh, M. A., Sheppard, A. R., & Zhao, K. An evaluation of self-reported mobile phone use compared to billing records among a group of engineers and scientists. *Bioelectromagnetics* ***32*(1)**, 37-48 (2011).

73. *Singh, V. K., & Jain, A. Toward harmonizing self-reported and logged social data for understanding human behavior. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 2233-2238). (2017).
74. *Timotijevic, L., Barnett, J., Shepherd, R., & Senior, V. Factors influencing self-report of mobile phone use: the role of response prompt, time reference and mobile phone use in recall. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* ***23*****(5)**, 664-683. (2009).
75. *Tokola, K., Kurttio, P., Salminen, T., & Auvinen, A. Reducing overestimation in reported mobile phone use associated with epidemiological studies. *Bioelectromagnetics: Journal of the Bioelectromagnetics Society, The Society for Physical Regulation in Biology and Medicine, The European Bioelectromagnetics Association* ***29*****(7)**, 559-563 (2008).
76. *van Berkel, N., Goncalves, J., Lovén, L., Ferreira, D., Hosio, S., & Kostakos, V. Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. *International Journal of Human-Computer Studies* ***125***, 118-128 (2019).
77. *Wilmer, H. H., Hampton, W. H., Olino, T. M., Olson, I. R., & Chein, J. M. Wired to be connected? Links between mobile technology engagement, intertemporal preference and frontostriatal white matter connectivity. *Social cognitive and affective neuroscience* **14(4**), 367-379 (2019).
78. *Yuan, N., Weeks, H. M., Ball, R., Newman, M. W., Chang, Y. J., & Radesky, J. S. How much do parents actually use their smartphones? Pilot study comparing self-report to passive sensing. *Pediatric research* ***86*****(4)**, 416-418 (2019).
79. Moyer, A., & Finney, J. W. Rating methodological quality: toward improved assessment and investigation. *Accountability in research* **12(4)**, 299-313 (2005).
80. Crowe, M., & Sheppard, L. A review of critical appraisal tools show they lack rigor: alternative tool structure is proposed. *Journal of clinical epidemiology,* **64(1)**, 79-89 (2011).
81. Protogerou, C., & Hagger, M. S. A case for a study quality appraisal in survey studies in psychology. *Frontiers in psychology* ***9*** (2019).
82. Protogerou, C., & Hagger, M. S. A checklist to assess the quality of survey studies in psychology. *Methods in Psychology* **3** (2020).
83. Rodgers, M. A., & Pustejovsky, J. E. Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological methods* (2020).
84. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software,* **36(3),** 1–48 (2010).

85. Geyer, K., Ellis, D. A., Shaw, H., & Davidson, B. I. Open source smartphone apps and tools for measuring, quantifying, and visualizing screen time. Preprint at PsyArXiv: https://doi.org/10.31234/osf.io/eqhfa (2020).
86. Nunnally, J. *Psychometric theory 3rd Edition* (McGraw-Hill Education, 1994).
87. Carlson, K. D., & Herdman, A. O. Understanding the impact of convergent validity on research results. *Organizational Research Methods* **15(1)**, 17–32 (2012).
88. Social Media Addiction Reduction Technology Act, S. 2314, 116th Congress. https://www.congress.gov/bill/116th-congress/senate-bill/2314/all-info (2019).
89. Salganik, M. J. *Bit by bit: Social research in the digital age* (Princeton University Press, 2019).
90. Sen, I., Floeck, F., Weller, K., Weiss, B., & Wagner, C. (2019). A total error framework for digital traces of humans. Preprint at *arXiv arXiv:1907.08228v3* (2019)
91. Dienlin, T. et al. An Agenda for Open Science in Communication. *Journal of Communication* **71 (1)**, 1 – 26 (2020)
92. Bishara, A. J., & Hittner, J. B. Reducing bias and error in the correlation coefficient due to nonnormality. Educational and psychological measurement **75(5)**, 785-804 (2015).
93. Lazer, D. M., et al. Computational Social Science: Obstacles and Opportunities. *Science* **369 (6507)**, 1060 – 1062 (2020).
94. Sewall, C. R., & Parry, D. A., What is the role of depression in the discrepancy between estimated and actual iPhone use? An exploratory analysis using cubic response surface analysis. Preprint at OSF Preprints: https://doi.org/10.31219/osf.io/e9pfa
95. Van Smeden, M., Lash, T. L., & Groenwold, R. H. Reflection on modern methods: Five myths about measurement error in epidemiological research. *International Journal of Epidemiology* **49(1)**, 3444 – 347 (2020).
96. Kaye, L. K., Orben, A., Ellis, D. A., Hunter, S. C., & Houghton, S. The Conceptual and Methodological Mayhem of "Screen Time." *International Journal of Environmental Research and Public Health* **17(10)**, (2020).
97. Meier, A., & Reinecke, L. Computer-Mediated Communication, Social Media, and Mental Health: A Conceptual and Empirical Meta-Review. *Communication Research.* (2020).
98. Büchi, M. A Proto-Theory of Digital Well-Being. Preprint at OSF Preprints https://doi.org/10.31219/osf.io/k3e2j (2020).
99. Harris, B., Regan, T., Schueler, J., & Fields, S. A. Problematic mobile phone and smartphone use scales: A systematic review. *Frontiers in Psychology* **11**, (2020).

100. Laconi, S., Rodgers, R. F., & Chabrol, H. The measurement of internet addiction: A critical review of existing scales and their psychometric properties. *Computers in Human Behavior* **41,** 190–202 (2014).

101. Webster, J., & Watson, R. Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly* **26(2).** (2002).

102. Hedges, L. V., Tipton, E., & Johnson, M. C. Robust variance estimation in meta-regression with dependent effect size estimates. *Research synthesis methods* **1(1)**, 39–65 (2010).

103. Fisher, Z., & Tipton, E. Robumeta: An R-package for robust variance estimation in meta-analysis. Preprint at arXiv arXiv:1503.02220v1 (2015).

104. Fu, R. et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. *Journal of clinical epidemiology* **64(11)**, 1187-1197 (2011).

105. Gilpin, A. R. Table for conversion of kendall's tau to spearman's rho within the context of measures of magnitude of effect for meta-analysis. *Educational and psychological measurement* **53(1)**, 87–92 (1993).

106. Rupinski, M. T., & Dunlap, W. P. Approximating Pearson product-moment correlations from Kendall's tau and Spearman's rho. *Educational and Psychological Measurement* **56**, 419–429 (1996).

107. Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. *Introduction to meta-analysis* (John Wiley & Sons, 2011).

108. Cohen, J. *Statistical power analysis for the behavioral sciences* (Academic Press, 1988).

109. Friedrich, J. O., Adhikari, N. K., & Beyene, J. The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: a simulation study. *BMC Medical Research Methodology* **8(1),** 32 (2008).

110. Lajeunesse, M. J. On the meta-analysis of response ratios for studies with correlated and multi-group designs. *Ecology* **92(11)**, 2049–2055 (2011).

111. Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology* **2**, 85– 112 (2016).

112. Pustejovsky, J. E. clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections (R Package Version 0.4.2) (2017).

113. Egger, M., Smith, G. D., Schneider, M., & Minder, C. Bias in meta-analysis detected by a simple, graphical test. *BMJ: British Medical Journal,* **315(7109)**, 629–634 (1997).

114. Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. Contour- enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of clinical epidemiology* **61(10**), 991–996 (2008).

Acknowledgements

We wish to acknowledge all of the researchers who made their data available for inclusion in this review. We are especially grateful for those who completed extra analyses to provide us with relevant data and statistics. Additionally, we acknowledge the role played by those who shared our public calls for relevant data and the valuable input provided by the reviewers of this paper. The authors received no specific funding for this work.

Author contributions

D.A.P and B.I.D conceived the study. D.A.P, B.I.D, C.J.R.S, J.T.F, and H.M collected the data. D.A.P analysed the data with input from D.S.Q. D.A.P, B.I.D, C.J.R.S, J.T.F, and H.M wrote the first draft of the paper. All authors (D.A.P, B.I.D, C.J.R.S, J.T.F, H.M, and D.S.Q) discussed the results and contributed to revision of the final manuscript.

Competing interests

The author(s) declare that there are no conflicts of interest with respect to the authorship or the publication of this article.

Figure legends

Figure 1. PRISMA flow diagram for the study inclusion process. A total of 47 records fulfilled the eligibility criteria.

Figure 2. **Forest plot of the effect sizes for studies included in the meta-analysis for the association between self-reported and logged measures of digital media use.** Individual Pearson's *r* estimates are depicted by filled squares, with the square sizes indicating the relative weight of each effect size estimate in the meta-analysis. The filled diamond represents the overall summary effect size ($r = 0.38$, 95% CI [0.33, 0.42], $p < 0.001$). The error bars and diamond width represent the 95% CIs for the effect sizes. The dashed reference line at the intercept for $r = 0.5$ represents the point from which the magnitude of the association would be sufficient to conclude that the measures are appropriate substitutes for one another. RE = Random effects model. RVE = Robust variance estimation (conducted with a correlated effects weighting scheme).

Figure 3. **Contour-enhanced funnel plots.** The plot depicts the relationship between the observed effect sizes (on the x-axis) and their standard errors (on the y-axis) for comparisons concerning digital media use (A), problematic use (B) and reporting accuracy (C). The vertical lines indicate the estimated summary effect size. The shaded bands represent the significance contours indicated in the legend and each black dot represents an observed effect size. Visual inspection of all three plots does not indicate asymmetry, nor does it indicate evidence of publication bias as there is no obvious overrepresentation of effect sizes in the highlighted significance contours.

Figure 4. **Forest plot of the effect sizes for studies included in the meta-analysis for the association between self-reported problematic use and logged measures of use.** Individual Pearson's *r* estimates are depicted by filled squares, with the square sizes indicating the relative weight of each effect size estimate in the meta-analysis. The filled diamond represents the overall summary effect size ($r = 0.25$, 95% CI [0.20, 0.29], $p < 0.001$). The error bars and diamond width represent the 95% CIs for the effect sizes. The dashed reference line at the intercept for $r = 0.5$ represents the point from which the magnitude of the association would be sufficient to conclude that the measures are appropriate substitutes for one another. RE = Random effects model. RVE = Robust variance estimation (conducted with a correlated effects weighting scheme).

Figure 5. Forest plot of the effect sizes for studies included in the meta-analysis for the ratio of means between self-reported and logged measures of digital media use. The results are represented on a log scale. Individual response ratios (ratio of means) are depicted by filled squares, with the square sizes indicating the relative weight of each effect size estimate in the meta-analysis. The filled diamond represents the overall summary effect size (R = 1.21, 95% CI [0.94, 1.54], p = 0.129). The error bars and diamond width represent the 95% CIs for the effect sizes. The dashed reference line at the intercept for 1.0 represents a 1:1 ratio between self-reported and logged digital media use, with values below one indicating under-reporting and values above one indicating over-reporting of digital media use. RE = Random effects model. RVE = Robust variance estimation (conducted with a correlated effects weighting scheme).

Tables

Table 1. Digital media usage correlations moderator and subgroup analyses.

Moderator	*k*	*r*	β	95% CI	*p*
Medium			-0.03	[-0.14, 0.09]	0.621
Social media	13	0.35		[0.27, 0.43]	< 0.001
Phone	49	0.38		[0.31, 0.45]	< 0.001
Self-report form			0.14	[-0.16, 0.42]	0.265
Scales	6	0.24		[0.00, 0.46]	0.048
Single estimates	60	0.39		[0.34, 0.43]	< 0.001
Self-report category			-0.002	[-0.13, 0.13]	0.978
Usage duration	47	0.38		[0.33, 0.43]	< 0.001
Usage volume	19	0.34		[0.25, 0.43]	< 0.001

Note. *k* = number of included effect size estimates; *r* = Pearson correlation coefficient; β = metaregression coefficient from a model in which a categorical moderator with two levels was entered as a predictor; 95% CI corresponds to the β coefficient for moderators or the *r* values for individual moderator levels; *p* corresponds to the β coefficient for moderators, or the subgroup analysis for individual moderator levels.

Table 2. Reporting accuracy subgroup analyses.

Moderator	*k*	*R*	95% CI	*p*
Medium				
Social media	5	2.89	[0.18, 46.04]	0.241
Phone	41	1.07	[0.84, 1.35]	0.574
Self-report category				
Usage duration	35	1.29	[1.01, 1.66]	0.044
Usage volume	14	0.80	[0.57, 1.11]	0.162

Note. *k* = number of included effect size estimates; *R* = risk ratio; 95% CI corresponds to the *R* values for individual moderator levels; *p* corresponds to the subgroup analysis for individual moderator levels.

Methods

Protocol and Registration

To pre-register our expectations and methodology, our systematic review protocol was made publicly accessible prior to data collection [42]. All materials required to reproduce the results of the study are available on the Open Science Framework (https://osf.io/dhx48/). While we provide formal exploratory research questions and hypotheses in our study protocol, for the sake of brevity, here we simply provide an overview of our a priori expectations for the meta-analysis, before outlining the details of our data collection and analysis procedures.

Given the accuracy and validity issues with self-report measures of media use, we expected the association between self-reported measures of media use and measures produced from digital trace data to be positive, but only small-to-medium in magnitude. To understand if the association between self-reports and logged measures is affected by characteristics of the medium or the self-report measure, we explored whether it is moderated by (a) the medium (i.e., social media, smartphones, the Internet, computers, gaming), (b) the form of self-report measure (i.e., a single estimate or a scale), or (c) the category of media use (i.e., volume of interactions or duration of usage).

In addition to considering associations between measures explicitly concerning media usage, acknowledging that, despite concerns over validation procedures [97, 98] and questionable relations between the constructs assessed and usage [19], scales assessing problematic media use (including excessive usage among other conceptualisations) are frequently adopted to make claims about media usage itself [22, 23, 69], we investigated the association between such measures and logged measures of digital media use. For this separate analysis we also expected the association between self-reported measures of problematic media use and usage measures produced from digital trace data to be positive but small to medium in magnitude.

Our final aim concerned the accuracy of self-report measures, relative to equivalent logged measures of digital media use. To this end, we assessed whether participants typically under- or over-report their digital media use compared to equivalent logged measures. To understand if there are factors that systematically affect accuracy, we investigated if there is evidence indicating that measurement error is systematically related to either the medium or the category of media use involved in a comparison.

Eligibility Criteria

We restricted inclusion to studies that collected both self-reported and logged measures of digital media use. For self-reports, eligible scales or single estimates should have either concerned use in general (i.e., volume or duration) or problematic use (i.e., excessive usage or other conceptions of problematic use). These self-report and logged measures should have concerned use of either social media, games, a mobile phone, the Internet in general, or a computer. For general usage measures, we only considered comparisons between self-report measures that concerned either the total or average duration (e.g., minutes, hours) or volume (e.g., number of pickups, number of logins, number of phone calls etc.) of media use and equivalent logged measures for the same period (e.g., daily, weekly etc.). In addition to these criteria, we restricted inclusion to studies published since 2007 (inclusive), the initial release year for the iOS operating system (with the release of Android in the following year), and a time from which use of social networking services gained widespread popularity. We also restricted inclusion to studies reported in English. While we excluded studies that explicitly targeted clinical populations, no further restrictions were placed on participant populations, nor were restrictions placed on publication status.

Information Sources and Search Strategy

To identify relevant published studies, we conducted an automated search on five broad bibliographic databases: PubMed, Scopus, PsychInfo, Communication & Mass Media Complete, and the ACM Digital Library. To target unpublished (grey) literature we used the ProQuest Dissertations & Theses A&I database. A generic search string was developed in consultation with an academic librarian at Stellenbosch University and, for each database, was adjusted as required. The search string includes four clauses, with at least one matching term required for each clause. The first clause includes terms relating to various forms of eligible media (e.g., social media OR Internet OR phone OR games, etc.). The second and third clauses relate to logged data (e.g., server logs OR track, etc.) and self-report measures (e.g., survey OR self-report OR questionnaire, etc.), respectively. The fourth clause includes terms relating to media use (e.g., use OR usage OR behaviour, etc.). The full search strings (applied to the title, abstract, and keywords fields or just the abstract field if restricted) and search dates are available through the OSF (https://osf.io/dhx48/). In addition to the automated search, a manual search was conducted within five relevant journals (Human Communication Research; Cyberpsychology, Behavior and Social Networking; Communication Methods and Measures; International Journal of Human-Computer Studies; Media Psychology). Following assessment for eligibility, the included studies were supplemented by 'backward' and 'forward' search procedures [101] using the Google Scholar search engine. Finally, we made public calls for relevant unpublished data and papers on Twitter (these tweets were viewed approximately 10,000 times) and the Psychological Methods Discussion Group on Facebook.

Study Selection

After executing the automated search procedure, two authors conducted the manual search. Five authors independently screened the resulting titles and abstracts against the inclusion criteria. The full texts of included studies were then retrieved and screened. Any disagreements were discussed and, if needed, an additional author was consulted. Finally, two authors conducted forward and backward reference-list searches from the included studies. The outcomes of these selection procedures are described at the outset of the results section.

Data Collection

Relevant data were extracted from eligible studies and entered into a spreadsheet. Elements extracted include publication year, a description of the study population involved, study sample size, the source of logged and self-reported data, the form of media use recorded, and measurement produced (e.g., total use, average use, etc.), and the duration for which logged data was acquired. To enable the analysis of convergent validity, effect sizes were extracted from reported correlation analyses for associations between self-reported and logged measures of media use as well as for correlations between problematic use and logged measures. For estimates of use, we only included comparisons for equivalent actions, time periods, and forms (e.g., average phone use per day, total weekly social media use, or daily phone pickups etc.) while, for problematic use scales, we included reported associations with logged measures for the duration or volume of use for any of the five targeted media (e.g., total phone time, average phone pickups, etc.). Both Pearson's product moment correlation coefficients (r) and Spearman's rank-ordered correlation coefficients (r_s) were extracted.

To analyse under- or over-reporting, we extracted measures of central tendency and variability for self-reported estimates that explicitly concern either the duration or the volume of media use reported on a continuous scale and logged measures for equivalent outcomes. To perform moderator analyses, we coded the medium as either 'phone', 'gaming', 'social media', 'computer', or 'Internet'. This categorisation was based on the source of log-tracked data and, in instances in which overlap existed (e.g., social media on a phone), we coded the most specific medium known. Self-report measures were coded to capture one of two outcomes: 'use' or 'problematic use', reflect one of two forms: 'scale' or 'single estimate', and represent one of two categories of use: 'duration' or 'volume' (i.e., use instances).

If reported data were insufficient to compute the necessary effect sizes, we contacted the corresponding authors to request ad hoc analyses or for further descriptive statistics. If, after two attempts the relevant data were still not available, and relevant values were represented in plots in a paper, we used a web plot digitizer (WebPlotDigitizer: https://apps.automeris.io/wpd/) to convert plotted representations into numeric values. If no response was received from corresponding authors and relevant plots were not available to be digitized, the comparison was excluded.

Data Items

To analyse usage correlations the analysis only included effect sizes for correlations between logged usage and self-report measures that explicitly concerned media use. For these analyses, if a study reported correlations for both logged overall use (total or average duration or volume) and logged use of specific smartphone applications or websites, to avoid nested correlations, we excluded correlations involving individual applications or websites and only included comparisons for overall indications of use. However, if an otherwise eligible comparison was reported and no overall use metric was available, comparisons for specific use types were included. Furthermore, if no comparison with overall use was reported, with the exception of social media and gaming, we excluded comparisons that involved aggregations of different applications or websites into higher-level categories (i.e., use of navigation applications, use of video platforms, use of fitness applications etc.). To analyse correlations for measures concerning problematic use, the analysis only included effect sizes for correlations between logged media use and self-reported problematic use. To investigate measurement accuracy, we only considered single point estimates for overall use duration or use instances for a given medium that were provided on a continuous scale. For this investigation we included relevant reported sample sizes, correlations, as well as descriptive statistics (means and standard deviations) for self-reports and equivalent log measures.

Quality of Evidence Assessment

As an addition to our original protocol, to assess the quality of evidence in the included studies, we used the quality of survey studies in psychology (Q-SSP) checklist [82]. Given shortcomings in many existing assessment tools and mismatches with non-medical or experimental research, this checklist, comprising 20 items (item and scoring descriptions are available at https://osf.io/5aepd), was developed to evaluate the quality of psychological studies adopting survey designs. While our targeted body of research typically involves behavioural tracking in addition to survey methods, the Q-SSP nonetheless largely covers relevant quality domains pertinent to this sample. Where necessary, we amended the items or the scoring scheme to fit our scope. An overall quality score, represented as a percentage, is derived on the basis of the proportion of YES scores out of the total applicable items for a given study. Depending on the number of applicable items, studies are required to achieve a score of approximately 70% to be rated as 'acceptable' in quality, while scores less than this threshold suggest that the study may be of 'questionable' quality.

To better suit our specific research context, as is common [81], we made a number of amendments to the Q-SSP checklist. First, noting that many studies in this regard set out objectives or aims rather than specific research questions or hypotheses, for item 1 (the reporting of hypotheses or research questions) we also accepted the former as eligible statements. For item 11 (the reporting of measures in full) we only considered the provision of the self-report measures in the report or any supplementary materials. For studies conducted entirely online (i.e., data collection occurred through MTurk, Prolific, or another platform), items 13 (information about the persons who collected the data) and 14 (information about the context of data collection) were coded as not applicable. For item 15 (information about the duration of data collection), if existing data were provided by the participants (i.e., through data donation), the not applicable code was used. For item 12 (measure validity), given the focus of the present investigation and the emphasis on developing an understanding of measurement validity, this item was coded as not applicable for all studies. Similarly, for item 19 (participant debrief), noting Protogerou and Hagger [82], as the included studies did not involve any form of participant deception, the not applicable code was also used for all studies. Given these amendments, while the original checklist includes between 20 and 16 items, our checklist could include between 18 and 13 items. Therefore, as Protogerou and Hagger [82] recommend, we extended the original scoring scheme to account for these differences. The final study quality assessment sheet is available at: https://osf.io/kcshv/. Because two of the 47 papers were included on the basis of unpublished raw data received directly from the authors, the quality assessment was only conducted for the remaining 45 papers. Three authors independently assessed each study using the Q-SSP checklist, with any disagreements resolved through discussion.

Summary Measures and Synthesis of Results

All analyses were performed with the R statistical programming language (v. 4.0.2). A complete list of the packages used in the analysis is provided in the analysis code available through the OSF (deviating from the protocol, robust variance estimation was conducted with the robumeta package rather than the metafor package as specified). Three distinct meta-analyses were conducted. In the first, we focused on correlations between self-reported and logged media use. In the second, the analysis concerned the degree of under- or over-reporting. In the third, we focused on correlations between self-reported problematic use and logged use. For all analyses we adopted an a priori statistical significance level of α = .05. To account for variance inflation resulting from dependent observations for different measures for the same participants (i.e., some studies provided more than one estimate for the meta-analysis), we used cluster-robust variance estimation (RVE) based on the sandwich method with adjusted estimators for small samples and a correlated effects weighting scheme with the default assumed value of r = 0.80 [102, 103]. For all moderator analyses, acknowledging that there is no widely accepted minimum number of effects required, noting previous recommendations [104], we specified a minimum requirement of four included effects per moderator level.

For the correlational meta-analyses, to stabilise the variances, raw effect sizes were transformed into normalised correlation coefficients (Fisher's z). Effects originally reported as Spearman's r_s were first transformed to Pearson's r and then transformed to Fisher's z for synthesis with the effect sizes originally reported using Pearson's r. Deviating from our preregistration in which we had specified the use of Gilpin's [105] conversion tables for the transformation from r_s to r, we used the following equation specified in Rupinski and Dunlap [106] to perform this transformation and approximate Pearson's r: $r = 2\sin(r_s(\pi/6))$. For reporting, we performed Fisher's z-to-r transformation [107].

For both correlational meta-analyses, we estimated random-effects models to calculate overall summary effect sizes. To interpret the outcomes of the correlational meta-analyses, in-line with Cohen [108], we took correlation coefficients of .1 to be small, .30 to be medium, and .50 or greater to be large effect sizes, respectively. However, noting our aim of investigating convergent validity, acknowledging Carlson and Herdman's [87] recommendations, we considered correlation coefficients above 0.7 to indicate strong evidence of convergent validity, between 0.5 and 0.7 to indicate acceptable convergent validity, and below 0.5 to be inadequate to support convergent validity between the two measurement forms.

To investigate measurement accuracy, we first determined the proportion of comparisons that are indicative of accurate, under-reported, or over-reported media use. For this analysis, we used a margin of error of 5% or more above the tracked measure to indicate over-reporting, 5% or more below to indicate under-reporting, and mean estimates within 5% of the logged measure to be accurate. To quantify the magnitude of the difference in means produced using the different measurement forms, given the within-subjects nature of the analysis and the existence of a true ratio scale with a natural zero point [107], we calculated the log transformed ratio of means [109, 110], and estimated the sampling variance accounting for the correlation between measurements [84]. These unitless effect sizes were then synthesized by estimating a random effects model and then back transformed for reporting (This ratio of means is commonly known as the response ratio R in Ecology research). In this analysis, a value of one corresponds to an equal ratio between self-reported and logged measures, while values less than one indicate under-reporting and values greater than one indicate over-reporting. The magnitude of the outcome represents the ratio of self-reported to logged media use.

Risk of Bias Across Studies

To account for study quality and assess potential biases due to 'small-study effects', which can include publication bias, we visually inspected funnel plot symmetry and performed Egger's regression test [113] for each of the three primary meta-analyses. To visualize possible publication bias, we used a contour-enhanced funnel plot which superimposes notable areas of statistical significance (i.e., $p = 0.1$, $p = 0.05$, $p = 0.01$). An over-representation of effect sizes in the highlighted areas is indicative of possible publication biases [113]. As a further sensitivity analysis, if a model included effect sizes reported in both peer-reviewed and pre-publication studies, we conducted meta-regression moderator analyses to determine if effect sizes reported in peer-reviewed studies differ from pre-publication studies (e.g., preprints, unpublished data, or papers under review). Finally, as an additional post hoc sensitivity analysis, if a model included effect sizes that were included using the web plot digitiser, we synthesized the relevant effects excluding these effect sizes to determine whether our results were robust to this inclusion method.

Additional Analyses

To consider possible sources of heterogeneity in the observed correlations and investigate factors that affect the relationship between self-reported and logged media use, three categorical moderator analyses were conducted. The first concerned the effect of the medium on the correlation (i.e., whether effects differ between studies investigating correlations for social media use, phone use, or gaming for instance). The second considered the potential moderating effect of the measure category (either usage volume or duration), while the third concerned the form of self-report measure (scale or single estimate). For each moderator category, in addition to meta-regression models, we estimated separate random effects models to produce summary effect sizes for each subgroup.

For the analysis of response accuracy, to account for possible sources of heterogeneity, we planned two categorical moderator analyses, estimating random effects models to produce summary weighted effect sizes for each subgroup. In the first, we examined whether the results differed based on the category of use estimated (e.g., use duration or use volume). In the second, we examined whether they differ by the medium.

In addition to these pre-planned moderator analyses, for both the analysis of usage correlations and reporting accuracy, three additional post hoc exploratory moderator analyses were conducted. In the first, we investigated whether the findings were impacted by the population type involved in an analysis. We coded the study samples into five population categories: adolescents; adults; students; general (the sample includes individuals from multiple populations); and unknown. The second additional moderator analysis concerned the method through which tracking data was acquired. We coded the tracking methods into four categories: third party tools; built-in tools; custom tools developed for research purposes; and operator or platform data. The third post hoc moderator analysis concerned the data collection design and, for this analysis, we coded the designs into three categories: data donations (i.e., participants provided the researchers with access to data that had already been collected); direct tracking (i.e., participants installed a tracking tool as part of the study); and operator or platform supplied data (i.e., data on participants' usage were acquired from a platform or network operator). Descriptive statistics for the data underlying these three additional moderator analyses are available in Extended Data Figure 3. To perform an omnibus test for moderators with more than two levels, following Tanner-Smith et al. [111] and Pustejovsky [112], we performed Approximate Hotelling-Zhang (HTZ) tests with small sample corrections using the club sandwich package (Pustejovsky, 2017). Finally, for the analysis of usage correlations and reporting accuracy, we ran post hoc multiple moderator analyses in which all a priori moderators were included simultaneously in the model. For these analyses, as with the a priori moderator analyses, we only included moderator levels with a sufficient number of effects available.

Across all of the pre-planned and post hoc moderator analyses, an important caveat merits noting. While we follow standard procedures, the statistical power of the moderator analyses is limited by the quantity of available evidence reported in primary studies. For this reason, while the results provide an accurate summary of current knowledge, we encourage caution in their interpretation.

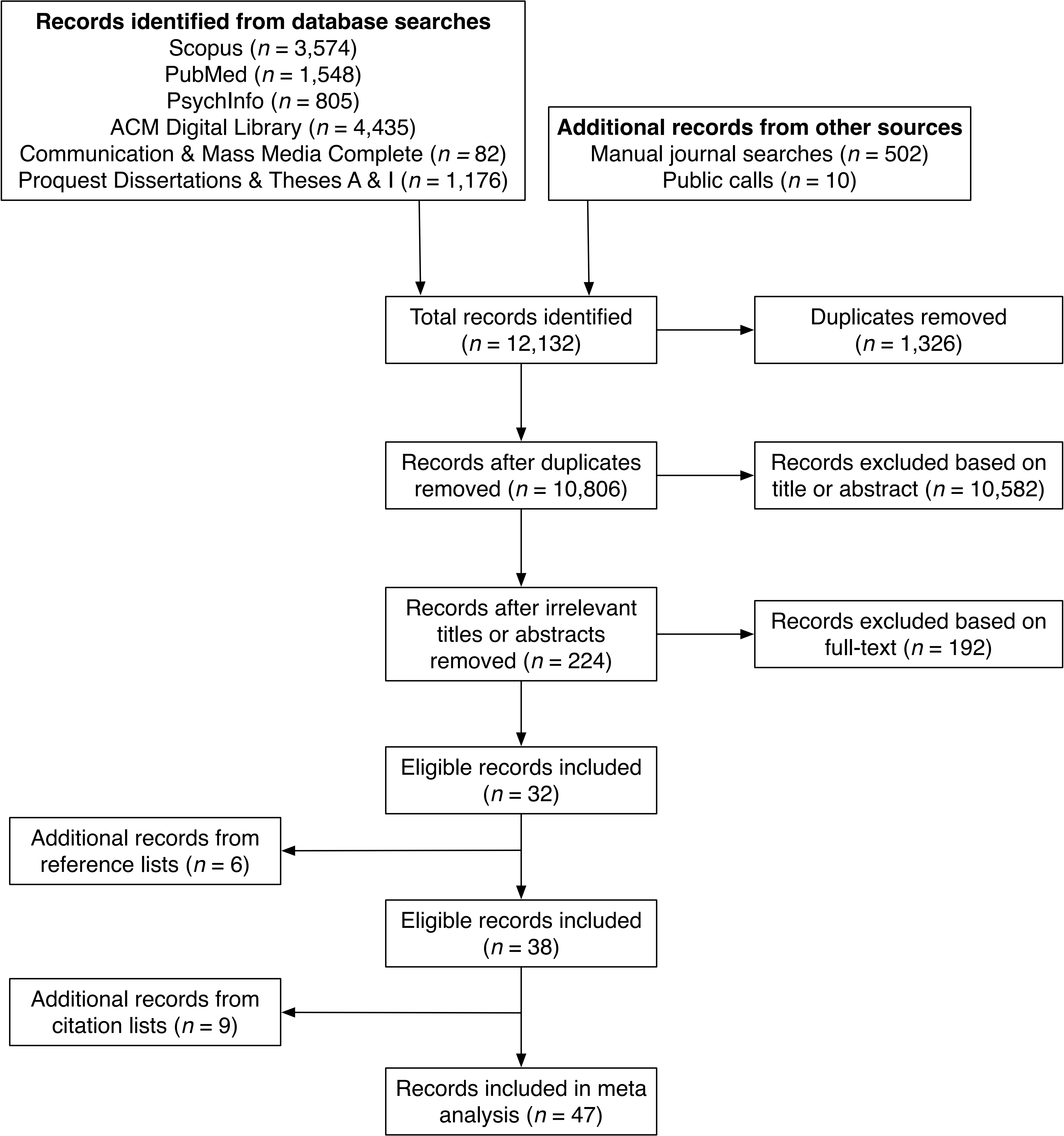
For the three primary meta-analyses, to examine the variance and heterogeneity among effects, we computed Q and I^2, interpreting statistically significant Q values to indicate heterogeneity and I^2 values of approximately 25%, 50%, and 75% to indicate low, moderate, and high heterogeneity, respectively. To determine if the analyses were impacted by any outliers, we conducted outlier and influence diagnostics for the original models (i.e., Cook's distance, covariance ratios, diagonal elements of the hat matrix) using the metafor package [84] and performed leave-one-out sensitivity re-analyses without any identified outliers. Equivalence testing using the two one-sided test (TOST) procedure was also applied to assess evidence for the absence of meaningful effects. A smallest effect size of interest of $r = 0.1$ was used to determine equivalence bounds (i.e., a lower bound of -0.1 and a higher bound of 0.1). The results of the TOST procedure are presented in the Supplementary Information.

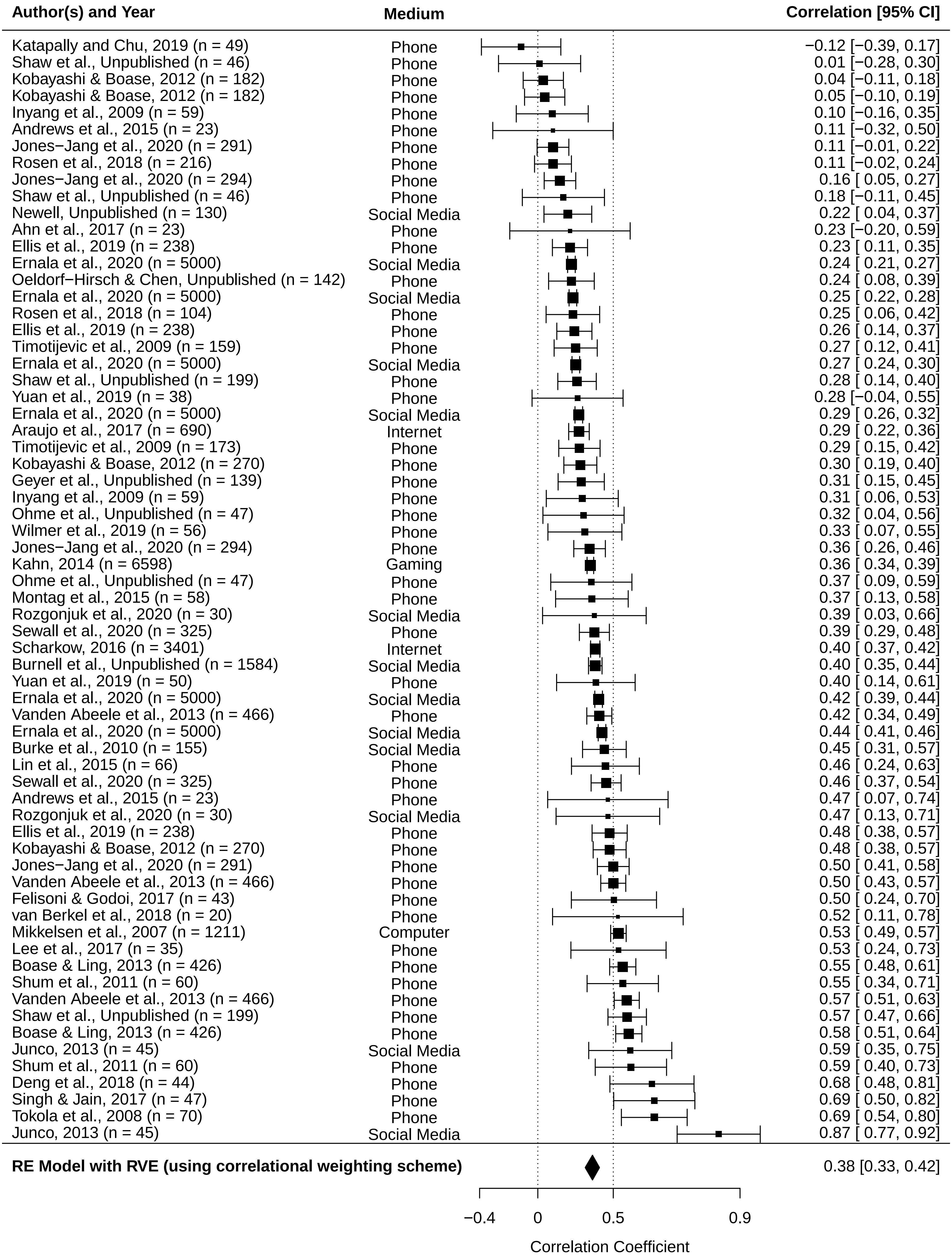
Data Availability

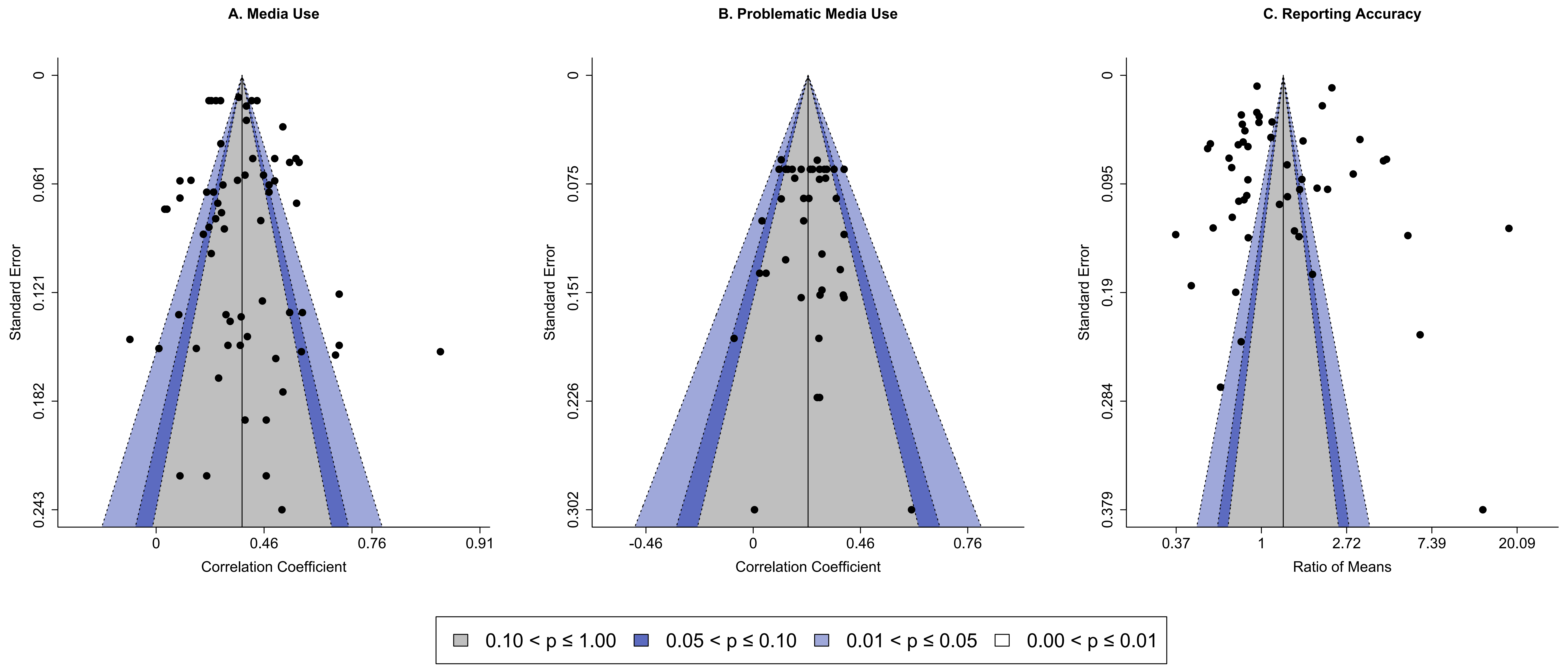
The raw and processed data are available on the Open Science Framework website (https://osf.io/dhx48/). These data include all extracted effect sizes, study-descriptives, and descriptive statistics. In cases where raw data was provided by study authors, as with all included studies, we only provide the necessary descriptive statistics and effective sizes used to compute the summary statistics in the meta-analyses, and do not share these original authors' data. The data have been assigned a unique identifier: 10.17605/OSF.IO/JS6YE

Code Availability

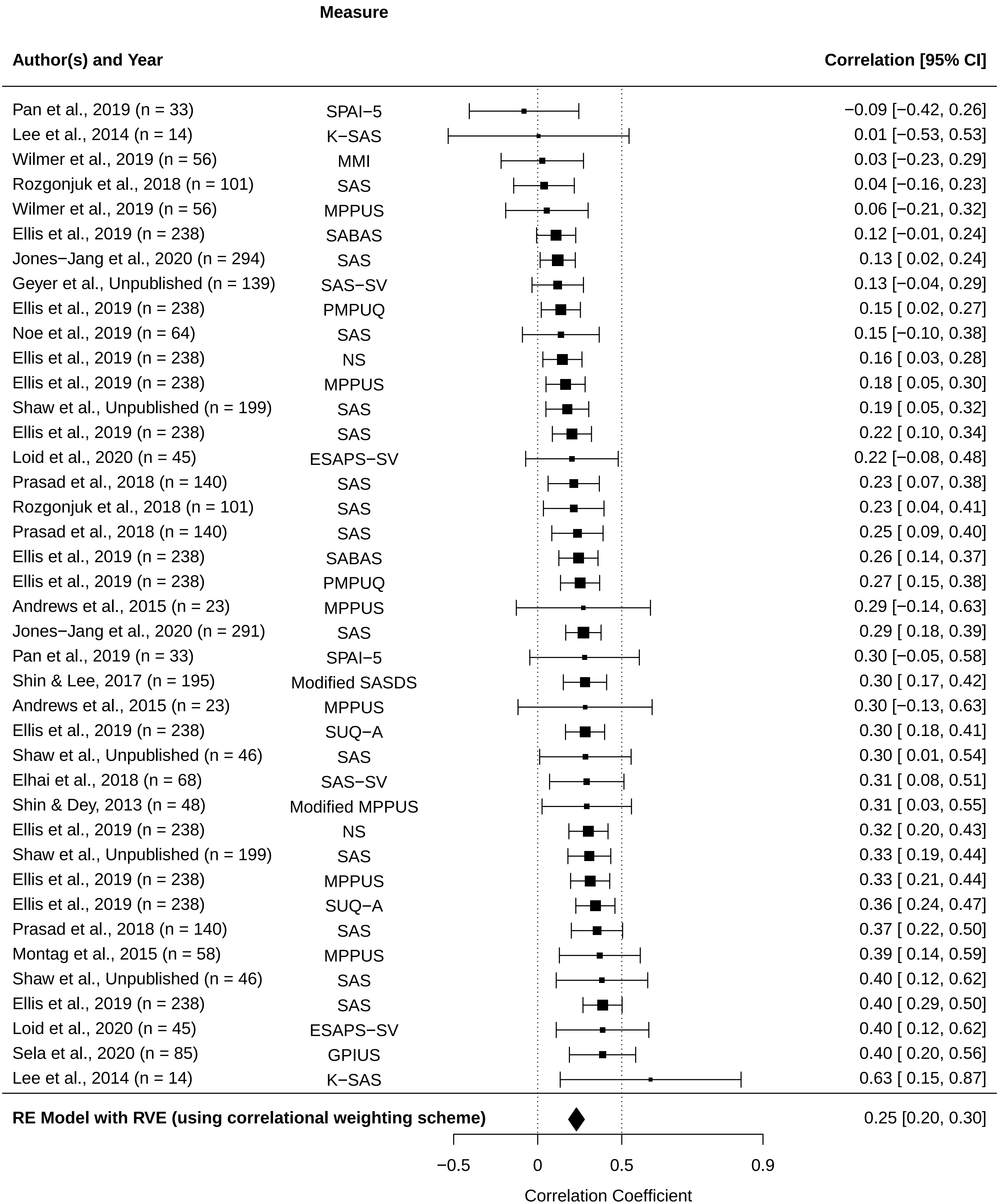
The code (written in the R statistical language) used to analyse the relevant data is provided on the Open Science Framework website (https://osf.io/dhx48/). All materials needed to reproduce the analyses are available at this link.

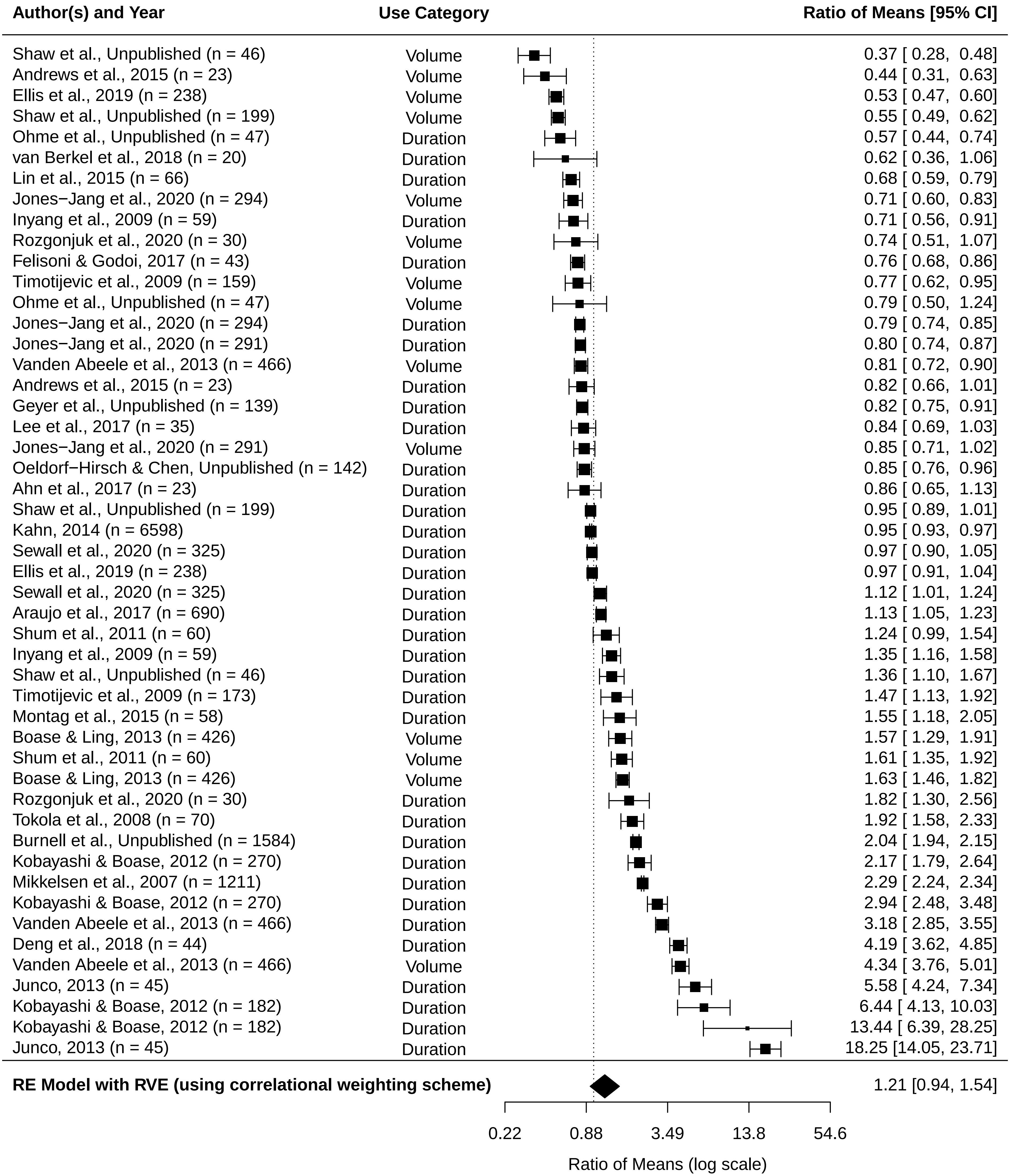

Records identified from database searches
Scopus (n = 3,574)
PubMed (n = 1,548)
PsychInfo (n = 805)
ACM Digital Library (n = 4,435)
Communication & Mass Media Complete (n = 82)
Proquest Dissertations & Theses A & I (n = 1,176)
Additional records from other sources
Manual journal searches (n = 502)
Public calls (n = 10)
Total records identified (n = 12,132)
Duplicates removed (n = 1,326)
Records after duplicates removed (n = 10,806)
Records excluded based on title or abstract (n = 10,582)
Records after irrelevant titles or abstracts removed (n = 224)
Records excluded based on full-text (n = 192)
Eligible records included (n = 32)
Additional records from reference lists (n = 6)
Eligible records included (n = 38)
Additional records from citation lists (n = 9)
Records included in meta analysis (n = 47)


Author(s) and Year
Medium
Correlation [95% CI]
Katapally and Chu, 2019 (n = 49) Phone −0.12 [−0.39, 0.17]
Shaw et al., Unpublished (n = 46) Phone 0.01 [−0.28, 0.30]
Kobayashi & Boase, 2012 (n = 182) Phone 0.04 [−0.11, 0.18]
Kobayashi & Boase, 2012 (n = 182) Phone 0.05 [−0.10, 0.19]
Inyang et al., 2009 (n = 59) Phone 0.10 [−0.16, 0.35]
Andrews et al., 2015 (n = 23) Phone 0.11 [−0.32, 0.50]
Jones−Jang et al., 2020 (n = 291) Phone 0.11 [−0.01, 0.22]
Rosen et al., 2018 (n = 216) Phone 0.11 [−0.02, 0.24]
Jones−Jang et al., 2020 (n = 294) Phone 0.16 [0.05, 0.27]
Shaw et al., Unpublished (n = 46) Phone 0.18 [−0.11, 0.45]
Newell, Unpublished (n = 130) Social Media 0.22 [0.04, 0.37]
Ahn et al., 2017 (n = 23) Phone 0.23 [−0.20, 0.59]
Ellis et al., 2019 (n = 238) Phone 0.23 [0.11, 0.35]
Ernala et al., 2020 (n = 5000) Social Media 0.24 [0.21, 0.27]
Oeldorf−Hirsch & Chen, Unpublished (n = 142) Phone 0.24 [0.08, 0.39]
Ernala et al., 2020 (n = 5000) Social Media 0.25 [0.22, 0.28]
Rosen et al., 2018 (n = 104) Phone 0.25 [0.06, 0.42]
Ellis et al., 2019 (n = 238) Phone 0.26 [0.14, 0.37]
Timotijevic et al., 2009 (n = 159) Phone 0.27 [0.12, 0.41]
Ernala et al., 2020 (n = 5000) Social Media 0.27 [0.24, 0.30]
Shaw et al., Unpublished (n = 199) Phone 0.28 [0.14, 0.40]
Yuan et al., 2019 (n = 38) Phone 0.28 [−0.04, 0.55]
Ernala et al., 2020 (n = 5000) Social Media 0.29 [0.26, 0.32]
Araujo et al., 2017 (n = 690) Internet 0.29 [0.22, 0.36]
Timotijevic et al., 2009 (n = 173) Phone 0.29 [0.15, 0.42]
Kobayashi & Boase, 2012 (n = 270) Phone 0.30 [0.19, 0.40]
Geyer et al., Unpublished (n = 139) Phone 0.31 [0.15, 0.45]
Inyang et al., 2009 (n = 59) Phone 0.31 [0.06, 0.53]
Ohme et al., Unpublished (n = 47) Phone 0.32 [0.04, 0.56]
Wilmer et al., 2019 (n = 56) Phone 0.33 [0.07, 0.55]
Jones−Jang et al., 2020 (n = 294) Phone 0.36 [0.26, 0.46]
Kahn, 2014 (n = 6598) Gaming 0.36 [0.34, 0.39]
Ohme et al., Unpublished (n = 47) Phone 0.37 [0.09, 0.59]
Montag et al., 2015 (n = 58) Phone 0.37 [0.13, 0.58]
Rozgonjuk et al., 2020 (n = 30) Social Media 0.39 [0.03, 0.66]
Sewall et al., 2020 (n = 325) Phone 0.39 [0.29, 0.48]
Scharkow, 2016 (n = 3401) Internet 0.40 [0.37, 0.42]
Burnell et al., Unpublished (n = 1584) Social Media 0.40 [0.35, 0.44]
Yuan et al., 2019 (n = 50) Phone 0.40 [0.14, 0.61]
Ernala et al., 2020 (n = 5000) Social Media 0.42 [0.39, 0.44]
Vanden Abeele et al., 2013 (n = 466) Phone 0.42 [0.34, 0.49]
Ernala et al., 2020 (n = 5000) Social Media 0.44 [0.41, 0.46]
Burke et al., 2010 (n = 155) Social Media 0.45 [0.31, 0.57]
Lin et al., 2015 (n = 66) Phone 0.46 [0.24, 0.63]
Sewall et al., 2020 (n = 325) Phone 0.46 [0.37, 0.54]
Andrews et al., 2015 (n = 23) Phone 0.47 [0.07, 0.74]
Rozgonjuk et al., 2020 (n = 30) Social Media 0.47 [0.13, 0.71]
Ellis et al., 2019 (n = 238) Phone 0.48 [0.38, 0.57]
Kobayashi & Boase, 2012 (n = 270) Phone 0.48 [0.38, 0.57]
Jones−Jang et al., 2020 (n = 291) Phone 0.50 [0.41, 0.58]
Vanden Abeele et al., 2013 (n = 466) Phone 0.50 [0.43, 0.57]
Felisoni & Godoi, 2017 (n = 43) Phone 0.50 [0.24, 0.70]
van Berkel et al., 2018 (n = 20) Phone 0.52 [0.11, 0.78]
Mikkelsen et al., 2007 (n = 1211) Computer 0.53 [0.49, 0.57]
Lee et al., 2017 (n = 35) Phone 0.53 [0.24, 0.73]
Boase & Ling, 2013 (n = 426) Phone 0.55 [0.48, 0.61]
Shum et al., 2011 (n = 60) Phone 0.55 [0.34, 0.71]
Vanden Abeele et al., 2013 (n = 466) Phone 0.57 [0.51, 0.63]
Shaw et al., Unpublished (n = 199) Phone 0.57 [0.47, 0.66]
Boase & Ling, 2013 (n = 426) Phone 0.58 [0.51, 0.64]
Junco, 2013 (n = 45) Social Media 0.59 [0.35, 0.75]
Shum et al., 2011 (n = 60) Phone 0.59 [0.40, 0.73]
Deng et al., 2018 (n = 44) Phone 0.68 [0.48, 0.81]
Singh & Jain, 2017 (n = 47) Phone 0.69 [0.50, 0.82]
Tokola et al., 2008 (n = 70) Phone 0.69 [0.54, 0.80]
Junco, 2013 (n = 45) Social Media 0.87 [0.77, 0.92]
RE Model with RVE (using correlational weighting scheme) 0.38 [0.33, 0.42]
−0.4
0
0.5
0.9
Correlation Coefficient



A. Media Use
B. Problematic Media Use
C. Reporting Accuracy
Standard Error
0
0.061
0.121
0.182
0.243
0
0.46
0.76
0.91
Correlation Coefficient
0
0.075
0.151
0.226
0.302
-0.46
0
0.46
0.76
Correlation Coefficient
0
0.095
0.19
0.284
0.379
0.37
1
2.72
7.39
20.09
Ratio of Means
0.10 < p ≤ 1.00
0.05 < p ≤ 0.10
0.01 < p ≤ 0.05
0.00 < p ≤ 0.01


Measure
Author(s) and Year
Correlation [95% CI]
Pan et al., 2019 (n = 33) SPAI−5 −0.09 [−0.42, 0.26]
Lee et al., 2014 (n = 14) K−SAS 0.01 [−0.53, 0.53]
Wilmer et al., 2019 (n = 56) MMI 0.03 [−0.23, 0.29]
Rozgonjuk et al., 2018 (n = 101) SAS 0.04 [−0.16, 0.23]
Wilmer et al., 2019 (n = 56) MPPUS 0.06 [−0.21, 0.32]
Ellis et al., 2019 (n = 238) SABAS 0.12 [−0.01, 0.24]
Jones−Jang et al., 2020 (n = 294) SAS 0.13 [0.02, 0.24]
Geyer et al., Unpublished (n = 139) SAS−SV 0.13 [−0.04, 0.29]
Ellis et al., 2019 (n = 238) PMPUQ 0.15 [0.02, 0.27]
Noe et al., 2019 (n = 64) SAS 0.15 [−0.10, 0.38]
Ellis et al., 2019 (n = 238) NS 0.16 [0.03, 0.28]
Ellis et al., 2019 (n = 238) MPPUS 0.18 [0.05, 0.30]
Shaw et al., Unpublished (n = 199) SAS 0.19 [0.05, 0.32]
Ellis et al., 2019 (n = 238) SAS 0.22 [0.10, 0.34]
Loid et al., 2020 (n = 45) ESAPS−SV 0.22 [−0.08, 0.48]
Prasad et al., 2018 (n = 140) SAS 0.23 [0.07, 0.38]
Rozgonjuk et al., 2018 (n = 101) SAS 0.23 [0.04, 0.41]
Prasad et al., 2018 (n = 140) SAS 0.25 [0.09, 0.40]
Ellis et al., 2019 (n = 238) SABAS 0.26 [0.14, 0.37]
Ellis et al., 2019 (n = 238) PMPUQ 0.27 [0.15, 0.38]
Andrews et al., 2015 (n = 23) MPPUS 0.29 [−0.14, 0.63]
Jones−Jang et al., 2020 (n = 291) SAS 0.29 [0.18, 0.39]
Pan et al., 2019 (n = 33) SPAI−5 0.30 [−0.05, 0.58]
Shin & Lee, 2017 (n = 195) Modified SASDS 0.30 [0.17, 0.42]
Andrews et al., 2015 (n = 23) MPPUS 0.30 [−0.13, 0.63]
Ellis et al., 2019 (n = 238) SUQ−A 0.30 [0.18, 0.41]
Shaw et al., Unpublished (n = 46) SAS 0.30 [0.01, 0.54]
Elhai et al., 2018 (n = 68) SAS−SV 0.31 [0.08, 0.51]
Shin & Dey, 2013 (n = 48) Modified MPPUS 0.31 [0.03, 0.55]
Ellis et al., 2019 (n = 238) NS 0.32 [0.20, 0.43]
Shaw et al., Unpublished (n = 199) SAS 0.33 [0.19, 0.44]
Ellis et al., 2019 (n = 238) MPPUS 0.33 [0.21, 0.44]
Ellis et al., 2019 (n = 238) SUQ−A 0.36 [0.24, 0.47]
Prasad et al., 2018 (n = 140) SAS 0.37 [0.22, 0.50]
Montag et al., 2015 (n = 58) MPPUS 0.39 [0.14, 0.59]
Shaw et al., Unpublished (n = 46) SAS 0.40 [0.12, 0.62]
Ellis et al., 2019 (n = 238) SAS 0.40 [0.29, 0.50]
Loid et al., 2020 (n = 45) ESAPS−SV 0.40 [0.12, 0.62]
Sela et al., 2020 (n = 85) GPIUS 0.40 [0.20, 0.56]
Lee et al., 2014 (n = 14) K−SAS 0.63 [0.15, 0.87]
RE Model with RVE (using correlational weighting scheme) 0.25 [0.20, 0.30]
−0.5
0
0.5
0.9
Correlation Coefficient


Author(s) and Year
Use Category
Ratio of Means [95% CI]
Shaw et al., Unpublished (n = 46) Volume 0.37 [0.28, 0.48]
Andrews et al., 2015 (n = 23) Volume 0.44 [0.31, 0.63]
Ellis et al., 2019 (n = 238) Volume 0.53 [0.47, 0.60]
Shaw et al., Unpublished (n = 199) Volume 0.55 [0.49, 0.62]
Ohme et al., Unpublished (n = 47) Duration 0.57 [0.44, 0.74]
van Berkel et al., 2018 (n = 20) Duration 0.62 [0.36, 1.06]
Lin et al., 2015 (n = 66) Duration 0.68 [0.59, 0.79]
Jones−Jang et al., 2020 (n = 294) Volume 0.71 [0.60, 0.83]
Inyang et al., 2009 (n = 59) Duration 0.71 [0.56, 0.91]
Rozgonjuk et al., 2020 (n = 30) Volume 0.74 [0.51, 1.07]
Felisoni & Godoi, 2017 (n = 43) Duration 0.76 [0.68, 0.86]
Timotijevic et al., 2009 (n = 159) Volume 0.77 [0.62, 0.95]
Ohme et al., Unpublished (n = 47) Volume 0.79 [0.50, 1.24]
Jones−Jang et al., 2020 (n = 294) Duration 0.79 [0.74, 0.85]
Jones−Jang et al., 2020 (n = 291) Duration 0.80 [0.74, 0.87]
Vanden Abeele et al., 2013 (n = 466) Volume 0.81 [0.72, 0.90]
Andrews et al., 2015 (n = 23) Duration 0.82 [0.66, 1.01]
Geyer et al., Unpublished (n = 139) Duration 0.82 [0.75, 0.91]
Lee et al., 2017 (n = 35) Duration 0.84 [0.69, 1.03]
Jones−Jang et al., 2020 (n = 291) Volume 0.85 [0.71, 1.02]
Oeldorf−Hirsch & Chen, Unpublished (n = 142) Duration 0.85 [0.76, 0.96]
Ahn et al., 2017 (n = 23) Duration 0.86 [0.65, 1.13]
Shaw et al., Unpublished (n = 199) Duration 0.95 [0.89, 1.01]
Kahn, 2014 (n = 6598) Duration 0.95 [0.93, 0.97]
Sewall et al., 2020 (n = 325) Duration 0.97 [0.90, 1.05]
Ellis et al., 2019 (n = 238) Duration 0.97 [0.91, 1.04]
Sewall et al., 2020 (n = 325) Duration 1.12 [1.01, 1.24]
Araujo et al., 2017 (n = 690) Duration 1.13 [1.05, 1.23]
Shum et al., 2011 (n = 60) Duration 1.24 [0.99, 1.54]
Inyang et al., 2009 (n = 59) Duration 1.35 [1.16, 1.58]
Shaw et al., Unpublished (n = 46) Duration 1.36 [1.10, 1.67]
Timotijevic et al., 2009 (n = 173) Duration 1.47 [1.13, 1.92]
Montag et al., 2015 (n = 58) Duration 1.55 [1.18, 2.05]
Boase & Ling, 2013 (n = 426) Volume 1.57 [1.29, 1.91]
Shum et al., 2011 (n = 60) Volume 1.61 [1.35, 1.92]
Boase & Ling, 2013 (n = 426) Volume 1.63 [1.46, 1.82]
Rozgonjuk et al., 2020 (n = 30) Duration 1.82 [1.30, 2.56]
Tokola et al., 2008 (n = 70) Duration 1.92 [1.58, 2.33]
Burnell et al., Unpublished (n = 1584) Duration 2.04 [1.94, 2.15]
Kobayashi & Boase, 2012 (n = 270) Duration 2.17 [1.79, 2.64]
Mikkelsen et al., 2007 (n = 1211) Duration 2.29 [2.24, 2.34]
Kobayashi & Boase, 2012 (n = 270) Duration 2.94 [2.48, 3.48]
Vanden Abeele et al., 2013 (n = 466) Duration 3.18 [2.85, 3.55]
Deng et al., 2018 (n = 44) Duration 4.19 [3.62, 4.85]
Vanden Abeele et al., 2013 (n = 466) Volume 4.34 [3.76, 5.01]
Junco, 2013 (n = 45) Duration 5.58 [4.24, 7.34]
Kobayashi & Boase, 2012 (n = 182) Duration 6.44 [4.13, 10.03]
Kobayashi & Boase, 2012 (n = 182) Duration 13.44 [6.39, 28.25]
Junco, 2013 (n = 45) Duration 18.25 [14.05, 23.71]
RE Model with RVE (using correlational weighting scheme) 1.21 [0.94, 1.54]
0.22
0.88
3.49
13.8
54.6
Ratio of Means (log scale)

Moderator	*k*	*r*	*F*	95% CI	*p*
Population*			0.42	-	0.745
Adults	38	0.41	-	[0.33, 0.48]	<0.001
General	4	0.37	-	[0.19, 0.53]	0.023
Student	15	0.37	-	[0.26, 0.48]	<0.001
Unknown	7	0.35	-	[0.23, 0.46]	<0.001
Sampling category			0.90	-	0.423
Data donation	16	0.35	-	[0.29, 0.40]	<0.001
Direct tracking	30	0.36	-	[0.26, 0.46]	<0.001
Supplied data	20	0.40	-	[0.33, 0.47]	<0.001
Logging collection method†			1.4	-	0.279
Built in tool	16	0.35	-	[0.29, 0.40]	<0.001
Custom built tool	15	0.29	-	[0.15, 0.42]	<0.001
Operator or platform data	20	0.40	-	[0.33, 0.47]	<0.001
Third party tool	14	0.45	-	[0.27, 0.60]	<0.001

Moderator	*k*	*R*	*Exp(β)*	*F*	95% CI	*p*
Population*			1.01	-	-	0.969
Adults	32	1.22	-	-	[0.89, 1.69]	0.196
Student	11	1.24	-	-	[0.64, 2.40]	0.468
Sampling category			-	3.4	-	0.066
Data donation	14	1.24	-	-	[0.66, 1.21]	0.412
Direct tracking	24	1.31	-	-	[0.84, 2.04]	0.214
Supplied data	11	1.46	-	-	[1.03, 2.08]	0.039
Logging collection method			-	2.85	-	0.074
Built in tool	14	0.89	-	-	[0.66, 1.21]	0.412
Custom built tool	14	0.95	-	-	[0.60, 1.51]	0.827
Operator or platform data	11	1.46	-	-	[1.03, 2.08]	0.039
Third party tool	10	1.91	-	-	[0.81, 4.50]	0.113

Descriptor	*k* (%)		
	Media usage	Problematic usage	Reporting accuracy
Population			
Adolescents	2 (3.03)	1 (2.50)	2 (4.08)
Adults	38 (57.58)	25 (62.50)	32 (65.31)
General	4 (6.06)	2 (5.00)	3 (6.12)
Student	15 (22.73)	12 (30.00)	11 (22.45)
Unknown	7 (10.61)	-	1 (2.04)
Sampling category			
Data donation	16 (24.24)	18 (45.00)	14 (28.57)
Direct tracking	30 (45.46)	22 (55.00)	24 (48.98)
Supplied data	20 (30.30)	-	11 (22.45)
Logging collection method			
Built in tool	16 (24.24)	18 (45.00)	14 (28.57)
Custom built tool	15 (22.73)	12 (30.00)	14 (28.57)
Operator or platform data	20 (30.30)	-	11 (22.45)
Third party tool	14 (21.21)	10 (25.00)	10 (20.41)
Other*	1 (1.52)	-	-