Computers in Human Behavior Reports 4 (2021) 100142

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Somebody's Watching Me: Smartphone Use Tracking and Reactivity

Roland Toth [a,*], Tatiana Trifonova [b]

[a] Freie Universität Berlin, Germany
[b] Max-Planck-Institute for Human Development, Germany

ARTICLE INFO

Keywords:
smartphone
Tracking
Observation
Reactivity
Bias

ABSTRACT

Like all media use, smartphone use is mostly being measured retrospectively with self-reports. This leads to misjudgments due to subjective aggregations and interpretations that are necessary for providing answers. Tracking is regarded as the most advanced, unbiased, and precise method for observing smartphone use and therefore employed as an alternative. However, it remains unclear whether people possibly alter their behavior because they know that they are being observed, which is called reactivity. In this study, we investigate first, whether smartphone and app use duration and frequency are affected by tracking; second, whether effects vary between app types; and third, how long effects persist. We developed an Android tracking app and conducted an anonymous quasi-experiment with smartphone use data from 25 people over a time span of two weeks. The app gathered not only data that were produced after, but also prior to its installation by accessing an internal log file on the device. The results showed that there was a decline in the average duration of app use sessions within the first seven days of tracking. Instant messaging and social media app use duration show similar patterns. We found no changes in the average frequency of smartphone and app use sessions per day. Overall, reactivity effects due to smartphone use tracking are rather weak, which speaks for the method's validity. We advise future researchers to employ a larger sample and control for external influencing factors so reactivity effects can be identified more reliably.

Smartphones are used by a large portion of society and have become pervasive in everyday media use (Newzoo, 2019). The device also found its place in media use research and introduced new methodological challenges (e.g., Bayer, Campbell, & Ling, 2016; Harari et al., 2019; Kaye, Orben, Ellis, Hunter, & Houghton, 2020). As with many other subjects of interest, asking people about media use retrospectively in questionnaires is the dominant data collection tool in the social sciences (Griffioen, Rooij, Lichtwarck-Aschoff, & Granic, 2020; e.g.,; Guthrie, 2010). However, with technological advancements, more sophisticated assessment methods were developed and employed.

Tracking is the technologically assisted, automatic, passive, and precise observation of behavior while or shortly after it occurs. It can take different forms – for example, log file analysis is used for assessing websites visited (Scharkow, 2016) and phone system logs document the use of smartphones in general and specific apps, which is a function that is already implemented in operating systems like Android (Harari et al., 2019). Research consistently showed that questionnaire data on the frequency and duration of media use differ from such tracking data to a worrying extent, which also applies to smartphone use (for an overview, see: Parry et al., 2021). This suggests that questionnaires do not represent media use adequately – given the assumption that tracking is the objective baseline all other assessment methods need to be checked against. However, this is not necessarily true, as people tend to alter their behavior when they are aware of being observed. In psychology, this effect is known as reactivity (Gittelsohn, Shankar, West, Ram, & Gnywali, 1997).

Smartphone tracking data may therefore also be biased because people react to being observed in the first place. Then again, smartphone use is oftentimes initiated habitually, and thus, unconsciously (e.g., Schnauber-Stockmann & Naab, 2019), which speaks against this conclusion. Therefore, we investigate the following question: *Are smartphone use tracking data biased due to reactivity?*

To address this issue, we conducted a quasi-experimental, anonymous tracking study with 25 Android users for two weeks. Tracking was performed with an Android app that was developed specifically for this study. It does not only capture recent use after installation, but also past use ocurring prior to it. This allowed us to juxtapose and compare unbiased use before and potentially biased use after installation.

* Corresponding author. Garystrasse 55, 14195, Berlin, Germany.
E-mail address: roland.toth@fu-berlin.de (R. Toth).

https://doi.org/10.1016/j.chbr.2021.100142
Received 27 June 2021; Received in revised form 15 September 2021; Accepted 23 September 2021
Available online 29 September 2021

1. Smartphone use measures

In questionnaires, media use is mostly assessed with measures of frequency and duration of use. Frequency is operationalized in terms of subjective assessments, for example, *never* to *all the time* (Marty-Dugas, Ralph, Oakman, & Smilek, 2018), days per week (Lopez-Fernandez, Männikkö, Kääriäinen, Griffiths, & Kuss, 2018) or as an absolute frequency within a specified time frame (van Berkel et al., 2019). Duration is operationalized with regard to fixed time frames and a frame of reference, for example, *less than 30 min* to *more than 3 h per day* (Chang et al., 2018) or, similarly to frequency, in minutes per day or week (Lemola, Perkinson-Gloor, Brand, Dewald-Kaufmann, & Grob, 2014). However, there is a major problem with this approach. Questionnaire data on media use lack absolute ground truth (van Berkel et al., 2019) as people have to aggregate lots of information from a possibly long period of time and countless use episodes in order to answer such questions. This goes along with multiple cognitive issues. Respondents need to 1) understand the question properly, 2) recall respective behavior, 3) infere and estimate the frequency or duration of use, 4) allocate their estimation within the scale of the question, and might 5) ultimately still answer in a biased way due to social desirability (Schwarz & Oyserman, 2001). As a result, answers are often quite different from the information that is actually of interest and results in under- or overestimations (e.g., Naab, Karnowski, & Schlütz, 2018; Valkenburg & Peter, 2013, p. 200). This is especially problematic when assessing smartphone use due to the typically high frequency and short duration of use episodes. With technological advancements, more objective, passive observation techniques were developed.

Passive observation is regarded as the most valid method for assessing media use (Vandewater & Lee, 2009, p. 9), as it enables the collection of behavioral data without the need of subjective assessment. Tracking can be considered its modern implementation that does not require researchers or obvious observation tools like cameras to be present. Study participants are only made aware of the observation setting in the beginning of the study, but not during data collection. Methodologically, it is therefore a blend of an "undisguised naturalistic observation, where the participants are made aware of the researcher presence and monitoring of their behavior" and a "disguised naturalistic observation" where researchers "make their observations as unobtrusively as possible so that participants are not aware that they are being studied" (Price, Jhangiani, Chiang, Leighton, & Cuttler, 2017, p. 121). As opposed to advanced survey methods like time diaries (Thulin & Vilhelmson, 2007) or the Experience Sampling Method (ESM) (Csikszentmihalyi, 2014), tracking is supposed to be unobtrusive and independent of people's own perception and interpretation and can therefore be considered the best option for assessing quantitative metrics of media use, namely duration and frequency (e.g., Boase & Ling, 2013; Scharkow, 2016).

Phone use tracking was not only possible since the smartphone's release. For example, network providers always kept track of customers' incoming and outgoing calls and messages and such data were used in research before (e.g., Cohen & Lemish, 2003). With the advent of smartphone technology, however, researchers soon recognized the device's potential not only for interpersonal and mass communication per se, but also from a methodological perspective, as it enabled the automated collection of more refined use data (Raento, Oulasvirta, & Eagle, 2009). Since the iPhone was introduced in 2007 and marked the inception of the smartphone as we know it today (Jackson, 2018), tracking was employed in many studies for collecting precise data on smartphone use with regard to total use and the use of specific applications. These data were then used to check relationships with other concepts of interest, such as personality traits, college course performance, and social connectedness (e.g., Andrews, Ellis, Shaw, & Piwek, 2015; Harari et al., 2019; Rosen et al., 2018; Walker, Koh, Wollersheim, & Liamputtong, 2015). Among others, it was shown that smartphone use is characterized by frequent and short use episodes because the device is basically permanently active and in use (e.g., Klimmt, Hefner, Reinecke, Rieger, & Vorderer, 2019; Rosen et al., 2018).

While smartphone use tracking is used in studies for various purposes and applied reasonably, it is rarely questioned or evaluated regarding its validity. One of the very few studies that explicitly investigated methodological implications of methods that take advantage of mobile devices dealt with ESM. ESM is an advanced survey method that focuses on in-situ measurement of situational and emotional contexts while or shortly after they occur, minimizing the risk of retrospection biases (Csikszentmihalyi, 2014). In a comparison between survey and ESM data on time use, Sonnenberg, Riediger, Wrzus, and Wagner (2011) rightfully noted that interpretations of the differences between both methods implicitly assume the superiority of ESM, although in principle it might well be possible that ESM data are actually more error-prone than survey data. However, as the authors argue, it is hard to falsify this possibility as there is a lack of "empirical evidence on the performance of ESM" (p. 24). Analogous to this issue, we would hereby like to provide such evidence for the tracking method. While many studies have shown that questionnaire data on the frequency and duration of media use differ from respective tracking data, much less attention was devoted to the more general problem of the validity of tracking data themselves – namely, whether even tracking data are valid representations of the concepts they are supposed to measure. In other words, the question is not whether tracking is the most accurate method for time-related use measures in theory (which it should be, due to the lack of participants' active involvement) but whether the method produces biased data to begin with.

2. Reactivity

Reactivity can constitute a potential threat to the validity of research results. It usually occurs "when actors change their behavior due to the presence of an observer" (Gittelsohn et al., 1997, p. 182).

Several forms of reactivity can be distinguished. Each highlights a particular aspect of behavior change (Barnes, 2010). The most prominent form is social desirability, which especially applies to sensitive topics assessed in questionnaires or interviews. It emerges when participants intentionally demonstrate (allegedly) positive behavior and conceal behavior that they perceive as socially inappropriate (e.g., Fang, Wen, & Prybutok, 2014; Jensen & Hurley, 2005; Krumpal, 2013). Another form is the Hawthorne Effect, which originally suggested that factory workers' productivity increased when observed, regardless of manipulation or experimental condition (Adair, 1984; Barnes, 2010; Lied & Kazandjian, 1998). However, in numerous studies, the Hawthorne Effect is used to refer to any change in participants' behavior and therefore as an equivalent to the term reactivity (Barnes, 2010; McCambridge, Wilson, Attia, Weaver, & Kypri, 2019).

Lots of research on reactivity dealt with subjects such as medical personnel professional behavior (Eckmanns, Bessert, Behnke, Gastmeier, & Rü; Leonard & Masatu, 2006; Mangione-Smith, Elliott, McDonald, & McGlynn, 2002), patients' performance (Berthelot, Nizard, & Maugars, 2019; Bouchet, Guillemin, & Briançon, 1996; Feil, Grauer, Gadbury-Amyot, Kula, & McCunniff, 2002; McCambridge et al., 2019), academic performance (Adair, 1984; Cook, 1962; Haddad, Nation, & Williams, 1975), and voting behavior (Gerber, Green, & Larimer, 2008; Granberg & Holmberg, 1992). An improvement of people's behavior due to observation was also shown concerning indoor air pollution (Barnes, 2010) and electricity consumption (Schwartz, Fischhoff, Krishnamurti, & Sowell, 2013).

There is only little research on reactivity concerning media use. For instance, it was shown that both children and adults alter their television viewing behavior in presence of parents or researchers, respectively (Christakis & Zimmerman, 2009; Otten, Littenberg, & Harvey-Berino, 2010). Also, former Nielsen panelists were used as research subjects in order to rule out reactivity due to the panelists' adjustment to being observed (Taneja & Viswanathan, 2014).

As mentioned before, the smartphone became a crucial and ubiquitous device for digital communication in the course of few years. Also, tracking of smartphone use emerged as an accessible and rich data source that is strongly intertwined with people's everyday lives and should therefore provides access to behavioral data that are not biased by subjective assessment. At first glance, it is tempting to trust in the quality of such data. However, it remains unclear whether the tracking itself leads to a disruption of usual use patterns and collected data therefore do not reflect usual use.

Smartphones offer two layers of use – the use of the device itself and the use of apps, as representations of gratifications and possibilities offered by the device (Schnauber-Stockmann & Naab, 2019; Turkle, 2008). We therefore ask the following research question:

RQ 1: Does smartphone use tracking lead to reactivity concerning smartphone and app use?

Some media uses are more sensitive than others. For example, it was shown that social desirability affects self-reports about the use of pornography (Valkenburg & Peter, 2013, p. 200). Chances are that uses of certain app types (e.g., social media, dating apps, gaming, entertainment) might be more affected by reactivity than others due to social desirability. We are not aware of existing research that investigated perceived social desirability, intimacy or privacy concerns for specific app types. However, it was shown that smartphone users perceive permissions granted for accessing the phone features multimedia storage, SMS, camera, microphone, and GPS sensor as particularly sensitive (Furini, Mirri, Montangero, & Prandi, 2020). While many apps nowadays require access to these features, they are most prominently requested by instant messaging (e.g., WhatsApp), social media (e.g., Facebook), and gaming (e.g., Pokémon Go) apps.

We do not yet know whether these app types are associated with more reactivity while being tracked due to the sensible permissions, or even less reactivity due to the desensitization regarding privacy invasion experienced by users of these app types anyways. Considering that investigating app types was shown to be more feasible than investigating specific apps (David, Roberts, & Christenson, 2018, p. 271), we therefore pose the following research question:

RQ 2: Is reactivity regarding the use of instant messaging, social media, and gaming apps different to overall reactivity?

Lastly, research suggests that reactivity effects decrease with time due to participants' habituation to the setting (Cousens, Kanki, Toure, Diallo, & Curtis, 1996; Harris, 1982; Wu, 2013). Some studies demonstrate that the change in participant behavior occurs on the first day of observation and fades away over the course of some days (Gittelsohn et al., 1997; Leonard & Masatu, 2006; Schmitz, Stanat, Sang, & Tasche, 1996). Therefore, reactivity effects are more likely at the beginning of the observation period than later (McCambridge, Witton, & Elbourne, 2014) which does not appear to vary with observational duration and frequency (Harris, 1982). However, there is also evidence for reactivity effects with no signs of habituation whatsoever (Harris, 1982; Kypri, Langley, Saunders, & Cashell-Smith, 2007; Murray, Swan, Kiryluk, & Clarke, 1988).

The smartphone is a highly versatile device with many and short use episodes which may quickly distract people from the observation setting. Therefore, it is likely that potential smartphone use tracking reactivity effects do not last long. How long the time frame of reactivity is, though, is up for debate. If the app types mentioned in RQ 2 are affected by reactivity, we will investigate the persistence of these effects, too. This leads to our last research question:

RQ 3: How long do reactivity effects on smartphone and app use persist?

Duration and frequency were shown to measure different elements of smartphone use quantity (Andrews et al., 2015; Wilcockson, Ellis, & Shaw, 2018). For this reason, we investigate all research questions with regard to both duration and frequency.

3. Method

In this study, we used phone system logs for data collection. Still, we use the term tracking for better readability and as a reference to passive observation methods in general.

For answering our research questions, we developed an Android app, *A Tricky Tracker* (ATT), that collected data that were produced before as well as after its installation. This way, we could compare them to each other and derive insights on behavioral alterations caused by tracking. Due to the complex structure of the data involved, our overall aim was to exclude data only when absolutely necessary and separately for each individual analysis, so we could leverage all available data the best way possible. We report all data exclusions, all manipulations, and all measures.

3.1. Procedure

ATT accessed a log file implemented within the Android operating system which stores all actions occurring on the device, so-called *events*. Events are further categorized by *event types*, which are listed in the official Android documentation (Google, 2019c). See Table 1 for an overview of event types relevant for our analysis.[1] We describe the use of each of these event types for specific measures in the section Measures.

After installation, ATT regularly synchronized event data from said log file and produced one data set each that contained all events that were captured since the last synchronization. Additionally, it actively and regularly registered whether the device was locked, unlocked, shut down, or booted, with a respective time stamp, as this information was not logged in the form of events on devices running Android 8 and lower. In these versions, it is therefore impossible to identify precisely when the device was locked, unlocked, shut down, or booted in the past. Devices running Android 9, however, logged locking and unlocking instances without the need of additional implementation (Google, 2019c). For this reason, we had to limit ourselves to data from Android 9 devices.

All resulting data were saved in the format *.json*. For this study, ATT was set up in a way that all data were regularly synchronized with a virtual Linux server. We used pseudonymized, unique identifiers for each individual device so we could tell them apart in the analysis while at the same time preserving participants' anonymity.

However, the benefit of ATT did not only lie in capturing current use during tracking. For this study, it was of crucial importance to assess data that were definitely not biased due to observation and that were comparable to data produced after participation began. The Android log file typically contains data from up to two weeks in the past. Therefore,

Table 1
Event types used for data preparation.

Event type	Description
1	Activity resumed
2	Activity paused
17	Keyguard shown
18	Keyguard hidden
26	Device shutdown
27	Device booted

[1] Keyguard corresponds to the phone lock screen (Google, 2020b). In Android versions below 9, *activity resumed* corresponds to *activity moved to the foreground* and *activity paused* to *activity moved to the background*.

right after participants installed ATT and accepted all policies, ATT accessed event data that were produced earlier. Participants were informed about this feature before and during installation, but given the opportunity to cancel the installation and participation anytime. However, as they neither knew of the study itself nor ATT's features before participation, these data can be considered truly free from any kind of methodological bias. As such, they represent the baseline for regular behavior in this study and allow for holding subsequent behavior (after the installation of ATT) against it. While retrospective smartphone use data were used in research before (e.g., David et al., 2018) they were not yet used for investigating reactivity to tracking by juxtaposing them with subsequent, "real-time" tracking data.

The main view of the app contained a timer that counted down from 14 days to let participants know when the study would end. In a menu, participants could view the researchers' contact information and review the data privacy statement.

We used ATT to conduct an anonymous quasi-experiment with 25 participants, which is a similar sample size used in previous smartphone use tracking studies (e.g., Caine, 2016; van Berkel et al., 2019). Data collection took place between December 12, 2019 and January 11, 2020.

3.2. Participants

We were interested in a rather unspecific issue that potentially affects any smartphone user and did therefore not impose any restrictions concerning participants' social characteristics. We recruited a convenience sample through different channels, including the SoSci Panel (SoSci Panel, 2020), survey websites like surveycircle.com, advertisements at the local university, and the researchers' private (social) networks. Due to financial limitations, we could not provide incentives for participation. We would like to note that recruitment for a study that involves tracking methods is aggravated by a high inhibition threshold and effort necessary for the installation of an app and agreeing to being observed for weeks, even when offered an incentive (Andrews et al., 2015).

We set up a dedicated website containing important information about the study, measures of data protection, and a registration form. People who registered automatically received an Email containing a link to an anonymous questionnaire with questions concerning demographic features for sample description, a link to the installation file of ATT (in *.apk* format), and a detailed installation guide. Unfortunately, many more people participated in the survey than ultimately in the tracking procedure. It was not possible for us to link tracking and questionnaire data because this would have made individual identification possible and therefore not complied with anonymity. For this reason, we could not identify which of the completed questionnaires actually belonged to persons who participated in the tracking. The questionnaire results indicated that people interested in participation were 48% female, with a mean age of 36 years ($SD = 44.26$). During installation, participants were again informed about the study procedure and required to accept a data privacy statement. After installation, participants were not notified or interrupted by ATT at all during the period of data collection, eliminating a potential additional source of reactivity (van Ballegooijen et al., 2016). After two weeks of data collection, participants received a notification from the app asking them to uninstall it.

3.3. Measures

3.3.1. Smartphone and app use session

A use session indicated the time span between the first and the last events of a consistent use episode. With regard to smartphone use, event types 18 (keyguard hidden) and 27 (device booted) were considered first events, and event types 17 (keyguard shown) and 26 (device shutdown) last.

With regard to app use, event type 1 (activity resumed) was considered first events, and event types 2 (activity paused), 17 (keyguard shown), and 26 (device shutdown) last. As many apps feature multiple activities (Google, 2019a), any consistent sequence of activities performed within a single app without interruption was considered part of the same app use session.

3.3.2. Duration

We represented the duration of use sessions by calculating the difference between the time stamps of the start and the end of a use session. We calculated duration in seconds, minutes, and hours for different purposes.

3.3.3. Frequency

The frequency of use sessions was the number of use sessions per device occurring within a specified time period.

3.3.4. Time frame

The time frame indicated whether a use session took place before (*0*) or after (*1*) the installation of ATT. To ensure readability, we call the former *pre-installation* and the latter *post-installation* smartphone/app use from here on out. We assigned each smartphone and app use session a time frame by checking whether the time stamp of the start of the session was smaller or larger than the time stamp of the installation.

3.3.5. Day

We assigned each use session an integer that represented the day it took place relative to the installation date of ATT (e.g., *-8* for the eighth day before installation; *5* for the fifth day after installation).

3.3.6. App type

We automatically assigned all apps a type according to the Google Play Store, which was done similarly for iPhone use data before (David et al., 2018). To achieve this, we used the Python library Google-Play-Scraper (JoMingyu, 2020). Apps that could not be categorized were assigned the type "Other."

3.4. Data preparation

We programmed a parser in Python that extracted all relevant data from the .json files and then transformed and merged them into a single data frame. Each row of this data frame represented a single event on one device (e.g., moving a specific app to the foreground). Variables included the Android version of the device, the event type and time stamp of the event, and the package name (Google, 2019b) of the app performing it.

As data collection took place at the end of the calendar year, some data were generated during or between the Christmas holidays and New Year's Eve (December 24–26, December 31 - January 1). It is likely that mobile communication behavior is different during these time spans (Vandewater & Lee, 2009, p. 10). People may use their smartphones more than usual for communicating with family and friends – then again, they might use them less than usual so they can enjoy some quality time with their peers in person. To be on the safe side and to account for the possibility that smartphone use might be affected the day before Christmas, too, we marked all smartphone use sessions that took place between and including December 23 and January 1. As such, we were able to investigate and account for possible noise in the data during analysis.

We then iteratively aggregated and transformed the data frame such that each row represented one use session. Our approach was already applied similarly by Harari et al. (2019). Two data frames were created this way – the first containing *smartphone use* sessions, the second containing *app use* sessions, as defined in the section Measures.

3.4.1. Smartphone use

Data from four participants were excluded from the smartphone use data set. One did not provide pre-installation data at all. One accounted for most exceedingly long, seemingly uninterrupted smartphone use sessions (up to 12 h). Two participated only for (part of) the first day after having installed ATT.

Finally, we excluded all data from the first and last days provided by each participant in order to omit incomplete data for these days, especially with regard to use frequency.

3.4.2. App use

Regarding app use, we assumed that we could use data from all Android versions, as moving an app to the foreground or background is always captured without additional implementation (Google, 2019c). For good measure, we additionally considered the data on screen activity and shutdowns/boots captured by ATT itself.

Following this approach, even data generated after the installation of ATT contained (seemingly) uninterrupted app use sessions that lasted extremely long (e.g., 12 h). Further investigation showed that interruptions of app uses through screen locks and shutdowns were probably not always captured properly. Andrews et al. (2015) faced similar problems. For this reason, we created another version of this data set that only included data from Android 9 devices. Hence, we could use existing event types that indicate locks, unlocks, shutdowns and boots without accessing the checks implemented in ATT for both pre- and post-installation use.

Even then, there were very long, consistent app uses without interruptions (up to 9 h). Those were probably still instances where internal Android mechanisms failed to register screen locks. In the end, we settled for a cutoff value of 5 h, which Andrews et al. (2015) considered very long use. Below that, there were still long use sessions – however, they took place in apps where long, uninterrupted use sessions are reasonable, e.g., Pokémon GO, YouTube, or Twitch.

We excluded events regarding some system-related apps and functions as recommended in previous research (Jones, Ferreira, Hosio, Goncalves, & Kostakos, 2015). We then applied the same transformations and data exclusions already applied to smartphone use, save the device that accounted for most exceedingly long smartphone use sessions. Finally, we merged both into a single data frame for analysis and tagged them with a dedicated variable for distinction.

3.5. Data analysis

RQ 1 deals with the question whether smartphone and app use are affected by participation in tracking. We investigated this question for duration and frequency.

We analyzed smartphone and app use duration on the basis of single use sessions. However, in our data, these were not independent from one another as each was associated with a specific device. This does not meet the assumptions of regression modeling (Field, Miles, & Field, 2012, p. 957). Multilevel modeling was neither applicable, as a higher number of top-level units (around 30) is recommended for this analysis method (Hox & McNeish, 2020). Also, while we expected individual use differences between devices, we were not interested in explaining them in this study. For this reason, we accounted for all variance due to the differences between devices by adding one dummy variable each to multiple regression, resulting in a fixed-effects model. We did the same regarding app types for analyses of app use duration, as each app use session is not only tied to an individual, but also an app type. As it turned out, 10.34% of the data were generated during the time span declared as the holiday season. In order not to discard valuable data, we decided to control for the holidays in analyses of duration instead of omitting them.

As a second step, we investigated possible reactivity concerning the frequency of smartphone and app use. Frequency depends on a time frame of reference, which is why we could not investigate it on the level of individual use sessions the same way we did with duration. Instead, we aggregated the average number of smartphone and app use sessions per day for pre- and post-installation use separately for each participant. Due to the non-normal distribution of the difference between pre- and post-installation use frequencies and the low number of observations/days within the aggregated data set, we performed the non-parametric Wilcoxon signed-rank test for paired samples, which has less strict assumptions than a corresponding *t*-test (Field et al., 2012, p. 957). Due to the necessary aggregations, we excluded all holiday data from analyses of use frequency and we could not control for individuals and app types. This led to yet another problem: Some participants generated data that were assigned to both holidays and non-holidays on the same day, as days were operationalized in relation to the exact installation instance in this study, not actual calender days. Therefore, similarly to the exclusions of the first and last days of data collection per participant, we excluded all corresponding data for frequency analyses as to avoid biases due to partial data removals per day. For consistency, we also applied these same exclusions to duration data for visualizations and descriptive analyses, but not for regression analysis.

RQ 2 asks whether effects found in instant messaging, social media and gaming apps were different to the general findings of RQ 1. Due to the low sample size, we first checked whether these types were among the app types used the longest and the most so that sufficient occurrence was given (see section Results). We employed a similar approach as for RQ 1, but only considered app use data. For minimizing the problem of multiplicity, we calculated false discovery rates (FDR).

RQ 3 is concerned with the persistence of potential reactivity effects. Considering the sample size, we decided to apply the same tests from RQ 1 to two-day-intervals of post-installation use and compare each of them to overall pre-installation use to identify possible patterns of effect changes. We chose two-day-intervals as a compromise between statistical power gained through a higher number of observations and the granularity necessary for identifying changes. Again, we controlled for multiplicity by calculating FDR.

After the exclusion of data from Android versions other than 9 and data cleansing, data from 12 devices were left, which happens to be a common sample size in studies on Computer-Human Interaction (CHI) (Caine, 2016). In total, the data comprised 14,330 smartphone use sessions and 43,053 app use sessions over a time span of up to 23 days. This time frame was shown to be more than sufficient for capturing both typical weekly usage and short, habitual checking behaviors (Wilcockson et al., 2018).

For analyses, coding, and typesetting, we used R [Version 4.0.2; R Core Team (2020)] and the R-packages *ggplot2* [Version 3.3.2; Wickham (2016)], *papaja* [Version 0.1.0.9997; Aust and Barth (2020)], and *tidyverse* [Version 1.3.0; Wickham et al. (2019)]. All data, analysis code and a visualization of individual participants' smartphone and app use over time can be found in the online supplementary material (OSM).

4. Results

On average, participants used their smartphones for 3.47 h (SD = 2.33) and 57.28 times (SD = 40.75) per day, excluding the holiday season. The distribution of session duration was strongly right-skewed (γ_{phone} = 10.35, γ_{apps} = 16.68) as a great majority of use sessions was fairly short (Mdn_{phone} = 46.04 s, Mdn_{apps} = 11.64 s). Therefore, we used median values for visualizing central tendencies with regard to average use session duration per day (see Fig. 1). For all analyses, we applied log transformation to duration, which resulted in a distribution much more similar to a normal distribution. This is a technique that was used in research with similar data before (e.g., van Berkel et al., 2019).

See Fig. 2 for a visualization of smartphone and app use frequencies per day. On Days 11, 12 and 13, only a single participant provided data (excluding data from the holiday season). For this reason, we excluded these days from all analyses of frequency.

Day 5 of app use strongly deviated from all other days concerning duration as well as frequency. Further investigation showed that this is

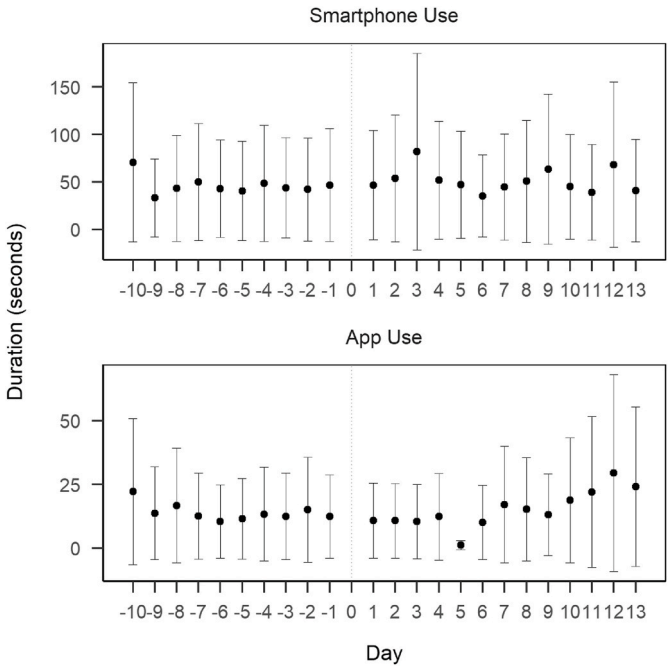

Fig. 1. Median session duration per day relative to the installation date. The dotted line represents the installation instance of ATT. Error bars represent median absolute deviation.

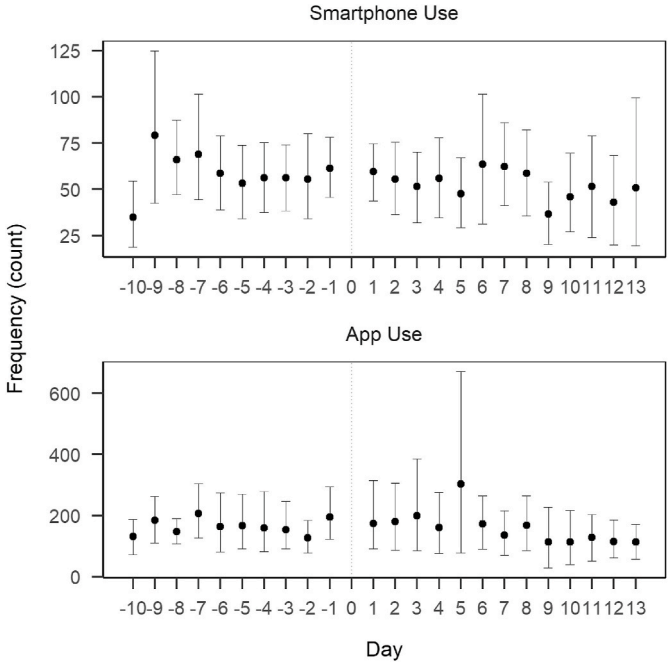

Fig. 2. Mean number of sessions per day relative to the installation date. The dotted line represents the installation instance of ATT. Error bars represent bootstrapped standard errors of the mean.

due to a single participant who registered both the shortest and, at the same time, the majority of all app use sessions that day by far (*Med* = 0.03 s, $n = 2094$). This was not the case for that same participant's smartphone use. While their average app use duration that day only deviated from the mean by 1.45 standard deviations, their average app use frequency deviated by 3.05 standard deviations. For this reason, we decided to exclude this participant's app use data from Day 5.

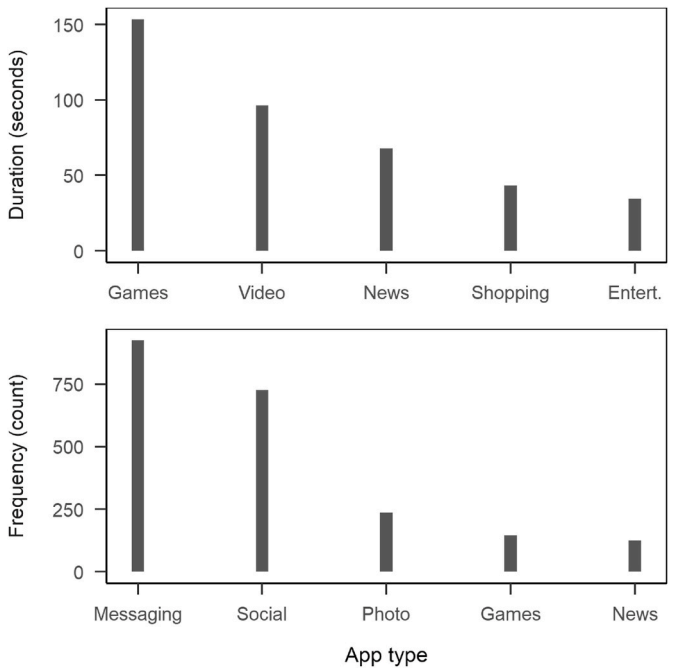

Fig. 3. Top five app types by median duration and mean frequency of use per device. Only apps that were used by at least half of participants were included, so as to increase external validity and decrease outlier influence.

See Fig. 3 for an overview of the five app types used the longest and most frequently. Consistent with previous research (David et al., 2018), gaming and watching videos took the longest and instant messaging and social media were accessed most frequently on average.

4.1. RQ 1

RQ 1 deals with the reactivity of participants due to tracking with regard to smartphone and app use duration and frequency.

Controlling for differences between individuals and the holidays, results showed that post-installation smartphone use session duration was significantly higher than pre-installation smartphone use session duration, $b = 0.09$, 95% CI [0.03, 0.14], $t(14316) = 3.12$, $p = .002$. The effect was rather weak and corresponded to an increase of about 9% (note that duration was log-transformed for analyses, hence the transformation and interpretation of the coefficient as percent change). Post-installation app use session duration, while additionally controlling for app type, was significantly lower than pre-installation app use session duration, $b = -0.04$, 95% CI [−0.08, 0.00], $t(43009) = -2.22$, $p = .027$. The effect corresponded to a decrease of 4%.

Post-installation smartphone use frequency (*Mdn* = 54.12) did not significantly differ from pre-installation smartphone use frequency (*Mdn* = 56.46), $r = -0.23$, $p = .266$. Neither did post-installation app use frequency (*Mdn* = 119.60) differ from pre-installation app use frequency (*Mdn* = 120.20), $r = -0.17$, $p = .376$.

4.2. RQ 2

RQ 2 deals with reactivity effects regarding instant messaging, social media, and gaming apps. We conducted separate analyses for session duration and frequency, respectively.

Results show that the duration of instant messaging app use increased significantly after the installation of ATT, $b = 0.07$, 95% CI [0.01, 0.12], $t(13753) = 2.23$, $p = .026$, $q = 0.038$, which corresponds to an increase of 17%. The effect was positive as opposed to the overall effect on app use duration. In contrast, social media app use duration decreased significantly, $b = -0.18$, 95% CI [−0.26, −0.11],

Table 2
Regression results for time frame predicting (log) duration of post-installation smartphone and app use sessions.

Days	*n*	*b*	95% CI	*t*(8315)	*p*	*q*
Smartphone use						
1–2	8329	−0.03	[− 0.12, 0.06]	−0.61	.542	.542
3–4	8170	0.16	[0.06, 0.25]	3.20	.001	.008
5–6	8150	0.05	[− 0.05, 0.15]	1.00	.319	.479
7–8	8140	0.14	[0.04, 0.23]	2.78	.005	.016
9–10	7855	0.05	[− 0.08, 0.17]	0.74	.460	.542
11–12	7822	0.13	[0.00, 0.25]	1.99	.047	.094
App use						
1–2	24437	−0.17	[− 0.23, − 0.11]	−5.81	< .001	.000
3–4	24348	−0.09	[− 0.15, − 0.03]	−3.12	.002	.002
5–6	23535	−0.30	[− 0.37, − 0.24]	−8.78	< .001	.000
7–8	23443	0.12	[0.05, 0.18]	3.49	< .001	.001
9–10	22589	0.30	[0.21, 0.39]	6.62	< .001	.000
11–12	22895	0.28	[0.20, 0.37]	6.44	< .001	.000

Note. Days represents the post-installation days considered for the analysis. Q-values represent p-values after FDR correction by use subset (smartphone or app).

$t(9980) = -5.12, p < .001, q < 0.001$, which corresponds to a decrease of 24%. The effect was also negative, but significantly larger than the overall effect on app use duration. We did not find a significant effect of tracking on gaming app use duration, $b = 0.13$, 95% CI [− 0.07, 0.33], $t(1570) = 1.30, p = .194, q = 0.194$. None of the three app types were affected by tracking with regard to frequency.

4.3. RQ 3

Finally, RQ 3 asks how long potential reactivity effects persist. See Table 2 for the results concerning smartphone and app use duration, Table 3 for smartphone and app use frequency, and Table 4 for instant messaging and social media app use duration.

Most notably, app use duration first decreased before increasing again after seven days. Smartphone use duration constantly increased after three days, although only two effects were significant. Regarding smartphone and app use frequency, we found consistently negative effects, none of which were significant. Instant messaging app use duration first decreased but then increased after five days. Social media app use duration first decreased but then returned to usual levels after nine days.

Table 3
Wilcoxon signed-rank test results for average use session frequency per day between pre- and post-installation use of smartphone and app use.

Days	*n*	*V*	95% CI	*r*	*p*	*q*
Smartphone use						
1–2	12	39	[− 12.62, 25.1]	0.00	1.000	1.000
3–4	12	46	[− 8.9, 40.57]	−0.10	.622	.778
5–6	11	43	[− 6.6, 12.51]	−0.17	.413	.778
7–8	10	35	[− 6.28, 11.85]	−0.15	.492	.778
9–10	6	19	[− 8.5, 83.7]	−0.48	.094	.469
App use						
1–2	13	58	[− 23.83, 41.3]	−0.16	.414	.518
3–4	13	53	[− 43.4, 47.75]	−0.09	.635	.635
5–6	12	55	[− 21.35, 28.42]	−0.24	.233	.518
7–8	11	44	[− 17, 61.03]	−0.19	.365	.518
9–10	6	18	[− 15.17, 218.56]	−0.41	.156	.518

Note. Days represents the post-installation days considered for the analysis. Q-values represent p-values after FDR correction by use unit (smartphone or app).

Table 4
Regression results for time frame predicting (log) duration of post-installation instant messaging and social media app use sessions.

Days	*n*	*b*	95% CI	*t*(7996)	*p*	*q*
Instant Messaging						
1–2	8011	−0.19	[− 0.28, − 0.09]	−3.81	< .001	.000
3–4	7854	−0.01	[− 0.11, 0.09]	−0.15	.880	.880
5–6	7613	0.16	[0.05, 0.26]	2.83	.005	.007
7–8	7657	0.20	[0.09, 0.31]	3.55	< .001	.001
9–10	7448	0.20	[0.06, 0.35]	2.70	.007	.008
11–12	7682	0.27	[0.15, 0.40]	4.25	< .001	.000
Social Media						
1–2	5319	−0.27	[− 0.37, − 0.16]	−5.04	< .001	.000
3–4	5460	−0.21	[− 0.31, − 0.11]	−4.05	< .001	.000
5–6	4923	−0.29	[− 0.43, − 0.16]	−4.18	< .001	.000
7–8	5054	−0.17	[− 0.30, − 0.05]	−2.74	.006	.009
9–10	4794	−0.01	[− 0.16, 0.14]	−0.09	.925	.925
11–12	4944	0.02	[− 0.17, 0.20]	0.19	.852	.925

Note. Days represents the post-installation days considered for the analysis. Q-values represent p-values after FDR correction.

5. Discussion

In this study, we investigated whether tracking people's smartphone use leads to changes in the quantitative metrics, namely duration and frequency, of that use. Further, we were curious whether changes in use, if present, affect specific app types in different ways. Lastly, we checked how long potential changes persist. We developed an Android app that could access smartphone use data from before and after its installation. We then conducted a quasi-experiment by tracking 25 people's smartphone use for up to 14 days.

We found that the duration of smartphone use sessions was slightly higher after ATT was installed. The duration of app use, however, was lower. Reasons for these rather small and contradictory overall effects can be seen in the analysis of effect persistence and reactivity effects regarding instant messaging and social media apps.

During the first two days, smartphone use duration was not affected by tracking at all. During some of the following days, it was increased and during others, it was not affected. In sum, this resulted in the small, positive overall effect we found. App use duration first decreased during the first two days, then increased as compared to pre-installation use after about seven days. The initial decrease outweighed the later increase, which is why the overall effect we found was negative.

It is likely that the decrease in app use duration shortly after the installation of ATT was caused by the tracking, which is consistent with previous research on reactivity (Gittelsohn et al., 1997; e.g.,; Harris, 1982; Wu, 2013). If it had made no difference, there would not have been a reason for any detectable decrease whatsoever during (and only during) the first few days and app use duration would either have stayed on the same level or already increased shortly after tracking started, just like smartphone use did. One might argue that day-to-day variations of use between individuals or the specific day of the week the installation took place introduced variance that led to effects caused by chance alone. However, considering that a dozen people provided data for the analysis and introduced variance, that installation dates varied, and that we controlled for differences between individuals and considered q-values, we argue that inter-individual changes should hardly be the cause of the effects we found.

The fact that the immediate decrease in app use duration was followed by a consistent and significant increase (up to 35%) with low FDR seven days later supports our assumption. It is, however, unlikely that this increase was associated with the tracking. Although we controlled for the influence of the holiday season, this extraordinary time frame probably still affected data that were produced longer than one day before it began.

Smartphone and app use frequency were not affected by ATT. The loss of statistical power due to the aggregation necessary for frequency analysis possibly concealed further significant effects. The question remains whether effects on smartphone and app use frequency, if they exists in the population, are consistent or change directions with time like we found regarding app use duration. Naab and Schnauber (2016) argue that habitual media use is usually initiated, but not necessarily performed, unconsciously. Therefore, post-installation app use was possibly still initiated unconsciously and therefore just as frequently as pre-installation app use due to habituation. The conscious performance of app use, however, was stopped sooner than usual as participants knew that their use was being tracked, resulting in shorter app use duration on average. On another note, every smartphone use session ultimately consists of one or multiple app uses. Considering that post-installation app use duration decreased for some days, app use frequency would need to increase during the same time frame for smartphone use duration to stay the same or even increase. This would indicate that people compensate for lower app use duration with higher frequency when they assume that their social reputation or privacy with regard to the device are in danger. These assumptions should be addressed in future research and might yield interesting results with regard to privacy behavior and use patterns.

One reason why tracking had more distinguishable effects on app than smartphone use duration is the number of app use sessions observed, which is roughly triple the number of smartphone use sessions and results in higher statistical power. We also argue that the importance of smartphones in everyday communication attenuates reactivity effects regarding the duration of their overall use and therefore shows in the use of specific apps and app types rather than overall device use. We investigated two of these app types. While instant messaging app use duration was higher after the installation of ATT, social media use duration was lower. Analysis of effect persistence showed that the same patterns already seen in overall app use duration applied to both app types. Instant messaging app use duration first decreased, but then increased after about five days. Social media app use duration first decreased, but returned to regular levels after about nine days. The reason for this might again be the holiday season: Conversing with family members and (close) friends is not only an important part of preparing for the upcoming holidays for many people, but also a popular use of instant messengers like WhatsApp (Church & De Oliveira, 2013). Meanwhile, people also use social media like Facebook to “interact with people they do not regularly see [and] chat with old acquaintances” (Whiting & Williams, 2013). As such, people might just use instant messaging more than social media apps in the context of an upcoming holiday that is centered around family and close friends, which is why instant messaging app use duration suddenly increased as time passed. The effects on social media apps, though, may therefore represent the essence of reactivity in this study. This also explains why we found tracking to affect app use duration negatively overall, but not smartphone use duration. As we could control for app types in the analysis of app use duration, such differences between types due to the holidays were canceled out and the results emphasize that the initial decrease in app use duration was indeed caused by reactivity. Earlier research showed that more common behavior seems to cause less reactivity while less common behavior causes more (Cousens et al., 1996). Since we did not analyze rarely used app types, we possibly overlooked other interesting effects. It is also important to note that app types assigned in the Google Play Store do not provide sufficient distinction between app types. For example, there arguably is an overlap between the types video and games and entertainment.

To sum it up, the results indicate that on average, the duration of app use decreases due to tracking for about six days. Smartphone use is not affected by tracking. One reason might be social desirability. People probably assume that using their device for longer periods of time without interruptions might be frowned upon. This especially holds for heavy users, as research showed that symptoms of smartphone addiction are positively associated with tendencies to fulfill social desirability (Herrero, Urueña, Torres, & Hidalgo, 2019, p. 86). Hence, they might end their use sessions slightly earlier than they usually would because the installation of ATT rendered the issue particularly salient for them. The effects found could also be a symptom of a lack of trust. Possibly, participants did not truly believe that participation was anonymous and no contents or private information were transmitted, resulting in less app use duration. Generally, the effects we found that could be attributed to the tracking were rather weak, which speaks for low reactivity and high convergent validity of smartphone use tracking data. This may be associated with the absence of interaction (Cousens et al., 1996) and minimum communication between researchers and participants (Schwartz et al., 2013; Taneja & Viswanathan, 2014) in our study. Since we did not perform observation in person and participants were not notified at all during the whole process after ATT was installed, reactivity could be expected to be low. This speaks for using smartphone use tracking under these circumstances. Expected gain is also considered a factor that contributes to the facilitation of reactivity effects (Barnes, 2010). No incentives were provided for participation in our study, which possibly further decreased reactivity.

5.1. Limitations and future research

The greatest constraint of this study is low sample size. This is due to the difficulty of recruitment for a study that involves passive observation. Our sample size is comparable to sample sizes in other CHI studies, which shows that many researchers in this field seem to struggle with this problem. Previous research showed that financial incentives are associated with higher willingness to participate in passive mobile tracking studies (Keusch, Struminskaya, Antoun, Couper, & Kreuter, 2019). Future research should therefore consider following that suggestion, keeping in mind possible negative consequences mentioned before. Due to the financial restrictions, we were dependent on people to participate out of sheer interest in the subject matter. This may not have mitigated reactivity effects too much, but it certainly impeded their identification due to low statistical power. While we found reactivity effects, they were only partially statistically significant and showed high FDR rates, where applicable. Especially regarding the investigation of different app types, a larger sample size would allow for better estimation of average use as app types are not always represented sufficiently. Another reason for our low sample size might be the requirement to install the app manually using a file that had to be downloaded, which possibly prevented less tech-savvy people from participating. This shows in the fact that about 60 people originally filled out the form on the project website and received the instruction email. Offering ATT on the Google Play Store might have reduced the problem. However, we avoided this approach in order to prevent possible participation in the project without visiting the project website for information or misusing the app for fun. The best compromise might be offering it on the Google Play Store as a closed test (Google, 2020c) or tying participation to a password sent out via Email. Also, it should be noted that the Google Play Store is subject to strict regulations concerning the sensitive data that are involved in tracking (Google, 2020d).

Future data collections of this kind should not immediately precede or even overlap with a holiday season. Even though we found effects that can reasonably be considered reactivity effects, it was technically impossible to tell apart the influence of the holiday season from the influence of tracking after some days.

Further, the lack of distinction between app types according to the Google Play Store may have slightly biased some results with regard to app types. We advise future research to either categorize apps manually based on existing research, or double-check the assignments Google Play Store provides.

Another caveat is the detection of locks, unlocks and shutdowns. As we showed, one cannot reliably assess recent smartphone and app use below Android version 9, even when actively tracking locks and unlocks

in the tracking app – let alone pre-installation use where this is not possible at all. Therefore, it is most feasible to restrict participation to devices running Android versions 9 or higher. Also, it seems like Android 10 introduced new opportunities and caveats for tracking apps to come (Google, 2020a).

Finally, we would like to offer a suggestion for future research employing smartphone use tracking methods. When interested in descriptive accounts of smartphone use quantities, researchers may resort to simply using pre-installation use data right away to get rid of reactivity completely. But when combining tracking with other methods like ESM to assess additional information on psychological or situational variables, or content analysis (e.g., De Vreese et al., 2017), tracking recent use is indispensable. Our findings therefore encourage the use of tracking data that were produced *at least seven days* after installation, just to be on the safe side. Of course, this number is not unrelated to our study setting, effectively involving only 12 people and the holiday season. If anything, decreased app use duration due to tracking was overshadowed by increased duration due to the upcoming holiday season. Accordingly, we advise researchers to wait even longer than seven days whenever possible.

5.2. Conclusion

In recent years, the smartphone has become a crucial means of communication and one of the most dominant ways to access digital content on the Internet. As such, investigating the validity of its measures is of utmost importance for communication science and many more fields of research. Smartphone use tracking allows for passive observation of use behavior in a way that has not been possible for any other medium before. The greatest advantages lie in the assessment of past use, as all events are being documented by the operating system at all times, and the ability to run tracking apps in the background. This yields the opportunity to juxtapose past and recent use behavior in order to investigate possible reactivity to the tracking, which has always been a problem in research using observation methods.

We found that the average duration of app use sessions decreases for some days due to the participation in tracking. The same applies to instant messaging and social media apps. Future research is needed to investigate effects of tracking on the average duration of smartphone use sessions, other app types, and both smartphone and app use session frequencies.

We hope this study motivates other researchers to give thought to the subject of reactivity concerning smartphone use tracking and conduct further investigations of possible implications for research on media use, privacy behavior, and research methods.

Declaration of competing interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chbr.2021.100142.

Author note

RT designed and conducted the study, the analysis, and wrote the manuscript. TT was the lead developer of the app. We would like to thank Prof. Daniela Schlütz from SoSci Panel for distributing our study among panel participants, Polina Guseva for literature research and inspiring discussions, and Harald Papp and Dennis Ricci for setting up and administering our server and database.

References

Adair, J. G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology, 69*(2), 334–345. https://doi.org/10.1037/0021-9010.69.2.334

Andrews, S., Ellis, D. A., Shaw, H., & Piwek, L. (2015). Beyond self-report: Tools to compare estimated and real-world smartphone use. *PloS One, 10*(10), Article e0139004. https://doi.org/10.1371/journal.pone.0139004. Retrieved from.

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja.

van Ballegooijen, W., Ruwaard, J., Karyotaki, E., Ebert, D. D., Smit, J. H., & Riper, H. (2016). Reactivity to smartphone-based ecological momentary assessment of depressive symptoms (MoodMonitor): Protocol of a randomised controlled trial. *BMC Psychiatry, 16*(1), 4–9. https://doi.org/10.1186/s12888-016-1065-5

Barnes, B. R. (2010). The Hawthorne Effect in community trials in developing countries. *International Journal of Social Research Methodology, 13*(4), 357–370. https://doi.org/10.1080/13645570903269096

Bayer, J. B., Campbell, S. W., & Ling, R. (2016). Connection cues: Activating the norms and habits of social connectedness. *Communication Theory, 26*(2), 128–149. https://doi.org/10.1111/comt.12090

van Berkel, N., Goncalves, J., Lovén, L., Ferreira, D., Hosio, S., & Kostakos, V. (2019). Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. *International Journal of Human-Computer Studies, 125* (December), 118–128. https://doi.org/10.1016/j.ijhcs.2018.12.002

Berthelot, J. M., Nizard, J., & Maugars, Y. (2019). The negative Hawthorne effect: Explaining pain overexpression. *Joint Bone Spine, 86*(4), 445–449. https://doi.org/10.1016/j.jbspin.2018.10.003

Boase, J., & Ling, R. (2013). Measuring mobile phone use: Self-report versus log data. *Journal of Computer-Mediated Communication, 18*(4), 508–519. https://doi.org/10.1111/jcc4.12021

Bouchet, C., Guillemin, F., & Briançon, S. (1996). Nonspecific effects in longitudinal studies: Impact on quality of life measures. *Journal of Clinical Epidemiology, 49*(1), 15–20. https://doi.org/10.1016/0895-4356(95)00540-4

Caine, K. (2016). Local standards for sample size at CHI. *Conference on Human Factors in Computing Systems - Proceedings*, 981–992. https://doi.org/10.1145/2858036.2858498

Chang, F. C., Chiu, C. H., Chen, P. H., Miao, N. F., Chiang, J. T., & Chuang, H. Y. (2018). Computer/mobile device screen time of children and their eye care behavior: The roles of risk perception and parenting. *Cyberpsychology, Behavior, and Social Networking, 21*(3), 179–186. https://doi.org/10.1089/cyber.2017.0324

Christakis, D. A., & Zimmerman, F. J. (2009). Young children and media: Limitations of current knowledge and future directions for research. *American Behavioral Scientist, 52*(8), 1177–1185. https://doi.org/10.1177/0002764209331540

Church, K., & De Oliveira, R. (2013). What's up with WhatsApp? Comparing mobile instant messaging behaviors with traditional SMS. In *MobileHCI 2013 - proceedings of the 15th international conference on human-computer interaction with mobile devices and services* (pp. 352–361). https://doi.org/10.1145/2493190.2493225

Cohen, A. A., & Lemish, D. (2003). Real time and recall measures of mobile phone use: Some methodological concerns and empirical applications. *New Media & Society, 5* (2), 167–183. https://doi.org/10.1177/1461444803005002002

Cook, D. L. (1962). The Hawthorne Effect in educational research. *Phi Delta Kappan, 44* (5), 116–122. Retrieved from http://www.jstor.org/stable/20342865.

Cousens, S., Kanki, B., Toure, S., Diallo, I., & Curtis, V. (1996). Reactivity and repeatability of hygiene behaviour: Structured observations from Burkina Faso. *Social Science & Medicine, 43*(9), 1299–1308. https://doi.org/10.1016/0277-9536(95)00380-0

Csikszentmihalyi, M. (2014). Flow and the foundations of positive psychology: The collected works of Mihaly Csikszentmihalyi. *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*, 1–298. https://doi.org/10.1007/978-94-017-9088-8

David, M. E., Roberts, J. A., & Christenson, B. (2018). Too much of a good thing: Investigating the association between actual smartphone use and individual well-being. *International Journal of Human-Computer Interaction, 34*(3), 265–275. https://doi.org/10.1080/10447318.2017.1349250

De Vreese, C. H., Boukes, M., Schuck, A., Vliegenthart, R., Bos, L., & Lelkes, Y. (2017). LLinking survey and media content data: Opportunities, considerations, and pitfall. *Communication Methods and Measures, 11*(4), 221–244. https://doi.org/10.1080/19312458.2017.1380175

Eckmanns, T., Bessert, J., Behnke, M., Gastmeier, P., & Rüden, H. (2006). Compliance with antiseptic hand rub use in I intensive care units: The Hawthorne Effect. *Infection Control & Hospital Epidemiology, 27*(9), 931–934. https://doi.org/10.1086/507294

Fang, J., Wen, C., & Prybutok, V. (2014). An assessment of equivalence between paper and social media surveys: The role of social desirability and satisficing. *Computers in Human Behavior, 30*, 335–343. https://doi.org/10.1016/j.chb.2013.09.019

Feil, P. H., Grauer, J. S., Gadbury-Amyot, C. C., Kula, K., & McCunniff, M. D. (2002). Intentional use of the Hawthorne effect to improve oral hygiene compliance in orthodontic patients. *Journal of Dental Education, 66*(10), 1129–1135. https://doi.org/10.1002/j.0022-0337.2002.66.10.tb03584.x.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Los Angeles; London; New Delhi: Sage. Retrieved from http://gso.gbv.de/DB=2.1/PPNSET?PPN=68436977X.

Furini, M., Mirri, S., Montangero, M., & Prandi, C. (2020). Privacy perception when using smartphone applications. *Mobile Networks and Applications, 25*(3), 1055–1061. https://doi.org/10.1007/s11036-020-01529-z

Gerber, A. S., Green, D. P., & Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review, 102* (1), 33–48. https://doi.org/10.1017/S000305540808009X

Gittelsohn, J., Shankar, A. V., West, K. P., Ram, R. M., & Gnywali, T. (1997). Estimating reactivity in direct observation studies of health behaviors. *Human Organization, 56* (2), 182–189. https://doi.org/10.17730/humo.56.2.c7x0532q2u86m207

Google. (2019a). *Introduction to activities*. Retrieved from https://developer.android.com/guide/components/activities/intro-activities.

Google. (2019b). *Set the application ID*. Retrieved from https://developer.android.com/studio/build/application-id.

Google. (2019c). *UsageEvents.Event*. Retrieved from https://developer.android.com/reference/android/app/usage/UsageEvents.Event.

Google. (2020a). *Behavior changes: All apps*. Retrieved from https://developer.android.com/about/versions/10/behavior-changes-all.

Google. (2020b). *KeyguardManager*. Retrieved from https://developer.android.com/reference/android/app/KeyguardManager.

Google. (2020c). *Set up an open, closed or internal test*. Retrieved from https://support.google.com/googleplay/android-developer/answer/3131213?hl=en.

Google. (2020d). *User data*. Retrieved from https://support.google.com/googleplay/android-developer/answer/9888076.

Granberg, D., & Holmberg, S. (1992). The Hawthorne effect in election studies: The impact of survey participation on voting. *British Journal of Political Science, 22*(2), 240–247. https://doi.org/10.1017/S0007123400006359

Griffioen, N., Rooij, M. van, Lichtwarck-Aschoff, A., & Granic, I. (2020). Toward improved methods in social media research. *Technology, Mind, and Behavior, 1*(1), 1–15. https://doi.org/10.1037/tmb0000005

Guthrie, G. (2010). *Basic research methods : An entry to social science research*. New Delhi, India: Sage Publications Pvt. Ltd. https://doi.org/10.4135/9788132105961.

Haddad, N. F., Nation, J. R., & Williams, J. D. (1975). Programmed student achievement: A Hawthorne effect? *Research in Higher Education, 3*(4), 315–322. https://doi.org/10.1007/BF00991248

Harari, G. M., Müller, S. R., Stachl, C., Wang, R., Wang, W., Bühner, M., et al. (2019). Sensing sociability: Individual differences in young adults' conversation, calling, texting, and app use behaviors in daily life. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000245

Harris, F. C. (1982). Subject reactivity in direct observational assessment: A review and critical analysis. *Clinical Psychology Review, 2*(4), 523–538. https://doi.org/10.1016/0272-7358(82)90028-9.

Herrero, J., Urueña, A., Torres, A., & Hidalgo, A. (2019). Smartphone addiction: Psychosocial correlates, risky attitudes, and smartphone harm. *Journal of Risk Research, 22*(1), 81–92. https://doi.org/10.1080/13669877.2017.1351472

Hox, J., & McNeish, D. (2020). Small samples in multilevel modeling. *Small Sample Size Solutions*, (February), 215–225. https://doi.org/10.4324/9780429273872-18

Jackson, K. (2018). *A brief history of the smartphone*. Retrieved from https://sciencenode.org/feature/Howdidsmartphonesevolve.php.

Jensen, J. D., & Hurley, R. J. (2005). Third-person effects and the environment: Social distance, social desirability, and presumed behavior. *Journal of Communication, 55* (2), 242–256. https://doi.org/10.1093/joc/55.2.242

JoMingyu. (2020). *Google-play-scraper*. Retrieved from https://github.com/JoMingyu/google-play-scraper.

Jones, S. L., Ferreira, D., Hosio, S., Goncalves, J., & Kostakos, V. (2015). Revisitation analysis of smartphone app use. In *UbiComp 2015 - proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing* (pp. 1197–1208). https://doi.org/10.1145/2750858.2807542

Kaye, L. K., Orben, A., Ellis, D. A., Hunter, S. C., & Houghton, S. (2020). The conceptual and methodological mayhem of "screen time". *International Journal of Environmental Research and Public Health, 17*(10). https://doi.org/10.3390/ijerph17103661

Keusch, F., Struminskaya, B., Antoun, C., Couper, M. P., & Kreuter, F. (2019). Willingness to participate in passive mobile data collection. *Public Opinion Quarterly, 83*, 210–235. https://doi.org/10.1093/poq/nfz007

Klimmt, C., Hefner, D., Reinecke, L., Rieger, D., & Vorderer, P. (2019). The permanently online and permanently connected mind. *Permanently Online, Permanently Connected, 18–28*. https://doi.org/10.4324/9781315276472-3

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality and Quantity, 47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9

Kypri, K., Langley, J. D., Saunders, J. B., & Cashell-Smith, M. L. (2007). Assessment may conceal therapeutic benefit: Findings from a randomized controlled trial for hazardous drinking. *Addiction, 102*(1), 62–70. https://doi.org/10.1111/j.1360-0443.2006.01632.x

Lemola, S., Perkinson-Gloor, N., Brand, S., Dewald-Kaufmann, J. F., & Grob, A. (2014). Adolescents' electronic media use at night, sleep disturbance, and depressive symptoms in the smartphone age. *Journal of Youth and Adolescence, 44*(2), 405–418. https://doi.org/10.1007/s10964-014-0176-x

Leonard, K., & Masatu, M. C. (2006). Outpatient process quality evaluation and the Hawthorne Effect. *Social Science & Medicine, 63*(9), 2330–2340. https://doi.org/10.1016/j.socscimed.2006.06.003

Lied, T. R., & Kazandjian, V. A. (1998). A Hawthorne strategy: Implications for performance measurement and improvement. *Clinical Performance in Quality Healthcare, 6*, 201–204. https://www.researchgate.net/publication/12948191_A_Hawthorne_strategy_Implications_for_performance_measurement_and_improvement.

Lopez-Fernandez, O., Männikkö, N., Kääriäinen, M., Griffiths, M. D., & Kuss, D. J. (2018). Mobile gaming and problematic smartphone use: A comparative study between Belgium and Finland. *Journal of Behavioral Addictions, 7*(1), 88–99. https://doi.org/10.1556/2006.6.2017.080

Mangione-Smith, R., Elliott, M. N., McDonald, L., & McGlynn, E. A. (2002). An observational study of antibiotic prescribing behavior and the Hawthorne effect. *Health Services Research, 37*(6), 1603–1623. https://doi.org/10.1111/1475-6773.10482

Marty-Dugas, J., Ralph, B. C. W., Oakman, J. M., & Smilek, D. (2018). The relation between smartphone use and everyday inattention. *Psychology of Consciousness: Theory Research, and Practice, 5*(1), 46–62. https://doi.org/10.1037/cns0000131

McCambridge, J., Wilson, A., Attia, J., Weaver, N., & Kypri, K. (2019). Randomized trial seeking to induce the Hawthorne effect found no evidence for any effect on self-reported alcohol consumption online. *Journal of Clinical Epidemiology, 108*, 102–109. https://doi.org/10.1016/j.jclinepi.2018.11.016

McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology, 67*(3), 267–277. https://doi.org/10.1016/j.jclinepi.2013.08.015

Murray, M., Swan, A. V., Kiryluk, S., & Clarke, G. C. (1988). The Hawthorne effect in the measurement of adolescent smoking. *Journal of Epidemiology & Community Health, 42* (3), 304–306. https://doi.org/10.1136/jech.42.3.304

Naab, T. K., Karnowski, V., & Schlütz, D. (2018). Reporting mobile social media use: How survey and experience sampling measures differ. *Communication Methods and Measures, 13*(2), 126–147. https://doi.org/10.1080/19312458.2018.1555799

Naab, T. K., & Schnauber, A. (2016). Habitual initiation of media use and a response-frequency measure for its examination. *Media Psychology, 19*(1), 126–155. https://doi.org/10.1080/15213269.2014.951055

Newzoo. (2019). *Global mobile market report*. Newzoo. Retrieved from https://newzoo.com/insights/trend-reports/newzoo-global-mobile-market-report-2019-light-version/.

Otten, J. J., Littenberg, B., & Harvey-Berino, J. R. (2010). Relationship between self-report and an objective measure of television-viewing time in adults. *Obesity, 18*(6), 1273–1275. https://doi.org/10.1038/oby.2009.371

Parry, D. A., Davidson, B. I., Sewall, C. J. R., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-021-01117-5

Price, P. C., Jhangiani, R. S., Chiang, I.-C. A., Leighton, D. C., & Cuttler, C. (2017). *Research methods in psychology*. Retrieved from https://opentext.wsu.edu/carriecuttler/.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Raento, M., Oulasvirta, A., & Eagle, N. (2009). Smartphones: An emerging tool for social scientists. *Sociological Methods & Research, 37*(3), 426–454. https://doi.org/10.1177/0049124108330005

Rosen, L. D., Mark Carrier, L., Pedroza, J. A., Elias, S., O'Brien, K. M., Lozano, J., et al. (2018). The role of executive functioning and technological anxiety (FOMO) in college course performance as mediated by technology usage and multitasking habits. *Psicologia Educativa, 24*(1), 14–25. https://doi.org/10.5093/psed2018a3

Scharkow, M. (2016). The accuracy of self-reported internet use—a validation study using client log data. *Communication Methods and Measures, 10*(1), 13–27. https://doi.org/10.1080/19312458.2015.1118446

Schmitz, B., Stanat, P., Sang, F., & Tasche, K. (1996). Reactive effects of a survey on the television viewing behavior of a telemetric television audience panel: A combined time-series and control-group analysis. *Evaluation Review, 20*(2), 204–229. https://doi.org/10.1177/0193841X9602000205.

Schnauber-Stockmann, A., & Naab, T. K. (2019). The process of forming a mobile media habit: Results of a longitudinal study in a real-world setting. *Media Psychology, 22*(5), 714–742. https://doi.org/10.1080/15213269.2018.1513850

Schwartz, D., Fischhoff, B., Krishnamurti, T., & Sowell, F. (2013). The Hawthorne effect and energy awareness. *Proceedings of the National Academy of Sciences of the United States of America, 110*(38), 15242–15246. https://doi.org/10.1073/pnas.1301687110

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation, 22* (2), 127–160. https://doi.org/10.1016/S1098-2140(01)00133-3

Sonnenberg, B., Riediger, M., Wrzus, C., & Wagner, G. G. (2011). Measuring time use in surveys. *SOEPpapers on Multidisciplinary Panel Data Research*, (390), 1–30. https://www.diw.de/documents/publikationen/73/diw_01.c.376621.de/diw_sp0390.pdf.

SoSci Panel. (2020). *Willkommen beim SoSci panel*. Retrieved from https://www.soscipanel.de.

Taneja, H., & Viswanathan, V. (2014). Still glued to the box? Television viewing explained in a multi-platform age integrating individual and situational predictors. *International Journal of Communication, 8*, 2134–2159. https://ijoc.org/index.php/ijoc/article/view/2841.

Thulin, E., & Vilhelmson, B. (2007). Mobiles everywhere: Youth, the mobile phone, and changes in everyday practice. *Young, 15*(3), 235–253. https://doi.org/10.1177/1103308807015003002

Turkle, S. (2008). Always-On/Always-On-You: The tethered self. In J. E. Katz (Ed.), *Handbook of mobile communication studies* (pp. 121–138). https://doi.org/10.7551/mitpress/9780262113120.003.0010

Valkenburg, P. M., & Peter, J. (2013). Five challenges for the future of media-effects research. *International Journal of Communication, 7*(1), 197–215. https://ijoc.org/index.php/ijoc/article/view/1962/0.

Vandewater, E. A., & Lee, S.-J. (2009). Measuring children's media use in the digital age: Issues and challenges. *American Behavioral Scientist, 52*(8), 1152–1176. https://doi.org/10.1177/0002764209331539

Walker, R., Koh, L., Wollersheim, D., & Liamputtong, P. (2015). Social connectedness and mobile phone use among refugee women in Australia. *Health and Social Care in the Community, 23*(3), 325–336. https://doi.org/10.1111/hsc.12155

Whiting, A., & Williams, D. (2013). Why people use social media: A uses and gratifications approach. *Qualitative Market Research: An International Journal, 16*(4), 362–369. https://doi.org/10.1108/QMR-06-2013-0041

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilcockson, T. D. W., Ellis, D. A., & Shaw, H. (2018). Determining typical smartphone usage: What data do we need? *Cyberpsychology, Behavior, and Social Networking, 21*(6), 395–398. https://doi.org/10.1089/cyber.2017.0652

Wu, L. (2013). Social network effects on productivity and job security: Evidence from the adoption of a social networking tool. *Information Systems Research, 24*(1), 30–51. https://doi.org/10.1287/isre.1120.0465