Innovating the measurement of parental digital behavior: A case study using metered data

Joshua Claassen
German Centre for Higher Education Research and Science Studies (DZHW)
Leibniz University Hannover

Melanie Revilla
RECSM-Universitat Pompeu Fabra

This document is a preprint and thus it may differ from the final version.

Acknowledgments
The authors are very grateful to Kieran Sargeant Rivilla and Tracy Silva for their help during data analysis and in preparing this study.

Funding
This project received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 849165).

Abstract
Parental Digital Behavior (PDB) is usually measured through survey self-reports struggling with recall error and social desirability. A promising alternative is the passive collection of digital traces, such as website visits, but there is little guidance on transforming traces into meaningful measurements of PDB. The present study identifies key analytical decisions involved in this methodological process and investigates their impact on research outcomes. We obtained approx. 183,000 traces from 800 participants in the Netquest metered panel in Spain, including website visits, search terms, and app usage, and classified these traces with two Large Language Models (LLMs). Conducting a "multiverse analysis," we created separate PDB measurements for 130 combinations of analytical decisions, and evaluated each measurement by estimating the proportion of variance explained through regression models. The 130 measurements of PDB differ substantially in terms of their explainability, that is, the proportion of explained variance. Some analytical decisions, including the type of trace considered (i.e., website visits, search terms, and/or app usage), are systematically associated with explainability, and should therefore be made carefully. This study expands the application field of metered data to PDB, laying the methodological foundation for a better understanding of internet-mediated parental information seeking. By providing novel insights on the classification and analysis of different types of traces, it also improves the general methodological toolkit for measuring digital behavior.

Keywords: Metered data, Digital trace data, Parental digital behavior, Large Language Models, Classification

Introduction

A growing number of (expectant) parents is engaging in Parental Digital Behavior (PDB) by turning to websites, search engines, and apps, supplementing traditional information sources, such as healthcare professionals (de Rosso et al., 2022; Jaks et al. 2019). PDB is associated with real-world parenting, including educational support (Livingstone et al., 2018), decision-making in the context of child vaccination (Charron et al., 2020), as well as nutrition and breastfeeding (Slomian et al., 2017). In addition, existing gender inequalities in care work (García-Mainar et al., 2011) seem to be reinforced in the digital sphere as women tend to engage in more PDB than men (Hiebert et al., 2021; Laws et al., 2019).

Although being increasingly studied, PDB is primarily measured with self-reports collected through surveys (Baumann et al., 2020; Lupton & Pedersen, 2016; Yudianto et al., 2023) that struggle with recall error (Tourangeau et al., 2000), satisficing (Krosnick, 1991), and social desirability bias (Neuberger, 2016). A promising alternative is represented by utilizing digital traces in the form of metered data (sometimes also called "web tracking data;" Revilla, 2022). Metered data refers to the collection of traces through tracking applications (so-called "meters") that participants actively install on their browsing devices (e.g., computers, tablets, and smartphones), gathering up to three types of traces: website visits (or URLs), search terms, and app usage (Bosch & Revilla, 2022a). These traces offer detailed insights into participants' PDB, such as website visits related to fetal development, online searches for infant nutrition, and the use of educational apps offering advice for (expectant) parents.

However, to draw substantial conclusions from metered data, researchers must first transform traces into meaningful measurements. This process involves numerous analytical decisions, but previous research has primarily focused on the collection of metered data rather than their processing and analysis. Thus, there are almost no best-practices yet. To address this research gap, we conducted a case study investigating how traces in the form of website visits, search terms, and app usage can be transformed into meaningful measurements of PDB. To this end, we obtained metered data alongside survey self-reports on (expectant) parenthood from 800 panelists of the Netquest metered panel in Spain in 2023 and 2024. The field period lasted up to 24 weeks. We systematically investigate how key analytical decisions, such as the type of trace considered, influence the explainability of PDB measurements, defined as the proportion of variance explained through regression models. Drawing on our empirical examinations, we derive evidence-based recommendations for future studies utilizing metered data.

Background and literature review

Opportunities and challenges of metered data

Metered data is increasingly being used to study various types of digital behavior, such as smartphone usage (Wenz et al., 2024), news and media consumption (Wirz et al., 2025), searches for health information (Bach & Wenz, 2020), online job searches (Ochoa & Revilla, 2025a), and ChatGPT adoption (Ochoa et al., 2026). While the measurement of PDB, to the

best of our knowledge, is still uncharted territory with respect to metered data, it offers several advantages over survey self-reports: Metered data allows researchers to measure PDB directly and in unprecedented granularity, avoiding recall error (Revilla et al., 2017; Wenz et al., 2024). Since participants install the meter only at the beginning of data collection, metered data may also reduce response burden and satisficing behavior (Revilla, 2022; Revilla et al., 2021). Societal norms and values related to parenting (Lee, 2023) represent another key point, potentially resulting in overreporting of PDB. In contrast to survey self-reports, metered data is less prone to social desirability bias because sensitive information is collected unobtrusively (Keusch et al., 2023).

Nonetheless, research on the collection of metered data (Adam et al., 2025; Revilla et al., 2017) points to potential error sources, such as non-participation bias (Keusch et al., 2022; Revilla et al., 2021) that may differ between countries (Revilla et al., 2016). In addition, previous research has identified tracking undercoverage (Bosch et al., 2025; Gonzáles-Bailón & Xenos, 2023; Ochoa & Revilla, 2025a) and device sharing (Revilla et al., 2017) as key error sources: When measuring PDB, participants may not install a meter on certain browsing devices (i.e., tracking undercoverage), resulting in an underestimation of PDB. Conversely, participants may jointly use a specific browsing device with their partner (i.e., device sharing), resulting in an overestimation of individual-level PDB. According to the error framework by Bosch and Revilla (2022a), the previously discussed error sources affect both the measurement of the concept of interest as well as the representation of the population of interest.

Transforming traces into measurements

The process of transforming traces into measurements of digital behavior has received far less attention than the mere collection of traces (see, for exceptions, Bosch, 2023; Ochoa & Revilla, 2025a; Ochoa & Revilla, 2025b; Reiss, 2023; von Hohenberg et al., 2024). Crucially, existing research points at numerous analytical decisions that researchers must make when analyzing traces in the form of metered data (Bosch, 2023).

Type of trace considered

First, researchers need to determine what type of trace the concept of interest produces (i.e., website visits, search terms, and/or app usage). It is plausible to assume that PDB produces all three types of traces, so that not considering one of these types would result in specification error (Bosch & Revilla, 2022a). A related consideration is the ambiguity of traces (Freelon, 2014; Hölzl et al., 2025) which may vary between types of traces. For example, conducting an online search on pregnancy-related health information must not necessarily be a representation of PDB, but could be motivated by having a pregnant friend or simply being interested in the topic, among various other reasons. In contrast, using a pregnancy-tracking app might be less ambiguous because apps require a higher initial effort (by actively downloading the app; Moungui et al., 2024). Consequently, specification error and different levels of ambiguity may cause metered data measurements to depend on the types of traces researchers consider.

Filtering, gathering, and classification of traces

For each type of trace considered, researchers need to identify traces that are relevant to the concept of interest. This can be achieved by specifying traces before data collection (i.e., a priori filtering; Bosch, 2023; Bosch & Revilla, 2022b) and/or classifying traces after data collection (i.e., post hoc classification; Reiss, 2023). The former approach adheres to the research ethical principle of "data minimization" (Adam et al., 2025) and possibly alleviates participants' privacy concerns (Makhortykh et al., 2022), as only traces that are (potentially) related to the study purposes are collected. Given that metered data can contain highly sensitive traces, such as website visits related to adult content (Martinez-Serra & Cardenal, 2025), this is a particularly important consideration. The creation of exhaustive lists of websites, search terms, and apps related to a concept of interest is often impossible, so that filtering can also be based on keywords (Sen et al., 2021). For example, as in the present study, researchers may specify that website URLs and search terms containing the term "pregnant" are to be collected. However, this likely results in a considerable amount of "bycatch," such as news articles about pregnant celebrities. Relatedly, many research questions require fine-grained measurements that cover different dimensions of the concept of interest (e.g., health-, education-, and entertainment-related PDB), going beyond the granularity of pre-defined lists and keywords. Other studies do not apply any filtering at all, so that full metered data is collected.

In such cases, traces must be (additionally) classified post-hoc, but the high volume of metered data typically renders the application of manual, human-based coding inefficient. Consequently, researchers primarily employ automated classification approaches. For example, several studies rely on the Webshrinker API (https://webshrinker.com/) to classify websites (Bach & Wenz, 2020; Kirkizh et al., 2024; Kulshrestha et al., 2021). Although Webshrinker is easy to implement, it is rather costly, does not cover search terms and apps, and requires researchers to use a pre-existing category scheme (von Hohenberg et al., 2024). Supervised machine learning models tailored to the concept of interest, in contrast, require a substantial amount of annotated data (Reiss, 2023) and often struggle to generalize beyond their training data. For example, in a recent study by van Hoof et al. (2025), contemporary machine learning models only identified between 28 and 67 percent of political and news-related search terms.

In light of these limitations, a more recently developed approach is represented by Large Language Models (LLMs) that are pre-trained on vast amounts of data and can be used without further task-specific training (Chae & Davidson, 2026). Previous research underscores the potential of LLMs when it comes to text-based classification tasks (Than et al., 2025). As different LLMs result in different classification outcomes, the LLM choice, including parameter settings such as temperature, represents a key analytical decision. In addition, some studies point at a rather low agreement between LLM-based classifications and manual, human-based coding (Revilla et al., 2026; von der Heyde et al., 2025), raising the question to what extent LLM-based classifications are in line with human judgement.

Dimension of the concept of interest

Researchers must also decide what dimension of the concept of interest they focus on. In the present study, we may create a single measurement of overall PDB as well as several measurements of more fine-grained dimensions, such as health-, education-, or entertainment-related PDB, having two important implications. First, at the trace level, measuring overall PDB would result in a binary classification task (i.e., PDB or no PDB), whereas measuring different dimensions of PDB would result in a multi-class classification task (e.g., health-, education-, or entertainment-related PDB). LLMs may perform differently for these two types of classification tasks (Kostina et al., 2025). Second, some dimensions, such as education-related PDB, might be particularly ambiguous. For example, learning websites and school information are not exclusively accessed by parents, but also by teachers, private tutors, and adult learners.

Aggregation rule

Finally, researchers are usually left with a dataset in long format, containing a high number of classified traces per participant. To transform these traces into measurements, researchers have to apply an aggregation rule. Specifically, depending on their research question, researchers may calculate the frequency or duration of traces within a particular category (e.g., health-related PDB). In addition, researchers may create absolute or relative measurements. For example, visiting 200 websites per week, including 20 websites classified as health-related PDB (absolute frequency), would result in a relative frequency of 10%. Importantly, previous research indicates that metered data measurements are sensitive to different aggregation rules (von Hohenberg et al., 2024).

Research questions and contributions

The present study aims to expand the application field of metered data to the measurement of PDB, while shedding light on the methodological process of transforming traces into measurements of digital behavior. To this end, we conducted a "multiverse analysis" (Steegen et al., 2016) by systematically varying four key analytical decisions: type of trace considered (i.e., website visits, search terms, and/or app usage), LLM employed for classification (i.e., Llama 3.2 3B or Mistral 7B), dimension focused on (i.e., overall PDB or specific dimensions of PDB), and aggregation rule applied (i.e., absolute or relative frequency and duration). We created separate measurements for all combinations of these analytical decisions and evaluated each measurement in terms of its explainability. Importantly, we define explainability as the adjusted R^2 values of a series of theoretically informed regression models, with the PDB measurements as dependent variables. These models included independent variables related to self-reported (expectant) parenthood, sociodemographic characteristics, and potential errors in the metered data. We investigate the following research questions:

How does the explainability of PDB measurements differ between different …
- … types of traces considered? (**RQ1**)
- … LLMs employed for classification? (**RQ2**)
- … dimensions of PDB focused on? (**RQ3**)
- … aggregation rules applied? (**RQ4**)

Overall, this study makes a two-fold contribution to the current state of research. First, it introduces a new application field for metered data, expanding the methodological toolkit available to social and behavioral scientists who study PDB. Second, it fills a gap in the methodological research on metered data, as it critically evaluates the impact of analytical decisions on research outcomes, deriving empirical-driven best practices. In doing so, our study contributes to the broader debate about the opportunities and challenges of digital trace data when it comes to measuring digital behavior.

Method

Data collection

We utilized data from the non-probability Netquest metered panel (https://netquest.com) in Spain. Panelists install a meter on at least one browsing device, collecting the URLs and timestamps of website visits and search terms. On mobile devices, the meter also collects data on app usage, which, in the present study, includes information on when and how long a specific app was used. To select participants for the present study, Netquest utilized the information of periodically conducted web surveys, including self-reports on (expectant) parenthood. Whenever possible, Netquest selected participants who fulfilled specific quality criteria (e.g., actively sharing traces every day). As the present study is part of a larger project, which aims to address various research questions, we drew two separate samples: One on expectant parenthood and one on parenthood. Importantly, this approach ensures that our study is based on participants in different stages of (expectant) parenthood. Both samples also contain participants who are not (expectant) parents, serving as a benchmark.

The first sample (N = 400; *expectant parenthood sample*) includes participants expecting a baby and participants not expecting a baby. To this end, Netquest leveraged the survey self-reports to identify all 166 participants in their panel expecting a baby. Based on their characteristics, Netquest created a matching sub-sample in terms of gender and age consisting of 234 participants not expecting a baby. Importantly, Netquest did not consider survey self-reports on parenthood when selecting participants for the *expectant parenthood sample*. Consequently, the sample includes participants with children of all ages (i.e., between 0 and 18 years as well as adult children).

In contrast, the second sample (N = 400; *parenthood sample*) includes participants having children up to the age of 12 years and participants having only adult or no children. Netquest again leveraged the survey self-reports, randomly selected 200 participants having children up to the age of 12 years, and created a matching sub-sample consisting of 200 participants having only adult or no children. Both samples were drawn in the previously described order, so that there are no duplicate participants.

Over a 24-week (*expectant parenthood sample*) and 12-week period (*parenthood sample*) in 2023 and 2024, Netquest extracted participants' metered data. For two participants, the data extraction period already started in 2021 and 2022, respectively, to increase the number of participants expecting a baby. The study was reviewed by an external ethics advisor and received approval. In addition, the study is part of a broader project which was approved by the

ethics committee of University Pompeu Fabra. Due to the sensitivity of the data, we employed a filtering approach, following the research ethical principle of "data minimization" (Adam et al., 2025). Specifically, we specified lists of websites, URL keywords, search terms, and apps potentially related to (expectant) parenthood, serving as the basis for data extraction. To gather relevant websites, we utilized multiple sources, including a web directory (i.e., https://www.curlie.org), website ranking (https://www.tranco-list.eu), as well as manual Google searches and the inspection of an existing metered dataset. To complement these sources, we instructed Netquest to extract all visits to websites whose URL contained at least one of 27 pre-specified keywords, including "baby," "pregnant," and "child." To create a list of search terms, we started with a set of general keywords (e.g., "pregnancy," "newborn," and "baby"), retrieved related search terms from Google Trends (https://trends.google.com), and enriched this list through collaborative brainstorming in our research team. Apps were collected by utilizing the data analysis platform UPUP (https://www.upup.com), allowing us to browse the top-ranking apps in Google's and Apple's app stores.

In total, Netquest extracted approx. 183,000 traces. In addition, Netquest provided us with the total number of website visits, search terms, and app usages, respectively, as well as their aggregated overall duration per participant, so that we could estimate the prevalence of PDB relative to participants' overall digital behavior.

LLM-based classifications

Our filtering approach resulted in a considerable amount of "bycatch," requiring further classification. We utilized the two open-weight LLMs Llama 3.2 3B (Meta, 2024) and Mistral 7B (Mistral AI, 2023) since they are frequently employed for text classification (Chae & Davidson, 2026; Christodoulou, 2024; Ei & Yongsiriwit, 2025; Mahajan & Rangaswamy, 2024) and can be deployed without any data sharing. The performance of LLMs heavily depends on the specific task (e.g., text classification), setting (e.g., temperature), and prompt design (e.g., few-shot). Therefore, we cannot say much a priori about the relative performance of Llama and Mistral.

The classification of websites was based on the text included in the URL. In addition, whenever possible, we scraped the URL's HTML title using the Selenium WebDriver (www.selenium.dev) in Python. As websites are constantly updated, it is often difficult to scrape URLs retrospectively (Dahlke et al., 2025). Therefore, we instructed both LLMs to prioritize the URL information to account for missing titles and potential discrepancies between the URLs and titles. The classification of search terms was simply based on their textual content. Finally, we scraped the apps' description texts from Google's and Apple's app stores, and we manually retrieved the store category for each app, so that the classification of apps was based on their names, store categories, and descriptions. We removed duplicate traces, so that each unique website (n = 56,822), search term (n = 11,309), and app (n = 115) was classified only once.

Our classification scheme was developed based on the metered data rather than using preconceived categories. The initial classification scheme consisted of seven substantive categories and an eighth category for unrelated traces, serving as the basis for the LLM-based

classifications: (1) Health (e.g., information on child and/or maternal health), (2) General (e.g., general child- and/or pregnancy-related information and support), (3) Social services (e.g., information on child benefits and parental leave), (4) Products (e.g., child- and/or pregnancy-related products), (5) Entertainment (e.g., TV shows, games, and music for children), (6) Education (e.g., learning websites and school information), (7) Activities (e.g., information on family trips and vacation with children), (8) Unrelated (e.g., news articles about pregnant celebrities). Based on these categories, we developed few-shot prompts for instructing Llama and Mistral. Although all prompts are based on the same eight categories, we tailored them to the respective task of classifying websites, search terms, and apps by including examples. The LLMs were operated on a secure university server through the Python library Ollama (https://github.com/ollama/ollama-python), using PyCharm Professional (version 2024.3). Following van Hoof et al. (2025), we deployed both LLMs with a temperature setting of 0.2. Importantly, to ensure that all categories are sufficiently prevalent in the metered data, we aggregated the initial categories into four top-level categories (or PDB dimensions) and a fifth category for unrelated traces. This final category scheme served as the basis for our analysis: (1) Health, (2) General, social services, activities, (3) Products, (4) Entertainment, education, (5) Unrelated.

As a quality assurance measure, a Spanish native speaker coded a randomly selected subset of 1% of the unique websites (n = 568), 5% of the unique search terms (n = 565), and 100% of the unique apps (n = 115), using the same initial classification scheme as the LLMs. The subset size was chosen based on time and budget constraints. After aggregating the manual codings into the final category scheme, containing the five categories outlined above, we calculated Cohen's kappa. Following the interpretative labels by Landis and Koch (1977), the results indicate an almost perfect agreement regarding apps (Llama = 0.82; Mistral = 0.91), moderate agreement regarding search terms (Llama = 0.60; Mistral = 0.49), and fair to moderate agreement regarding websites (Llama = 0.52, Mistral = 0.34).

Analytical strategy

To examine our research questions on the explainability of PDB measurements (**RQ1** to **RQ4**), it is essential to evaluate these measurements based on a sample with high variance in (expectant) parenthood. Therefore, we combined the *expectant parenthood sample* and *parenthood sample* (N = 800), allowing us to assign each participant a "profile" based on their youngest (unborn) child (if applicable): expectant parent, child between 0 and 2 years, child between 3 and 12 years, child between 13 and 17 years, or no (expectant) parent (including those having only adult children). This profile assignment was informed by participants' survey self-reports.

In total, we systematically varied four analytical decisions that are reported in Table 1: type of trace considered (**RQ1**), LLM employed (**RQ2**), dimension of PDB focused on (**RQ3**), and aggregation rule applied (**RQ4**). These analytical decisions resulted in 160 possible combinations, constituting our measurements (or dependent variables). However, we could not create 30 measurements that combine a single type of trace with relative frequency, as we lack disaggregated information on the duration of participants' overall website visits, search terms,

and app usage, respectively. For the remaining 130 measurements, we estimated linear regression models with the measurements as dependent variable and three blocks of independent variables.

First, we included dummy variables for participants' profile: expectant parent (1 = Yes), child between 0 and 2 years (1 = Yes), child between 3 and 12 years (1 = Yes), and child between 13 to 17 years (1 = Yes), with no (expectant) parent as reference. Relatedly, drawing on participants' survey self-reports, we included a dichotomous variable indicating whether participants have experienced a previous parenthood (1 = Yes), as well as another dichotomous variable indicating whether the assigned profile applies to participants' complete data extraction period (1 = Yes). Second, to control for standard sociodemographic characteristics, we included age (in years), female (1 = Yes), as well as medium (1 = Yes) and high education (1 = Yes) with low education as reference. Importantly, as existing studies indicate that women tend to engage in more PDB than men (Hiebert et al., 2021; Laws et al., 2019), we also included interaction terms between female and the dummy variables for participants' profile. Finally, we controlled for potential errors in the metered data by including a proxy variable for tracking undercoverage (1 = Yes). To this end, we used participants' survey self-reports on whether

Table 1. Overview of all analytical decisions that were systematically varied

Type of trace considered	LLM employed	Dimension of PDB focused on	Aggregation rule applied
- All three types	- Llama	- Overall	- Absolute frequency
- Website visits only	- Mistral	- Health	- Absolute duration
- Search terms only		- General, social services, activities	- Relative frequency
- App usage only		- Products	- Relative duration
		- Entertainment, education	

Note. Overall PDB was created by aggregating the four substantive top-level categories below. The fifth category for unrelated traces is not displayed because it does not represent a substantial dimension of PDB. In total, the analytical decisions resulted in 160 unique combinations (or measurements), of which 130 measurements could be empirically realized.

they frequently use a computer, smartphone, and tablet to connect to the internet, respectively, as well as paradata on participants' meter-equipped devices. Following Bosch et al. (2025), tracking undercoverage (1 = Yes) was then defined as self-reporting the use of at least one device type on which no meter is installed. Importantly, this is only a proxy variable, since participants provided information on whether they use a specific device type, but not on the number of devices. In addition, we included a numeric variable on participants' relative activity (in percent), indicating the number of days (relative to the length of the data extraction period), on which participants actively shared any traces. In doing so, we controlled for the fact that some participants may circumvent the installed tracking application by, for example, employing the incognito mode of their browser (Bosch & Revilla, 2022).

By examining the adjusted R^2 values of the 130 regression models, we determined the proportion of variance of each measurement that can be explained by the independent variables, representing our explainability measure. Since there is no gold-standard or benchmark for measuring PDB, we followed the notion that high-quality measurements of PDB should be (partially) explainable through theoretically associated variables, such as whether participants are (expectant) parents or not.

Drawing on the explainability measure, we investigate **RQ1** to **RQ4** in three consecutive steps. First, we visualize differences in the explainability of PDB measurements by plotting a bar chart with the 130 measurements on the x-axis and their adjusted R^2 values on the y-axis. Second, we present the ten measurements with the lowest adjusted R^2 values as well as the ten measurements with the highest adjusted R^2 values to show which combinations of analytical decisions have the lowest (and highest) explainability. Third, we present the results of a linear regression model using the 130 measurements as observations, their adjusted R^2 values as dependent variable, and the four analytical decisions as independent variables. This analysis allows us to identify systematic patterns between the analytical decisions under investigation and explainability across all 130 measurements. All analyses were performed in RStudio (version 2024.04.1).

Results

In the first step, we examine the explainability of the PDB measurements across all combinations of analytical decisions. To this end, Figure 1 shows a bar chart displaying the adjusted R^2 values of all 130 measurements, which are ordered from their lowest to highest value and categorized into quartiles. Overall, the adjusted R^2 values range between -0.01 and 0.19, and the first three quartiles do not exceed a value of 0.05. This indicates that the majority of measurements exhibits rather low explainability. Compared to the other quartiles, the fourth quartile shows by far the largest differences with adjusted R^2 values ranging between 0.05 and 0.19. In total, ten measurements reach adjusted R^2 values above 0.10. Importantly, the notable differences between the measurements indicate that the analytical decisions under investigation have a substantial impact on research outcomes, warranting closer examination.

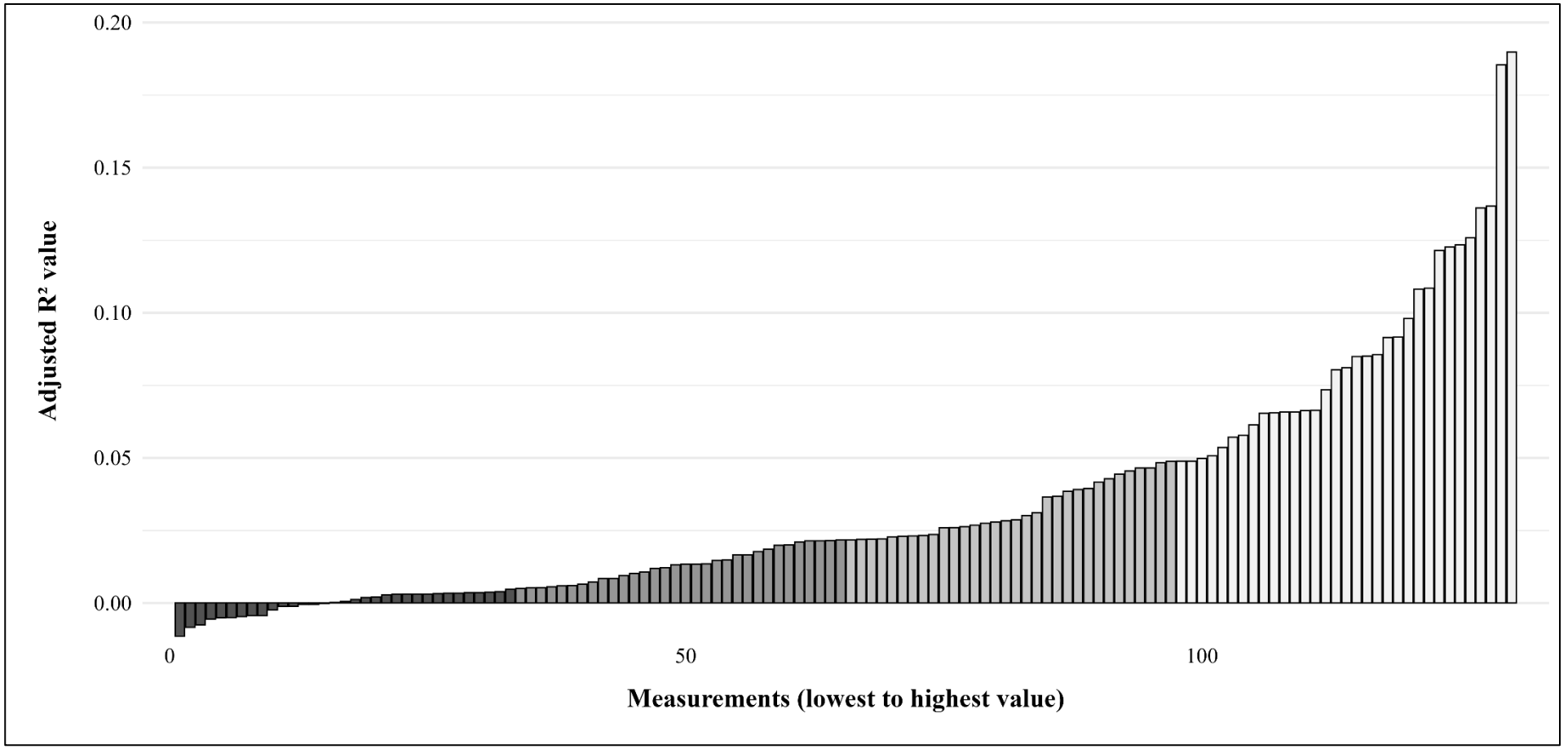

Figure 1. Adjusted R^2 values of all 130 realized PDB measurements

Note. Measurements are ordered from lowest to highest adjusted R^2 value (i.e., #1 to #130). Bars are colored according to their measurement's quartile, ranging from dark grey (1st quartile) to light grey (4th quartile).

In the second step, Table 2 shows the ten measurements with lowest and highest explainability, respectively, alongside their combinations of analytical decisions. Looking at the measurements with the lowest explainability, it is to observe that combinations including *All three types* (4/10) and *Website visits only* (4/10) are overrepresented, whereas combinations including *App usage only* do not occur at all (0/10). When it comes to the LLM employed, combinations with *Mistral* (8/10) occur more often than their counterparts with *Llama* (2/10). Furthermore, with respect to the dimension of PDB focused on, combinations including *General, social services, activities* seem to be overrepresented (6/10), and combinations including *Health* do not occur at all (0/10). Finally, all four aggregation rules are represented among the ten combinations, without any striking differences.

Now turning to the ten measurements with the highest explainability, combinations frequently include *App usage only* (8/10) but not *Website visits only* (0/10) or *Search terms only* (0/10). In contrast, both *Llama* and *Mistral* are included in the same number of combinations (5/10). There are also five pairs of identical combinations that only differ in the LLM employed, suggesting that the two LLMs may be at least partially interchangeable in the context of our measurements. However, with regard to the dimension focused on, Table 2 paints a different picture: The eight measurements with the highest explainability are exclusively based on combinations including *Health*, whereas the other dimensions do not occur at all, except for *Overall* (2/10). Finally, each aggregation rule is represented in two combinations, except for *Absolute frequency*, which is slightly overrepresented (4/10).

In the third step, to examine these patterns across all 130 measurements, we report the results of a linear regression model using the measurements as observations, their adjusted R^2 values as dependent variable, and the analytical decisions as independent variables. Figure 2 shows the results. In line with our previous analysis, the analytical decisions under investigation differ in their impact on explainability, except for the LLM employed (see Figure 2). Specifically, regarding the type of trace considered (using *All three types* as reference), *App usage only* is positively associated with explainability. With respect to the dimension of PDB focused on (using *Overall* as reference), *General, social services, activities* as well as *Entertainment, education* are negatively associated with explainability, whereas *Health* is positively associated with explainability. Furthermore, when it comes to the aggregation rule applied (using *Absolute frequency* as reference), *Relative frequency* and *Relative duration* are positively associated with explainability. In contrast, the LLM employed is not significantly associated with explainability.

Table 2. PDB measurements with lowest and highest explainability alongside their combinations of analytical decisions

	Type of trace considered	LLM employed	Dimension of PDB focused on	Aggregation rule	N	Adj. R^2
Lowest explainability						
#1	Search terms only	Mistral 7B	Products	Absolute duration	732	-0.01
#2	All three types	Mistral 7B	General, social services, activities	Relative duration	765	-0.01
#3	Search terms only	Mistral 7B	Overall	Absolute duration	732	-0.01
#4	Website visits only	Mistral 7B	Entertainment, education	Absolute duration	752	-0.01
#5	Website visits only	Llama 3.2 3B	General, social services, activities	Absolute frequency	752	-0.01
#6	Website visits only	Mistral 7B	General, social services, activities	Relative frequency	752	-0.01
#7	All three types	Llama 3.2 3B	Products	Relative duration	765	0.00
#8	Website visits only	Mistral 7B	General, social services, activities	Absolute frequency	752	0.00
#9	All three types	Mistral 7B	General, social services, activities	Relative frequency	765	0.00
#10	All three types	Mistral 7B	General, social services, activities	Absolute frequency	765	0.00
Highest explainability						
#121	App usage only	Mistral 7B	Overall	Absolute frequency	745	0.11
#122	App usage only	Llama 3.2 3B	Overall	Absolute frequency	745	0.11
#123	All three types	Llama 3.2 3B	Health	Relative duration	765	0.12
#124	App usage only	Llama 3.2 3B	Health	Relative frequency	745	0.12
#125	All three types	Mistral 7B	Health	Relative duration	765	0.12
#126	App usage only	Mistral 7B	Health	Relative frequency	745	0.13
#127	App usage only	Llama 3.2 3B	Health	Absolute frequency	745	0.14
#128	App usage only	Mistral 7B	Health	Absolute frequency	745	0.14
#129	App usage only	Llama 3.2 3B	Health	Absolute duration	745	0.19
#130	App usage only	Mistral 7B	Health	Absolute duration	745	0.19

Note. Adj. R^2 = Adjusted R^2 value. Measurements are ordered from lowest to highest R^2 value, and only the ten measurements with the lowest (#1 to #10) and highest explainability (#121 to #130) are displayed.

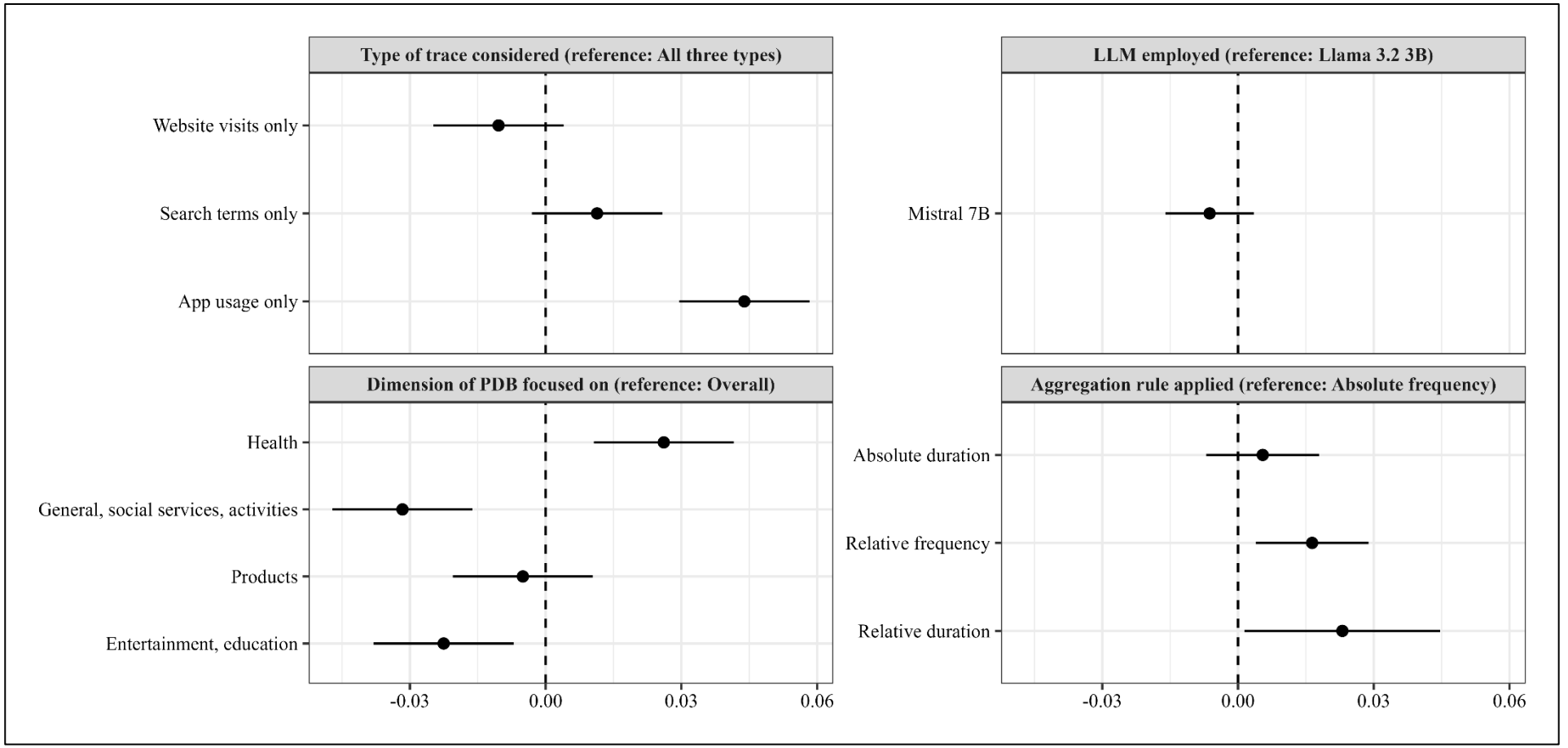

Figure 2. Linear regression model with the adjusted R^2 values as dependent variable

Note. Black dots = unstandardized coefficients. Black horizontal lines = 95% confidence intervals. N = 130 (i.e., number of measurements realized). The four analytical decisions were included as dummy variables, and the reference category is shown in parentheses, respectively.

Discussion and conclusion

The aim of this study was to investigate how traces in the form of website visits, search terms, and app usage can be transformed into meaningful measurements of PDB. To this end, we obtained up to 24 weeks of metered data alongside survey self-reports on (expectant) parenthood (N = 800) from Netquest's metered panel in Spain. By systematically varying the type of trace considered, LLM employed for classification, dimension of PDB focused on, and aggregation rule applied, we examined the impact of four key analytical decisions on the explainability of the measurements created.

Overall, we found that explainability differs considerably between PDB measurements. Across all 130 measurements (or combinations of analytical decisions), the adjusted R^2 values range between -0.01 and 0.19, and only ten measurements reach values above 0.10. These results suggest that the analytical decisions under investigation directly affect research outcomes and should therefore be made carefully. In other words, if researchers select arbitrary combinations of analytical decisions, there is a high risk of creating measurements with low explainability, threatening the validity and reproducibility of substantial results. We thus urge researchers analyzing digital trace data, such as metered and donated data, to document all analytical decisions made and, whenever possible, conduct robustness checks with alternative combinations of analytical decisions.

Importantly, the analytical decisions under investigation, except for the LLM employed, are systematically associated with explainability. Specifically, measurements based on *App usage only* are associated with higher explainability. One possible explanation is that website visits and search terms are more ambiguous than app usage, and that explainability is reduced by traces that appear to be PDB-related but are actually caused by other reasons, such as having a pregnant friend. Apps, in contrast, do not only require an active download, but are often tailored to specific target groups, such as expectant parents monitoring maternal and fetal health (Alsebayel et al., 2024). Consequently, participants who are not part of the app's explicit target group may see no reason to use such an app. However, this presumption lacks empirical evidence and needs further, more refined investigation. For example, it would be worthwhile to conduct additional qualitative interviews with participants to ask them about the context of selected traces (Gajardo et al., 2026).

When it comes to the dimension of PDB focused on, we found that *Health* is positively associated with explainability, whereas *General, social services, activities* and *Entertainment, education* are negatively associated with explainability. As for the type of trace considered, this may suggest that some dimensions of PDB are characterized by higher ambiguity. Particularly, *General, social services, activities* and *Entertainment, education* are comparatively broad dimensions encompassing a large variety of traces that can be interpreted in different ways. For example, website visits or online searches related to learning websites and school information must not necessarily be representations of PDB, as teachers, private tutors, and adult learners might access such content as well. At the same time, the positive association between *Health* and explainability is striking, given previous research showing that health information is often searched on behalf of others (Bernhörster & Reifegerste, 2025; Cutrona et al., 2015). Thus, to shed light on the mechanisms behind these associational patterns, we urge future studies to

collect additional contextual survey self-reports alongside metered data, such as questions on whom online information is typically searched for.

Turning to the aggregation rule applied, we observed that *Relative frequency* and *Relative duration* are positively associated with explainability. Relative measurements may result in higher explainability than absolute measurements because they account for the extent to which participants' digital behavior is actually collected through the meters installed. Put differently, some participants may deactivate the meter or employ their browser's incognito mode (Bosch & Revilla, 2022a) when, for example, searching for sensitive health information related to (expectant) parenthood. Since this would also reduce the overall digital behavior measured, relative measurements (partially) correct for the missing traces, whereas absolute measurements would underestimate PDB. Some research questions, such as those concerning exposure to health information, may require absolute measurements. Nevertheless, whenever possible, we strongly recommend that future researchers consider the creation of relative measurements when choosing an aggregation rule or conducting robustness checks.

In contrast to the other analytical decisions, the LLM employed was not associated with explainability, so that Llama and Mistral appeared at least partially interchangeable in the context of our measurements. Further analysis suggests that the two LLMs produced different classifications for about 40% of the unique traces. According to our results, these may not constitute systematic, but rather random differences that could possibly be attributed to the wide variety of possible interpretations of a given trace. However, these results should be interpreted cautiously as they are based on only two (open-weight) LLMs with comparatively small parameter sizes. Future studies should investigate the performance of other (proprietary) LLMs, including those with larger parameter sizes, such as GPT-5 (OpenAI, 2025).

Our study has methodological limitations, providing avenues for future research. First, we evaluated our PDB measurements by examining their explainability, but even our "best" measurement did not surpass an adjusted R^2 value of 0.19. The non-adjusted R^2 values reveal that the independent variables of our regression model could explain 21% of the best measurement's variance, leaving 79% unexplained. This could be due in part to unexplained variance within the (expectant) parenthood profiles. For example, some (expectant) parents may engage in forms of PDB that usually occur within apps and are thus not covered by our metered data approach, such as interacting in social media groups (Frey et al., 2022), watching YouTube videos (Knight et al., 2017), and seeking information through LLM-based chatbots (e.g., ChatGPT; Erol & Erol, 2026). The available survey self-reports indicated only the presence or absence of (expectant) parenthood and did not differentiate between (expectant) parents engaging in different forms of PDB. This may have limited the explanatory power of the regression models. To overcome this methodological limitation, future studies should collect direct and more fine-grained survey self-reports on PDB. In addition, to increase the coverage of PDB measurements, it would be worthwhile to supplement metered data with platform- and app-specific data donations (Boeschoten et al., 2022).

Second, and related to the previous point, Netquest identified a sufficient number of participants from the "no expectant parent," "parent," and "no parent" profiles, so that their selection was based on quality criteria, such as actively sharing traces on all days of the data extraction period. This was not possible for participants with the "expectant parent" profile,

requiring a full selection. As a result, these participants had fewer active meters. Although we included a proxy variable for tracking undercoverage as well as a variable on participants' relative activity in our regression models, this may not fully restore comparability, further reducing the explanatory power of our regression models.

Third, as we followed the research ethical principle of "data minimization" (Adam et al., 2025), metered data was pre-filtered based on lists of websites, URL keywords, search terms, and apps. However, these lists are not exhaustive, that is, they do not include all existing traces that are related to (expectant) parenthood. To examine whether and to what extent this affects our substantial results, it would be necessary to replicate our analyses, including the LLM-based classifications, based on an unfiltered metered dataset.

Finally, our study is limited by the general shortcomings of metered data. On the one hand, the metered panel used is highly selective. For example, a study by Revilla et al. (2021) found that male and younger participants are more likely to participate in the panel. On the other hand, our study is limited by tracking undercoverage and device sharing. We were able to partially control for tracking undercoverage in our regression models, but this was not possible for device sharing. In sum, our results may not generalize to the general internet population.

Overall, our study demonstrates the complexity of creating meaningful metered data measurements by highlighting the impact of various analytical decisions on research outcomes. In doing so, we lay the methodological foundation for future researchers utilizing metered data to, for example, investigate gender or socioeconomic differences in PDB. Importantly, the numerous "researcher's degrees of freedom" (Leitgöb & Keusch, 2026) associated with metered data as well as the inherent ambiguity of traces underscore that research outcomes based on digital trace data should not be considered "objective" or superior to those based on survey self-reports. Instead of replacing self-reports, we believe that digital trace data provides an alternative, supplementary epistemological perspective on social phenomena.

References

Adam, S., Makhortykh, M., Maier, M., Aigenseer, V., Urman, A., Gil Lopez, T., Christner, C., de León, E., & Ulloa, R. (2025). Improving the quality of individual-level web tracking: Challenges of existing approaches and introduction of a new content and long-tail sensitive academic solution. Social Science Computer Review, 45(5), 1050-1070. https://doi.org/10.1177/08944393241287793

Alsebayel, G., Troiano, G., & Harteveld, C. (2024). 'I believe the baby in the picture is my baby': User experiences with commercial pregnancy apps. Proceedings of the ACM on Human-Computer Interaction, 8, Article 266. https://doi.org/10.1145/3676511

Bach, R. L., & Wenz, A. (2020). Studying health-related internet and mobile device use using web logs and smartphone records. PLoS ONE, 15(6), Article e0234663. https://doi.org/10.1371/journal.pone.0234663

Baumann, I., Jaks, R., Robin, D., Juvalta, S., & Dratva, J. (2020). Parents' health information seeking behavior – does the child's health status play a role? BMC Family Practice, 21, Article 266. https://doi.org/10.1186/s12875-020-01342-3

Bernhörster, L., & Reifegerste, D. (2025). More help from friends and family? Trends in determinants of surrogate health information-seeking. Health Communication, 40(14), 3202-3212. https://doi.org/10.1080/10410236.2025.2502195

Boeschoten, L., Ausloos, J., Möller, J. E., Araujo, T., & Oberski, D. L. (2022). A framework for privacy preserving digital trace data collection through data donation. Computational Communication Research, 4(2), 388-423. https://doi.org/10.5117/CCR2022.2.002.BOES

Bosch, O. J. (2023). Validity and reliability of digital trace data in media exposure measures: A multiverse of measurement analysis. OSF Preprint. https://doi.org/10.31235/osf.io/fyr2z

Bosch, O. J., & Revilla, M. (2022a). When survey science met web tracking: Presenting an error framework for metered data. Journal of the Royal Statistical Society (Series A), 185(S2), S408–S436. https://doi.org/10.1111/rssa.12956

Bosch, O. J., & Revilla, M. (2022b). The challenges of using digital trace data to measure online behaviors: Lessons from a study combining surveys and metered data to investigate affective polarization. Sage Research Methods. https://doi.org/10.4135/9781529603644

Bosch, O. J., Sturgis, P., Kuha, J., & Revilla, M. (2025). Uncovering digital trace data biases: Tracking undercoverage in web tracking data. Communication Methods and Measures, 19(2), 157-177. https://doi.org/10.1080/19312458.2024.2393165

Charron, J., Gautier, A., & Jestin, C. (2020). Influence of information sources on vaccine hesitancy and practices. Médecine et Maladies Infectieuses, 50(8), 727-733. https://doi.org/10.1016/j.medmal.2020.01.010

Chae, Y., & Davidson, T. (2026). Large Language Models for text classification: From zero-shot learning to instruction-tuning. Sociological Methods & Research, 55(2), 501-567. https://doi.org/10.1177/00491241251325243

Christodoulou, C. (2024). NLPDame at Climate Activism 2024: Mistral sequence classification with PEFT for hate speech, targets and stance event detection. Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, 96-104. https://doi.org/10.18653/v1/2024.case-1.13

Cutrona, S. L., Mazor, K. M., Vieux, S. N., Luger, T. M., Volkman, J. E., & Rutten, L. J. F. (2015). Health information-seeking on behalf of others: Characteristics of 'surrogate seekers.' Journal of Cancer Education, 30, 12-19. https://doi.org/10.1007/s13187-014-0701-3

Dahlke, R., Kumar, D., Durumeric, Z., & Hancock, J. T. (2025). Quantifying the systematic bias in the accessibility and inaccessibility of web scraping content from URL-logged web-browsing digital trace data. Social Science Computer Review, 43(5), 1071-1086. https://doi.org/10.1177/08944393231218214

De Rosso, S., Nicklaus, S., Ducrot, P., & Schwartz, C. (2022). Information seeking of French parents regarding infant and young child feeding: Practices, needs, and determinants. Public Health Nutrition, 25(4), 879-892. https://doi.org/10.1017/S1368980021003086

Ei, L. M., & Yongsiriwit, K. (2025). Phishing email classification using few-shot learning with a lightweight LLM (Llama 3.2). 9th International Conference on Information Technology, 75-80. https://doi.org/10.1109/InCIT66780.2025.11276030

Erol, M., & Erol, A. (2026). Is it a reliable guide or a source to be approached with caution? GenAI in parents' information search processes regarding their children. European Journal of Pediatrics. https://doi.org/10.1007/s00431-026-07146-4

Freelon, D. (2014). On the interpretation of digital trace data in communication and social computing research. Journal of Broadcasting & Electronic Media, 58(1), 59-75. https://doi.org/10.1080/08838151.2013.875018

Frey, E., Bonfiglioli, C., Brunner, M., & Frawley, J. (2022). Parents' use of social media as a health information source for their children: A scoping review. Academic Pediatrics, 22(4), 526-539. https://doi.org/10.1016/j.acap.2021.12.006

Gajardo, C., Kalogeropoulos, A., Rossini, P., Mont'Alverne, C., & Nicholls, T. (2026). Evaluating the potential and limits of trace-based interviews when studying news use. Information, Communication & Society. https://doi.org/10.1080/1369118X.2026.2652514

García-Mainar, I., Molina, J. A., & Montuenga, V. M. (2011). Gender differences in childcare: Time allocation in five European countries. Feminist Economics, 17(1), 119-150. https://doi.org/10.1080/13545701.2010.542004

Gonzáles-Bailón, S., & Xenos, M. A. (2023). The blind spots of measuring online news exposure: A comparison of self-reported and observational data in nine countries. Information, Communication & Society, 26(10), 2088-2106. https://doi.org/10.1080/1369118X.2022.2072750

Hiebert, B., Hall, J., Donelle, L., Facca, D., Jackson, K., & Stoyanovich, E. (2021). 'Let me know when I'm needed': Exploring the gendered nature of digital technology use for health information seeking during the transition to parenting. Digital Health, 7, 1-8. https://doi.org/10.1177/20552076211048638

Hölzl, J., Keusch, F., & Sajons, C. (2025). The (mis)use of Google Trends data in the social sciences: A systematic review, critique, and recommendations. Social Science Research, 126, Article 103099. https://doi.org/10.1016/j.ssresearch.2024.103099

Jaks, R., Baumann, I., Juvalta, S., & Dratva, J. (2019). Parental digital health information seeking behavior in Switzerland: A cross-sectional study. BMC Public Health, 19, Article 225. https://doi.org/10.1186/s12889-019-6524-8

Keusch, F., Bach, F., & Cernat, A. (2023). Reactivity in measuring sensitive online behavior. Internet Research, 33(3), 1031-1052. https://doi.org/10.1108/INTR-01-2021-0053

Keusch, F., Bähr, S., Haas, G. C., Kreuter, F., Trappmann, M., & Eckman, S. (2022). Non-participation in smartphone data collection using research apps. Journal of the Royal Statistical Society (Series A), 185(S2), S225-245. https://doi.org/10.1111/rssa.12827

Kirkizh, N., Ulloa, R., Stier, S., & Pfeffer, J. (2024). Predicting political attitudes from web tracking data: A machine learning approach. Journal of Information Technology & Politics, 21(4), 564-577. https://doi.org/10.1080/19331681.2024.2316679

Knight, K., van Leeuwen, D. M., Roland, D., Moll, H. A., & Oostenbrink, R. (2017). YouTube: Are parent-uploaded videos of their unwell children a useful source of medical information for other parents? Archives of Disease in Childhood, 102, 910-914. https://doi.org/10.1136/archdischild-2016-311967

Kostina, A., Dikaiakos, M. D., Stefanidis, D., & Pallis, G. (2025). Large Language Models for text classification: Case study and comprehensive review. arXiv. https://arxiv.org/abs/2501.08457v1

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology, 5(3), 213-236. https://doi.org/10.1002/acp.2350050305

Kulshrestha, J., Oliveira, M., Karacalik, O., Bonnay, D., & Wagner, C. (2021). Web routineness and limits of predictability: Investigating demographic and behavioral differences using web tracking data. Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, 15(1), 327-338. https://doi.org/10.1609/icwsm.v15i1.18064

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174. https://doi.org/10.2307/2529310

Laws, R., Walsh, A. D., Hesketh, K. D., Downing, K. L., Kuswara, K., & Campbell, K. J. (2019). Differences between mothers and fathers of young children in their use of the internet to support healthy family lifestyle behaviors: Cross-sectional study. Journal of Medical Internet Research, 21(1), Article e11454. https://doi.org/10.2196/11454

Lee, T. T. (2023). Social class, intensive parenting norms and parental values for children. Current Sociology, 71(6), 964-981. https://doi.org/10.1177/00113921211048531

Leitgöb, H., & Keusch, F. (2026). Causal inference from digital behavioral data: Methodological implications. Cologne Journal of Sociology and Social Psychology, 77. https://doi.org/10.1007/s11577-026-01050-3

Livingstone, S., Blum-Ross, A., Pavlick, J., & Ólafsson, K. (2018). In the digital home, how do parents support their children and who supports them? Parenting for a Digital Future: Survey Report 1. Department of Media and Communications, The London School of Economics and Political Science, London, UK. http://eprints.lse.ac.uk/id/eprint/87952

Lupton, D., & Pedersen, S. (2016). An Australian survey of women's use of pregnancy and parenting apps. Women and Birth, 29(4), 368-375. http://dx.doi.org/10.1016/j.wombi.2016.01.008

Mahajan, S., & Rangaswamy, N. (2024). Extreme speech classification in the era of LLMs: Exploring open-source and proprietary models. In L. Garg, N. Kesswani, & I. Brigui (eds.), AI Technologies for Information Systems and Management Science (pp. 157-166). Springer. https://doi.org/10.1007/978-3-031-95017-9_15

Makhortykh, M., Urman, A., Gil-Lopez, T., & Ulloa, R. (2022). To track or not to track: Examining perceptions of online tracking for information behavior research. Internet Research, 32(7), 260-279. https://doi.org/10.1108/INTR-01-2021-0074

Martinez-Serra, A., & Cardenal, A. S. (2025). Who watches porn? Demographic insights from web tracking data. Computers in Human Behavior, 172, Article 108731. https://doi.org/10.1016/j.chb.2025.108731

Meta (2024). The Llama 3 herd of models. arXiv. https://doi.org/10.48550/arXiv.2407.21783

Mistral AI (2023). Mistral 7B. arXiv. https://arxiv.org/abs/2310.06825v1

Moungui, H. C., Nana-Djeunga, H. C., Anyiang, C. F., Cano, M., Postigo, J. A. R., & Carrion, C. (2024). Dissemination strategies for mHealth apps: Systematic review. JMIR mHealth and uHealth, 12, Article e50293. https://doi.org/10.2196/50293

Neuberger, L. (2016). Self-reports of information seeking: Is social desirability in play? Atlantic Journal of Communication, 24(4), 242-249. http://dx.doi.org/10.1080/15456870.2016.1208661

Ochoa, C., Melero, L. F., & Revilla, M. (2026). Tracing ChatGPT use in Spain: Differences among population groups in adoption and engagement. Human Behavior and Emerging Technologies. https://doi.org/10.1155/hbe2/9703008

Ochoa, C., & Revilla, M. (2025a). The utility of digital trace data to understand how people search for jobs online. Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 168(1), 65-92. https://doi.org/10.1177/07591063251349382

Ochoa, C., & Revilla, M. (2025b). Variability of a job search indicator induced by operationalization decisions when using digital traces from a meter. PLoS One, 20(12), Article e0338894. https://doi.org/10.1371/journal.pone.0338894

OpenAI (2025). OpenAI GPT-5 system card. arXiv. https://doi.org/10.48550/arXiv.2601.03267

Reiss, M. V. (2023). Dissecting non-use of online news. Systematic evidence from combining tracking and automated text classification. Digital Journalism, 11(2), 363-383. https://doi.org/10.1080/21670811.2022.2105243

Revilla, M. (2022). How to enhance web survey data using metered, geolocation, visual and voice data? Survey Research Methods, 16(1), 1-12. https://doi.org/10.18148/srm/2022.v16i1.8013

Revilla, M., Couper, M. P., Paura, E., & Ochoa, C. (2021). Willingness to participate in a metered online panel. Field Methods, 33(2), 202-216. https://doi.org/10.1177/1525822X20983986

Revilla, M., Ochoa, C., Höhne, J. K., & Couper, M. P. (2026). Transcribing and coding voice answers obtained in web surveys: Comparing three leading automatic speech recognition tools and human versus LLM-based coding. Journal of Survey Statistics and Methodology. https://doi.org/10.1093/jssam/smaf028

Revilla, M., Ochoa, C., & Loewe, G. (2017). Using passive data from a meter to complement survey data in order to study online behavior. Social Science Computer Review, 35(4), 521-536. https://doi.org/10.1177/0894439316638457

Revilla, M., Toninelli, D., Ochoa, C., & Loewe, G. (2016). Do online access panels need to adapt surveys for mobile devices? Internet Research, 26(5), 1209-1227. https://doi.org/10.1108/IntR-02-2015-0032

Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. Public Opinion Quarterly, 85(S1), 399-422. https://doi.org/10.1093/poq/nfab018

Slomian, J., Bruyère, O., Reginster, J. Y., & Emonts, P. (2017). The internet as a source of information used by women after childbirth to meet their need for information: A web-based survey. Midwifery, 48, 46-52. https://doi.org/10.1016/j.midw.2017.03.005

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. Perspectives on Psychological Science, 11(5), 702-712. https://doi.org/10.1177/1745691616658637

Than, N., Fan, L., Law, T., Nelson, L. K., & McCall, L. (2025). Updating "The future of coding": Qualitative coding with generative Large Language Models. Sociological Methods & Research, 54(3), 849-888. https://doi.org/10.1177/00491241251339188

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The psychology of survey response. Cambridge University Press.

Van Hoof, M., Trilling, D., Meppelink, C., Möller, J., & Loecherbach, F. (2025). Googling politics? Comparing five computational methods to identify political and news-related searchers from web browser histories. Communication Methods and Measures, 19(1), 63-89. https://doi.org/10.1080/19312458.2024.2363776

Von der Heyde, L., Haensch, A.-C., Weiß, B., & Daikeler, J. (2025). Using Large Language Models for coding German open-ended survey responses on survey motivation. Survey Research Methods, 19(4), 355-370. https://doi.org/10.18148/srm/2025.v19i4.8568

Von Hohenberg, B. C., Stier, S., Cardenal, A. S., Guess, A. M., Menchen-Trevino, E., & Wojcieszak, M. (2024). Analysis of web browsing data: A guide. Social Science Computer Review, 42(6), 1479-1504. https://doi.org/10.1177/08944393241227868

Wenz, A., Keusch, F., & Bach, R. L. (2024). Measuring smartphone use: Survey versus digital behavioral data. Social Science Computer Review, 43(5), 1030-1049. https://doi.org/10.1177/08944393231224540

Wirz, D. S., de Léon, E., Adam, S., & Makhortykh, M. (2025). Tracing knowledge gaps: Investigating the influence of education on news exposure and knowledge using digital trace data. The International Journal of Press/Politics. https://doi.org/10.1177/19401612251335372

Yudianto, B., Caldwell, P. H. Y., Nanan, R., Barnes, E. H., & Scott, K. M. (2023). Patterns of parental online health information-seeking behavior. Journal of Paediatrics and Child Health, 59, 743-752. https://doi.org/10.1111/jpc.16387