mda

methods, data, analyses

JOURNAL FOR QUANTITATIVE METHODS AND SURVEY METHODOLOGY

Volume 19, 2025 | 2

Exploring the Methodological Choices, Challenges and Solutions in Working with New Data Sources

Rachel Gibson and Trent D. Buskirk (Editors)

methods, data, analyses is published by GESIS – Leibniz Institute for the Social Sciences.

Managing editors: Barbara Felderer & Maximilian Linde
GESIS – Leibniz Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany

E-mail: mda@gesis.org
https://majournals.bib.uni-mannheim.de/mda

GESIS Leibniz Institute
for the Social Sciences

Copyediting: Désirée Nießen, Barbara Felderer, Maximilian Linde
Layout: Liebchen + Liebchen
ISSN 2190-4936 (Online)

Content

RESEARCH REPORTS

methods, data, analyses | Vol. 19(2), 2025, pp. 4-10

Editorial
An Overview of the Special Issue

Rachel Gibson[1] & Trent D. Buskirk[2]

[1] *University of Manchester*
[2] *Old Dominion University*

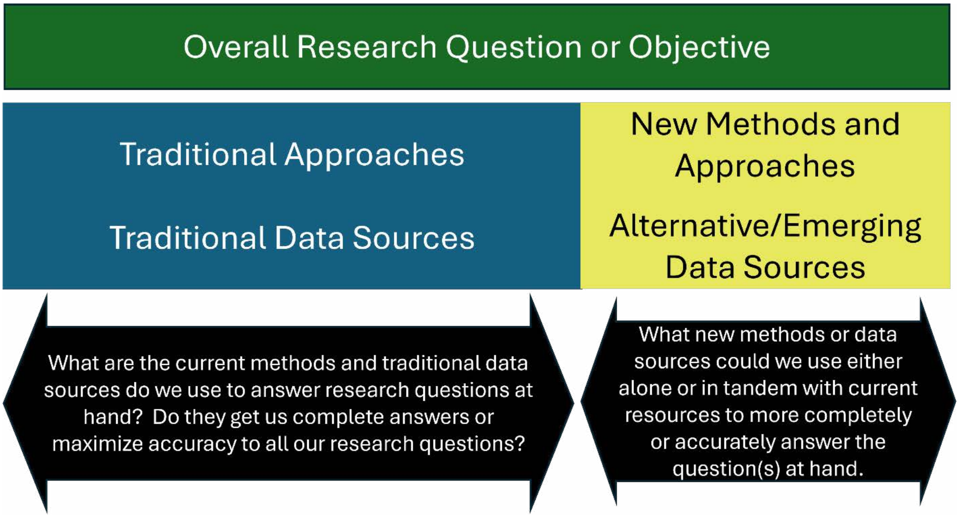
Across the quantitative social sciences, researchers increasingly face significant challenges and opportunities prompted by the arrival of new sources of very rich, highly granular, and often unstructured digital data. While traditional methods such as surveys and content analysis tools remain indispensable for measuring individual attitudes, behaviors, demographic characteristics, and media messaging online, they often struggle to capture the complex multimodal information streams and metadata generated by social media platforms, mobile devices, sensors, and tracking applications. Collecting and analyzing these diverse new forms of content, dynamic moment-to-moment behaviors, and naturally occurring interactions has become a pressing and exciting research task—one that holds real promise for answering longstanding research questions more completely and, in some cases, more accurately. Yet as access to these data grows, so too do the problems they pose in terms of representativeness, potential new sources of bias, complex preprocessing demands, and reproducibility.

One increasingly common response to these challenges is to link or "anchor" emerging data sources to more structured, researcher-designed forms of data, particularly surveys. Doing so offers two complementary benefits. First, traditional instruments can provide context, validation, and interpretability for new

The Special Issue is generated in part from the Digital Campaigning and Electoral Democracy (DiCED) Project. Grant agreement no. 833177 funded by "European Research Council (ERC) European Union's Horizon 2020 research and innovation programme

forms of behavioral or multimodal data. Second, emerging data sources can enrich surveys by extending coverage in time, modality, or behavioral detail, thereby filling gaps that conventional approaches alone may leave unaddressed.

We can frame the connections between the papers in this special issue using the motivating schematic depicted in Figure 1**.** Specifically, each study begins with a substantive research question for which traditional approaches and data sources—such as surveys, curated observational data or designed data, or conventional content analysis—could plausibly be used. However, in each case, the authors identify limitations in relying on these approaches alone, whether due to recall error, restricted temporal resolution, limited measurement scope, or difficulty capturing visual, behavioral, or contextual information. To address these limitations, the papers pursue two broad strategies. Some introduce **new analytical methods applied to existing data**, enabling researchers to model previously unaccounted-for sources of error or extract richer information from conventional inputs. Others incorporate **new or emerging data sources alongside established methods**, using digital traces, images, metadata, or real-time behavioral signals to improve the completeness, accuracy, or interpretability of the resulting analyses. In both cases, methodological innovation is driven not by novelty for its own sake, but by the goal of producing better answers to well-defined research questions.


Overall Research Question or Objective

Traditional Approaches

Traditional Data Sources

New Methods and Approaches

Alternative/Emerging Data Sources

What are the current methods and traditional data sources do we use to answer research questions at hand? Do they get us complete answers or maximize accuracy to all our research questions?

What new methods or data sources could we use either alone or in tandem with current resources to more completely or accurately answer the question(s) at hand.

Figure 1 A motivating schematic for thinking about the use of new and emerging data sources in conjunction with more traditional methods and sources.

This special issue brings together a set of papers that advance the field along precisely this shared dimension—**the deliberate extension of traditional quantitative approaches through new data, new methods, or both**—to reach more complete, accurate, or informative conclusions about individual preferences, activities, and attitudes. Each article is accompanied by a "reflective methodological appendix," in which authors are asked to "lift the hood" on their research process and document the choices, constraints, adaptations, and trade-offs encountered as their projects unfolded. These reflections make visible the iterative decision-making that typically remains hidden in published research and align closely with broader efforts to promote open and reflexive research practices within the social sciences. Importantly, the reflective appendices underscore a core motivation for this volume: emerging data sources come with significant qualifiers. None are genuinely "free." Across projects, researchers invested substantial effort in collection, cleaning, processing, linkage, and interpretation—often operating within constraints set by platform architectures, proprietary systems, device limitations, and finite computational infrastructure. Moreover, because these data depart from long-standing survey norms in structure, stability, and population coverage, authors repeatedly had to interrogate assumptions and adapt designs as the work unfolded. In these appendices we asked authors to go behind the veil of their papers—particularly the methods, data, and analyses components—to reflect on the pathway from inception to final publication. The result is a set of statements that make visible the decisions, routing, and particularly re-routing of research designs that would otherwise go unreported. Through this process, we hope this collection of papers and appendices can offer practical guidance to scholars embarking on similar projects and to contribute to a more transparent public dissection of the iterative dynamics that underpin research using new and emerging data sources.

Below we begin with an overview of the papers as a whole and how they provide alternative approaches to addressing a common challenge of linking established pre-digital data and methods with newer digital-exclusive versions. We also highlight their key findings as stand-alone pieces of research and their substantive value and contribution to their respective fields. We then turn our focus to the reflective appendices to identify the dilemmas and challenges that authors faced in investigating their research questions—and, importantly, how they addressed them in practice. We conclude by drawing out broader lessons learned from these experiences for future scholarship navigating the frontier of linked and augmented data analysis.

Emerging Data Types and New Modes of Observation

What unites the papers in this issue is a shared recognition that emerging data sources both complement and complicate survey-based research. Sometimes these sources can supplement surveys by filling coverage gaps or validating self-reports. In other cases, they introduce entirely new modes of measurement—visual, behavioral, or moment-triggered—that fundamentally reconfigure what social scientists can observe. In still others, they strain existing methodologies, requiring innovations in data processing, linkage, or research design. The five contributions showcased here span methodological innovations—from augmented Data Download Packages and real-time event-triggered surveys to mixture modeling for linkage errors and systematic coding of multimodal political appeals. Collectively, these studies address persistent challenges in data quality, representativeness, and analytical rigor by integrating diverse data streams, including digital trace data, visual content, and survey responses. Common themes include the pursuit of richer, more accurate insights into human behavior and communication, the development of tools to mitigate bias and error, and the expansion of research beyond traditional text-based and retrospective approaches.

The first paper by **Wedel, Ohme, and Araujo**, *Augmenting Data Download Packages—Integrating Data Donations, Video Metadata, and the Multimodal Nature of Audio-visual Content*, introduces Augmented Data Download Packages (aDDPs) as a novel way to enrich conventional digital trace data. By integrating survey responses, metadata, and multimodal content embeddings, aDDPs provide a more comprehensive view of user behavior. Using TikTok as a case study, the authors show how these enhanced packages enable nuanced analyses of engagement patterns and content classification, illustrating the potential of combining behavioral and self-reported data for social science research. Building on the theme of multimodality, the second paper by **Iglesias**, *Preferences, Participation, and Evaluation of Answering Questions About the Books Participants Have at Home Through Conventional and Image-Based Formats*, examines the role of visual data in survey design by comparing photo-based questions to conventional formats. Drawing on a large-scale mobile survey of Spanish parents, the study shows that while respondents generally prefer traditional questions, those who favor images engage more when given choice. Demographic and behavioral predictors of participation underscore the complexity of integrating visual tasks into surveys and highlight the need for adaptive designs that accommodate diverse respondent preferences. The third contribution by **Ochoa**, *Researching the Moment of Truth: An Experiment Comparing In-the-Moment and Conventional Web Surveys to Investigate Online Job Applications*, advances this discussion by exploring real-time, event-triggered surveys linked to metered data. Focusing on online job applications, the authors test whether surveys delivered immediately after detected events can improve data quality and reduce recall bias. The study finds strong acceptance of this approach and richer responses compared to conventional surveys, while also showing that memory-related errors may persist even under improved timing—highlighting both the value and limits of timely interventions for capturing accurate behavioral data.

While these papers focus on enhancing survey-based research through new data and measurement modes, the fourth study by **West, Slawski, and Ben-David**, *Improved Ensemble Predictive Modeling Techniques for Linked Social Media and Survey Data Sets Subject to Mismatch Error,* reminds us that linking data sources together, while expanding the scope of information, is not without error. This paper specifically addresses a critical methodological issue in linked datasets: mismatch errors introduced by probabilistic linkage. The authors propose a mixture modeling approach to adjust predictive modeling outputs when such errors occur, testing it on Twitter activity linked to survey-based ideology measures. Their method successfully recovers predictive performance, underscoring the importance of explicitly correcting linkage uncertainty in an era of increasingly complex data integration.

Finally, the fifth paper by **Cashell**, *Improving Assessments of Group-Based Appeals in Political Campaigns by Systematically Incorporating Visual Components of Ads*, extends the conversation on multimodality by introducing a coding scheme for visual and textual group-based appeals in political campaigns. Applying the schema to thousands of images from U.S. House races, the study reveals that indirect visual cues are as prevalent as direct mentions, often used in combination. By systematically capturing visual indicators, the framework reduces bias in measuring group targeting strategies and provides a more complete picture of modern political communication than using the current text-only approaches alone.

Taken altogether, these studies illustrate a shared focus on methodological innovation and data enrichment. They show how linking and augmenting data sources—whether through multimodal content, real-time triggers, or error-adjusted models—can overcome limitations of traditional approaches and open new avenues for research. As digital environments continue to evolve, the approaches showcased in this special issue chart a path toward more robust, adaptive, and multimodal research designs.

Reflections on the Reflective Appendices

The reflective appendices reveal a set of shared methodological themes that reinforce the central motivation of the special issue: the promise of emerging data comes tightly coupled with practical and inferential challenges that often surface most sharply after the initial research design is on paper. Across the five projects, authors repeatedly emphasize that key constraints emerged during data collection, processing, and integration—through lower-than-anticipated participation, uneven data completeness, technical constraints imposed by platforms or devices, and greater-than-expected demands for infrastructure and manual labor. These experiences reinforce that emerging data sources are rarely “plug-

and-play" alternatives to surveys; instead, they require substantial investment, flexibility, and explicit feasibility assessment as projects unfold. The appendices also underscore that some of the most consequential challenges—such as low uptake in burdensome tasks, unstable triggering conditions, linkage uncertainty, or inconsistencies in human coding—cannot be "fixed" purely in the analysis stage but instead shape what can credibly be claimed and what analyses are ultimately possible.

At the same time, the appendices highlight distinctive challenges that are specific to particular data types and integration strategies. Image-based approaches raise issues of respondent burden, privacy sensitivity, and the labor-intensity of classification, whether manual or automated. Behavioral trace and in-the-moment designs are highly dependent on platform architectures, URL stability, and operating-system capabilities, requiring sustained technical attention and ongoing monitoring of the data-generating process. Data donation work points to platform governance and user verification processes as practical bottlenecks, while also raising the likelihood of self-selection and the need to think carefully about what populations and phenomena can be studied credibly given participation patterns. Linked survey–social media analyses emphasize that even when linkage appears strong, mismatch error and linkage quality remain central threats to inference and must be modeled explicitly rather than treated as a secondary issue. Together, these reflections complement Figure 1 by showing that "adding" new data or methods does not simply expand what can be studied; it also introduces new constraints that must be documented and weighed as part of any fitness-for-use assessment.

Emerging Frameworks for Fitness-for-Use

Several papers demonstrate that emerging data may serve as enhancements rather than replacements for surveys. Visual data enriches communication research; multimodal DDPs broaden analytic possibilities; image-based responses provide detailed, ecologically grounded measurement for otherwise difficult-to-measure household inventories; and triggered surveys align measurement more closely with real-world behavior. Others underscore the need for caution and explicit adjustment when new forms of error are introduced, as in linkage mismatch error and predictive modeling. Across the issue, the resulting stance is pragmatic: emerging data sources neither uniformly surpass nor simply replicate surveys. Their value depends on context, construct, and research design—and on careful evaluation of fitness-for-use relative to the inferential goals at hand.

Developing a Culture of Methodological Reflexivity and Transparency

Beyond the substantive and methodological contributions of the five studies, the special issue advances open scientific practice by promoting “reflexive praxis”—researchers documenting the choices, complications, and adjustments shaping their approach and analysis. Each paper includes a peer-edited appendix offering insights into data access challenges, technical constraints, processing bottlenecks, linkage and measurement vulnerabilities, and the iterative redesign that often accompanies work with emerging sources. These reflections are intended as practical guides and as a step toward making the methodological pathway of digitally enabled social science more transparent and cumulative.

Conclusion

Together, the contributions demonstrate a future in which ‘old’ and emerging data sources and methods of data collection operate in tandem. Structured forms of data collected through surveys and established content analysis tools remain essential for representativeness, comprehensiveness and intentional measurement; emerging digital data provide new contextual and visual richness, temporal precision, and behavioral grounding. Yet integration requires methodological imagination, transparency, and rigor—and an honest accounting of the constraints and trade-offs involved. This special issue aims to contribute to that agenda by showing not only what these approaches can achieve, but also what is required to implement them responsibly and interpret them appropriately.

methods, data, analyses | Vol. 19(2), 2025, pp. 11-45 DOI: 10.12758/mda.2024.08

Augmenting Data Download Packages – Integrating Data Donations, Video Metadata, and the Multimodal Nature of Audio-visual Content

Lion Wedel[1], Jakob Ohme[1] & Theo Araujo[2]

[1] *Weizenbaum Institute for the Networked Society, Germany*

[2] *Amsterdam School of Communication Research, Netherlands*

Abstract

This research explores the potential of augmented Data Download Packages (aDDPs) as a novel approach to analyze digital trace data, using TikTok as a use case to demonstrate the broader applicability of the method. The study demonstrates how these data packages can be used in social science research to understand better user behavior, content consumption patterns, and the relationship between self-reported preferences and actual digital behavior.

We introduce the concept of aDDPs, which extend the conventional Data Download Packages (DDPs) by augmenting the collected data with survey data, metadata, content data, and multimodal content embeddings, among other possibilities - rendering aDDPs an unprecedentedly rich data source for social science research. This work provides an overview and guidance on collecting, augmenting DDPs, and analyzing the resulting aDDPs.

In a pilot study on 18 aDDPs, we use the combination of data components in aDDPs to facilitate research on user engagement behavior and content classification. We showcase the potential of the information breadth and depth that aDDPs depict by exploiting the combination of multimodal content embeddings, the users' watch history, and survey data. To do so, we train and compare uni- and multimodal classifiers, classify the 18 aDDPs' videos, and investigate the extent to which user engagement behavior impacts future content suggestions. Furthermore, we compare the users retrieved content with the users' self-reported content consumption.

Keywords: data download packages, augmentation, multimodality, TikTok, vertical videos, classification

TikTok is one of the fastest-growing social media platforms worldwide (Newman et al., 2023). In addition, its role in distributing information during the COVID-19 crisis and the Russian invasion of Ukraine, as well as the discussions around its Chinese ownership, manifests the understanding that the platform needs to be considered relevant for social media researchers of many fields (e.g., Basch et al., 2020; Primig et al., 2023). The European Commission has recently recognized this relevance, assigning TikTok the status of a very large online platform (VLOP), which can carry systemic risk for the European Union (*DSA*, 2023). As a vertical video platform (VVP), TikTok's main characteristics are short vertical videos (recorded in portrait mode) and the substantial reliance on algorithmic curation and passive use compared to other social media platforms (Hase et al., 2022). Unlike Twitter or Facebook, TikTok content is inherently multimodal beyond text and an occasional picture – consisting of audio-visual information. This creates new challenges and opportunities for computational social sciences and adjacent fields.

The EU General Data Protection Regulation (GDPR, 2016) allows users to demand the data TikTok has collected about them (TikTok, 2023b). Similar laws exist in countries and regions beyond the EU, such as Japan or Brazil (Boeschoten et al., 2020). The access explicitly allows sharing data with *"...third parties, such as social scientists."* (ibid., p. 4). This is the foundation to explore the potential of data donations for user-centered research purposes. Still, research utilizing Data Download Packages (DDPs) from video platforms like TikTok is sparse, given the expected difficulties of retrieving and analyzing the multimodal nature (i.e., moving images, audio, and text) of (vertical) videos. Specifically, it is difficult for social science research to understand exposure patterns based on data donations. It is, therefore, essential to develop new approaches to understand the content that, within the EU alone, around 135.9 million users are exposed to monthly (TikTok, 2023a).

This paper explores the potential of augmented DDPs (aDPPs) for social science researchers to study information exposure and conduct algorithmic auditing on TikTok. It presents a new approach, integrating TikTok DDPs with 1) survey data, 2) video metadata, 3) content data, and 4) the multimodal features of a TikTok post. Previous research has identified multiple challenges to arrive at a meaningful basis for social science research that allows the analysis of vertical video platform exposure data with DDPs (Boeschoten et al., 2021; Driel et al., 2022; Ohme et al., 2021). While we leave some of those unaddressed (e.g., sample biases and conversion rates of successful donation), we describe two challenges on the way to an augmented TikTok data download package: 1) the data donation

Direct correspondence to
Lion Wedel, Weizenbaum Institute for the Networked Society, Berlin, Germany
E-mail: lion.wedel@weizenbaum-institut.de

process and 2) the augmentation of DDPs. Subsequently, we provide solutions for tackling the described challenges in a pilot study. The concept of aDDPs is not limited to TikTok. It can serve as a guiding concept for research using data donations from any social media and content platform where the native DDP does not hold sufficient information to answer the proposed research questions.

In the following, we will first explain the background and relevance of the topic before we explain how TikTok data donations can be augmented with specifically multimodal content features. In the last exploratory part, we show how aDDPs can be used in social science research to answer substantial questions, such as how previous engagement affects future suggested content and whether user perceptions of their information consumption align with the empirical findings.

TikTok's Inherent Multimodality and the Potential of aDDPs

Over the last decade, multimedia content has increased in importance in delivering media messages to users and audiences. In this context, muti-modality describes the combination of different modes of content, such as *"… language, images, typography [or] layout …"* in a media format (Hiippala, 2017, p.421). Since their emergence, text and still images have been the predominant modes of content presentation on digital platforms, often in separated elements. With vertical video features such as Instagram Stories, Snapchat Spotlight, YouTube Shorts, and TikTok as the dominant vertical video-only platform, moving image is combined with audio tracks. This multimodality is further enhanced by integrating still images, icons, and text, such as hashtags or subtitles. This integration of different content modes in the format of a video challenges existing media analysis paradigms (e.g. Valkenburg, 2022) and calls for new approaches to preparing multimodal content for analysis. TikTok's platform logic is based on videos with audio and a description – thereby inherently multimodal (Hase et al., 2022).

Social scientists have a clear interest in researching video-only platforms such as TikTok but often retreat methodologically to qualitative methods (e.g., Mordecai, 2023; Zhou Ting, 2021), especially considering the complexity of multimodal data. Here, a set of contributors usually manually labels a sample of videos (e.g., Li & Kang, 2023; Ming et al., 2023; Ng & Indran, 2023; Yeung et al., 2022). Labeling posts for social science research aligns with a classification task in machine learning. Hence, the collection of DPPs and their augmentation are the first two steps. In the third step, a large-scale classification model is necessary to unfold the potential of aDDPs for critical research social scientists seek concerning video-only platforms.

A handful of contributions acknowledge the multimodality of TikTok videos, and the consequent contribution to the development of uni- and multimodal classifiers must be mentioned here. With *SexTok,* George & Surdeanu (2023) present a 1,000 video dataset on which they train separately a text and a video embedding-based classifier to predict one of three classes. Other pieces on TikTok videos extract text shown within the video or focus only on the audio feature to classify videos subsequently (e.g., Fiallos et al., 2021; Ibañez et al., 2021). Such work relies on one modality, ignoring the information depth other modalities could add. Kim et al. (2023) embrace TikTok as a multimodal platform but eventually reduce videos to thumbnails and audios to transcripts – in both cases, scrutinizing the information depth those modes might entail. Nevertheless, they showcase that using variables retrieved through pre-trained classifiers as the basis for scalable classification and subsequent analysis, such as hypothesis testing, is a feasible approach for research on TikTok and possibly other video-only platforms.

Research across domains has consistently shown that incorporating all available modalities improves the performance of classification tasks (e.g., Pandeya & Lee, 2021; Qi et al., 2023; Shang et al., 2021). Specifically in the application of social media posts, multimodal approaches have been proven to equalize weaknesses of unimodal representations in Instagram posts (Zeppelzauer & Schopfhauser, 2016). A truly multimodal classification approach to TikTok videos is presented by Shang et al. (2021), who take visual content, audio, video descriptions, and engagement data into account. It is trained and tested on 226 misleading and 665 non-misleading videos. However, they do not report on the performance of unimodal or non-neural network approaches – not ruling out that a multimodal neural network approach might be unnecessary. A comparison of different methods and modalities for the classification of fake news on TikTok is provided with *FakeSV* (Qi et al., 2023). They offer significant first evidence for the usefulness of multi-modal classification of Chinese (fake)-news TikTok videos.

A caveat for previous research is that they are trained and tested on datasets collected via hashtag, author, or event lists and/or are being hand-curated from the beginning. Those datasets only reflect a subset of the variety of videos users are possibly exposed to on TikTok. The classifiers trained on such data might not allow for a reliable classification of datasets that contain increased content variability, such as actual user trace data.

Collecting videos via a hashtag, keyword, or actor sample might tell us something about those topics and actors (and can serve to train a classifier). Still, it hardly tells us anything about the exposure to or impact of such content - what users consume and to what extent. Here, *data donations* present an excellent approach to gathering user-centric data that gives researchers access to watched videos. Two recent studies on TikTok base their findings on TikTok DDPs. They dive into analysis based on the raw DDPs and an accompanying survey, leaving questions of content exposure and multimodality unexplored (Goetzen et al., 2023; Zannettou et al., 2023). Hence, research has yet to use the full potential of TikTok DDPs to analyze exposure to multimodal content. However, the lack of understanding of content exposure and the multimodal nature of TikTok have posed two challenges for research on TikTok: 1) the facilitation of collecting TikTok DDPs and 2) the augmentation of said DDPs.

While we focus on the case of TikTok in this paper, the description also holds for digital platforms that are similarly multi-modal and have a vertical video feature, such as YouTube (Shorts) and Instagram (Reels). For those, an augmentation step is necessary for research incorporating the content level since the DDPs only contain metadata (Driel et al., 2022). For text-heavy platforms such as Facebook or Twitter, the DDPs already contain bigger parts of the content. However, these DDPs can, for example, be augmented with the full texts of articles users click on or post about. aDDPs are, hence, a generalizable approach that aims to increase the depth of available data for analysis, combining data not included in DDPs and corresponding survey data.

Challenge 1: The Data Donation Process & Available Frameworks

Digital trace data can be roughly differentiated into platform-centric and user-centric data. Platform-centric data is mainly gathered via APIs (often, this is publicly available data collected retrospectively without explicit user consent), while user-centric data is gathered either through tracking approaches on user devices (prospectively) or via data donations (Ohme et al., 2023). For TikTok, APIs or web scraping do not provide user-centric data. While they provide public data, private information such as the user's watch history and their behavior around each video is beyond their capabilities. Here, DDPs are the best option for collecting user-centric data to explore content exposure and the behavior of users.

DDPs provide an ecologically valid, non-reactive, reasonably scalable, and geographically independent data source – a combination of traits that no other user-centric data collection method provides (Driel et al., 2022; Ohme et al., 2023). DDPs represent the most complete available collection of user-centered digital trace data from TikTok available to date. Importantly, DDPs from TikTok give, at the time of our data collection in August 2023, the link to each video that was watched by a user - allowing for retrospective[1] data augmentation and making TikTok DDPs especially valuable for digital communication research (ibid.).

[1] Our current data collection has shown that the watch history contained in the DDPs only dates back half a year from the point of the data request. Other activities such as liking, commenting and private messages are present for the whole time of an account's existence.

To collect TikTok DDPs, a user must request the data as a JSON or TXT file and donate their data. The resulting *Data Download Package* (DDP) is a set of user-centric digital trace data (Ohme et al., 2023). The data donation, in general, can be facilitated in three different ways: First, researchers instruct the participants to install a desktop or mobile application that performs preprocessing steps locally and then sends the final DDP to the researchers' server (e.g, DataSkop, 2023). Second, researchers instruct the participants to upload the data directly to a server under their control, only to conduct data privatization and minimization afterward (e.g., Driel et al., 2022) or, third, use a web-based application that executes preprocessing steps on the participant's local machine, thereby only saving the final DDP to the researchers' database (e.g., Araujo et al., 2022; Boeschoten et al., 2023; Friemel & Pfiffner, 2023).

For the collection of TikTok DDPs, the third approach is ideal. It has the advantage of running the preprocessing locally, and current web applications are platform-independent – making the donation as easy and safe as possible for participants. Compared to the other two approaches, the threat of *compliance & consent error* (see Boeschoten et al., 2020) is mitigated as much as possible - *compliance* in the case of a dedicated desktop app that has to be installed and needs the user to transfer data between devices and *consent* in case of the direct data transfer – demanding the participant to donate not just the data required by researchers but also data such as address, name and personal messages. With *Port* (Boeschoten et al., 2023) and *DDM* (Pfiffner et al., 2022), at least two frameworks for a web app with the described advantages are in development and partially already published under open-access licenses to be used by researchers – the future of data donations is thereby set on web applications that allow for maximum privacy by minimal inconvenience for the donor.

For the current study, *Port* was employed, which allows for preprocessing on the participant's device, thereby mitigating privacy concerns for participants. Participants were recruited through a convenience sample, with a call for participation distributed via colleagues and student courses. Participants were initially led to an online survey that collected sociodemographic data and contained questions about their perception of the content they received on TikTok (further described in the section "Applying aDDPs in TikTok"). The survey also included detailed instructions on how to request their DDP from TikTok. During the survey, we generated a unique ID for each participant to link the survey data and the data donation. During the study, TikTok took up to three days to prepare the file (TikTok, 2023b). After three days, participants received an E-Mail with a personalized (via the ID) link to *Port*, where they found a manual on uploading their data donations. The ID is saved along with the data donations, allowing us to connect the survey data and the data donations later. 18 out of 42 (42.68%) recruited participants completed the process. Participants received an incentive of 20 € upon completion. The study received approval from the Ethical Review

Boards of the Weizenbaum Institute and the University of Amsterdam. An overview of the included information in the locally processed and donated DDPs can be found in Table 1.

While the data package that researchers retrieve is often only a subset of the DDP that the user has downloaded (depending on the preprocessing), we will continue to describe the donated data package as a (augmented) data download package since the subset that is augmented represents one to one the user trace data of the respective activities contained in the DDP (e.g., watch history).

Table 1 Description and collected variables for each activity beyond the timestamp. The timestamps always mark the beginning of the respective activity.

Activity	Description	Additional variables collected
Following	The user is following another user.	-
Favorites	The user is marking a video as a favorite.	Link to video
Logging in	The user is logging into their TikTok account.	Operating System
Searching	The user is searching TikTok with a search term.	-
Sharing	The user is sharing the present video in-app or externally.	-
Watching Videos	The user is watching a video.	Link to Video
Blocking	The user is adding another user to the block list.	-
Commenting	The user is commenting on a video.	-
Chatting	The user is writing a private message to another user.	-
Going Live	The user is starting a live stream.	-
Watching Livestreams	The user is watching a live stream	Link to Video
Posting Videos	The user is posting a video of their own.	Likes
Liking	The user is liking a video.	-

Challenge 2: Augmenting TikTok DDPs

TikTok DDPs provide a variety of insightful data points such as user activities (liking, sharing, watching), the users' app settings, and ad interests (Zannettou et al., 2023). Research has described different ways of linking DDPs with other data sources like survey data (e.g., Haim et al., 2023; Stier et al., 2020) and scraped metadata (e.g., video length, likes - Goetzen et al., 2023; Zannettou et al., 2023). Our suggested approach goes one step further. It proposes integrating audio-visual content features and their machine-readable multimodal feature embeddings (also multimodal representations) to the TikTok DDP, video metadata & survey data (see Figure 1). A resulting augmented data download package (aDDP) contains *survey data*, the *donated (subset) of* the *data download package*, *metadata* of a post (such as video length or number of likes), *content data* of a post (such as the video and audio file), and finally, *multimodal representations* of each post. These, ultimately, can serve as input for subsequent supervised and unsupervised machine learning tasks. Such aDDP combines the advantages of collecting initial user-centric data via DDPs with the richness of publicly accessible metadata and analyzable audio-visual content features. We do not stop with the augmentation via established computational methods in the social sciences (metadata scraping, natural language processing) but exploit the full depth of audio-visual content to facilitate state-of-the-art research. The concept of aDDPs provides a terminology that covers data linkage efforts and combines them with the advanced methodological opportunities of contemporary computational research.

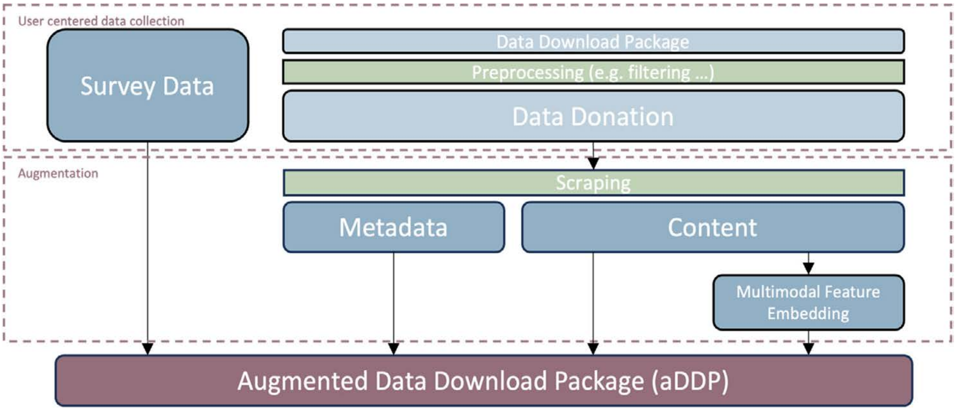

User centered data collection
Survey Data
Data Download Package
Preprocessing (e.g. filtering ...)
Data Donation
Augmentation
Scraping
Metadata
Content
Multimodal Feature Embedding
Augmented Data Download Package (aDDP)

Figure 1 Process of Data Donation Augmentation.

The augmentation with a survey during the initial data collection (e.g., Haim et al., 2023; Stier et al., 2020), as well as the initial collection of TikTok DDPs (e.g., Goetzen et al., 2023; Zannettou et al., 2023) itself, has been discussed previously; the other steps of the augmenting process demand a more detailed description and reflection to guide the present, and future research. Hence, in this paper, we focus on collecting metadata and content data and specifically explain the multimodal feature extraction for TikTok. This process, however, can be helpful in different projects in social science research that deal with multimodal content. To provide such guidance in an appropriate form, we will now go through the methodological decisions of the augmentation process. The substeps are exemplified with a pilot study of 18 data donors, showcasing the possibilities for empirical research based on aDDPs.

Collecting Meta and Content Data for TikTok

The TikTok Research API is not viable for our purpose because it only provides minimal metadata and no video or audio files – making other data sources necessary (Meßmer et al., 2023). At the same time, the terms of services forbid any other way of data augmentation in the case of using the API (TikTok, 2023). We thereby choose not to use the TikTok API.

Alternative public Python packages can facilitate the scraping instead, returning many more variables than the TikTok Research API, such as the *Pyktok* or *TikTok-Api* packages (Freelon, 2022/2023; Teather, 2019/2023). We found using the *TikTok-Api* package to be sufficiently reliable and convenient for data collection. To download the videos, we used a custom Python script. With the videos downloaded, the audio can be extracted with, e.g., the open-source Python package *moviepy* (Zulko, 2013/2023). For further information on the usage of the mentioned packages, please refer to their documentation.

In sum, the augmentation step of retrieving metadata and video data can currently not be sufficiently facilitated without programming and web scraping knowledge. As it comes with unofficial and custom scrapers, the scraping is volatile due to changes in website architecture. Custom scraping also poses a challenge to time management – because of its slowness and unreliability. Finally, scraping of content from the web poses legal questions. However, we deem our research in line with current EU legislation.[2]

An unsolvable circumstance of the current affordances for metadata and content data scraping is that we can not retrieve data for posts that are no longer available – be it for violations against the platforms' terms of service or the users' changed privacy settings. In our case, at the time of scraping, we could no

[2] The research is carried out by a non-profit research institute with the primary goal of scientific research. The scraping thereby falls under the exception granted by the DSM Directive for text and data mining. (Egger et al., 2022 p. 73-75)

longer retrieve data for 13.58% of the videos from the analyzed sample (1,821 out of 13,342 videos), which is similar to previous research on TikTok data donations (Zannettou et al., 2023).

Extract Feature Embeddings from TikTok Data

Depending on the research question and domain, the audio-visual content itself (e.g., manual content analysis) and meta-data can be used directly for analysis as they are. For machine learning tasks, there are two main options for representing the modalities: 1) using technical content characteristics such as cutting frequency or color spectrum (visual) and the loudness or dynamic complexity (audio) (e.g., Huddar et al., 2020; Ibañez et al., 2021; Lepa & Suphan, 2019; Syed et al., 2021) or 2) a vector representation retrieved from a pre-trained general-purpose model of the gathered modalities (e.g., Chiatti et al., 2019; Ram et al., 2020; Reeves et al., 2021). The latter approach (*transfer learning*) is at least equally good, often better for follow-up classification tasks compared to embeddings based on technical characteristics (Baltrusaitis et al., 2019; Zhang & Peng, 2022). This can be explained by them not being bound by the researchers' assumptions and knowledge of the possibilities surrounding each mode (Qi et al., 2023). Instead, the complexity of their training data binds the pre-trained models used to retrieve the embeddings. A typical training dataset for video representations is the *Kinetics 400* – a dataset that returns a vector of length 400 reflecting 400 human actions within the videos (Kay et al., 2017). The final layer and, even more, the last hidden layer – commonly larger and less impacted by the model's training classes - can be assumed to hold a sufficient number of latent characteristics of a video (or any other input modality) – superseding any hard-coded assumption made by the researchers.

The choice of how to retrieve the feature embedding is a core aspect of a multimodal classification task (Sleeman et al., 2021). Unlike in computer sciences, the models used to generate the embeddings should not merely be assessed based on their performance (Bender et al., 2021; Schwartz et al., 2020). When applied in the context of computational social science, the ease of implementation of a model becomes a significant factor. Since the performance difference between easily accessible pre-trained models and newer models that might be too recent to be accessible is usually in the lower one-digit percentages. Thus, the performance gain does not justify the added time spent on the implementation. Therefore, we suggest utilizing models that are easily importable in major machine learning libraries like *PyTorch* or *TensorFlow*. Both facilitate a hub of pre-trained models (*TensorFlow Hub*[3], *PyTorch Hub*[4]). Alternatively, platforms

[3] https://www.tensorflow.org/hub

[4] https://pytorch.org/hub/

such as *Hugging Face*[5]or *Kaggle*[6] are channels to source models that can be easily imported into common deep-learning frameworks in *Python*. More recent models are often only available as a set of scripts and files to be downloaded manually – which poses a significant inconvenience to researchers depending on their programming training. The following three subchapters will explain our embedding decisions.

Video

All state-of-the-art models are 3-dimensional convolutional neural networks, which differ in their performance only slightly across different classification tasks and training datasets (e.g., Huddar et al., 2020; Pandeya & Lee, 2021; Shang et al., 2021). Nevertheless, ideally, all current models should be tested if computationally feasible. For this research, we decided on *3D Resnet*[7]*,* a state-of-the-art model available via the *PyTroch Hub*. It is trained on the aforementioned *Kinetics 400* dataset and used in its pre-trained version without additional fine-tuning.

In line with the preprocessing requirements of *3D Resnet*, we sampled 32 frames from each video equally distributed over the video's length[8]. Depending on the application, other sample techniques can be helpful. Scene detection algorithms can identify sufficiently distinct parts of a video or maybe only the first 2 seconds of a video are of interest because the user has only watched those (Qi et al., 2023; Tian et al., 2019).

The second preprocessing requirement of *3D Resent* is that the single frames need to have dimensions of 256*256 pixels. Therefore, we squished the frames to the desired format – compared to cropping, this preserves more visual information from the original frame – even in reduced granularity (see Figure 2). Cropping would need previous knowledge of the area within the videos to focus on – which we do not have in the case of TikTok posts.

For each video, the preprocessed 32 frames are then fed into the *3D Resent* model, and the last hidden layer (length = 2304) is retrieved as the feature representation for the respective TikTok video. The resulting feature vector has two dimensions (2304x32) representing an embedding for each of the video's input frames. To retrieve an embedding for the whole video, the 2D vector is reduced to a 1D vector through element-wise aggregation, such as averaging (Selva et al., 2023).

[5] https://huggingface.co/

[6] https://www.kaggle.com/

[7] https://pytorch.org/hub/facebookresearch_pytorchvideo_resnet/

[8] If a video is 16 seconds long and has 30 frames per second we sample every 15th frame.

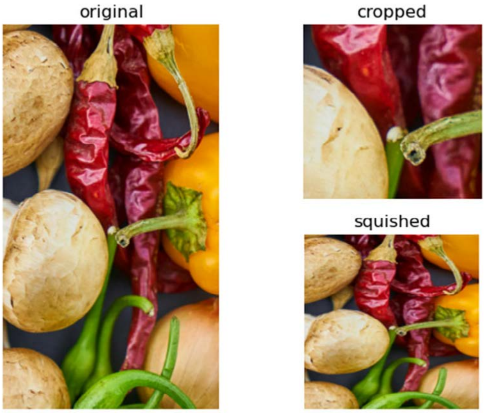


Figure 2 Imapact of cropping versus squishing on one example frame. We can see that the squished frame retains, unlike the cropped frame, information on the green & orange pepper. Original photo modification of Flat-lay Photography of Variety of Vegetables [E. Akyurt]. (204), under a Creative Commons [0] license.

Audio

TikTok videos' audios are heterogeneous – voice, music, and action-related acoustic signals are all possible. To acknowledge this variety, *VGGish* is used. *VGGish* is developed by *Google LLC* and trained on the *AudioSet* database. *AudioSet* is based on 2.1 million *YouTube* videos trained on 527 classes, from music over speech to lawn mowing (Hershey et al., 2017).

Like the video embedding, we extracted the feature representation based on the last hidden layer of the model (length = 4096). The embeddings returned reflect each second of the input audio and are aggregated to a 1D vector via element-wise average aggregation.

Text

The video descriptions are multi-lingual. Investigating a subset of videos[9], we find predominantly German (38.51%) & English (30.16%) descriptions. But also Korean, Arabic, Turkish, Russian & Cantonese content (together 15%). The language detection was conducted with *fasttext* (Joulin et al., 2016). A content classifier should be able to handle multi-lingual data, given that we cannot control the language of content in the DDPs. We use a state-of-the-art multi-lingual BERT model (Reimers & Gurevych, 2019). The model *distiluse-base-multilingual-cased-v1* is used since it supports 15 languages, including all mentioned above except

[9] The training dataset described later in this paper (N = 5,619).

Cantonese (1.5% of the descriptions). The output of the said model is not related to a classification taxonomy dictated via the training data but is supposed to serve as an input for further classification tasks. Therefore, we use only the final layer. *distiluse-base-multilingual-cased-v1*[10] returns a 1D vector of length 512.

After data collection and augmentation, each resulting aDDP (n = 18) consists of 1) the DDP, 2) corresponding survey data on sociodemographic characteristics and TikTok usage, 3) the raw content data (audio and video files that have been scraped), 4) metadata (length, likes, etc.), and 5) feature embeddings of the major modes a TikTok posts consists out of (visuals, audio & the textual description). All Python scripts used throughout the collection and augmentation process are made available open source (Wedel, 2024).

Applying aDDPs in TikTok Research

In the pilot study, we investigate the impact of user engagement behavior on the type of videos users encounter in their watch history. We use this exploratory question to showcase how aDDPs can be used in TikTok research and acknowledge that this is a proof-of-concept, not a study on its own. Results should, therefore, be interpreted accordingly. The user trace data under investigation are the 18 aDDPs, the collection and augmentation procedure of which has been described above.

As engagement behavior, we understand any action that signals a user paying attention to content. Here, we differentiate between passive (long watch time) and active (liking, sharing, etc.) engagement, along with the argumentation of first- and second-level exposure (Ohme & Mothes, 2020). The pilot study seeks to answer the following research questions concerning our 18 participants:

RQ1: Do users who show engagement behavior on informative videos receive more of such videos in future sessions/ within sessions?

RQ2: Does the users' self-reported consumption of informative videos align with actual digital trace data?

To facilitate research on the proposed questions, aDDPs are necessary because we need fine-grained user behavioral data (DDPs), survey data, and a database that allows us to classify each video with regard to whether it is informative or not (content data & multimodal feature representations).

However, DDPs do not let us know where the user has watched the videos on the platform. As of the time of data collection, TikTok holds two different feeds: the *for you feed* (algorithmically curated video suggestions) and the *following feed*

[10] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

(only videos published by creators a user follows). The *for you feed* is the default feed when opening the app and has been reported by TikTok as the dominant form of content consumption (TikTok, 2019). It is unclear to what extent suggestions from the following feed are also algorithmically suggested. Given that we can not distinguish between algorithmic and otherwise curated videos, we cannot certainly say that our results apply exclusively to the *for you feed*.

Methodology

To answer both research questions, we combine the different elements of the collected aDDPs. The DDPs were collected between the 18th of September 2023 and the 3rd of October from a German convenience sample, as described in the previous chapter on the TikTok DDP collection. The study sample comprised 18 individuals in Germany: 8 participants aged 16-26 and 10 aged 27-34. Most participants (15) held a university degree, while three did not. There were more females (9) than males (6), and three participants did not disclose their gender. The DDPs have been augmented as described in the respective previous section.

To facilitate content classification based on the multimodal feature embeddings, we train a classifier that categorizes the videos in the aDDPs into "informative" and "other" categories. The two categories are derived from the TikTok explore page classification. The TikTok explore page[11] is a website accessible via the TikTok desktop web interface. At the time of data collection, it consisted of 11 categories, where up to 200 videos were sorted within each category. The videos change constantly; to increase the dataset, we scraped the page repeatedly. The other ten categories are in contrast: *Dance & Music, Sports, Entertainment, Comedy & Drama, Cars, Fashion, Lifestyle, Pets & Nature, Relationships and Society.* The dataset is made available open source (Wedel, 2023). TikTok does not provide a description of these categories. A screening of the videos sorted under *Informative* shows mostly videos with tech, language, or finance tips and videos explaining scientific findings or history. We rely on the categorization being coherent enough to serve as a robust classification base for this proof-of-concept example. The chosen classification serves as an example of a prelabeled dataset that research needs to gather – either by manual labeling or using the limited number of videos labeled by TikTok.

The following sub-section guides through 1) the engagement measures based on the digital trace data, 2) the self-report-based engagement measures, and 3) the classifier training, including the subsequent classification of the videos in the aDDPs. We answer our RQs with binomial linear regression and the Pearson correlation coefficient.

[11] https://www.tiktok.com/explore

The aDDP-based engagement measures

Each aDDP is split into sessions using the time stamps included in the DDP. Each session represents a user's consecutive consumption of videos without a break. The information within TikTok DDPs does not allow us to decide on the sessions with absolute certainty. To detect the session breakpoints, we use a threshold of 105 seconds that Zannettou et al. (2023) derived from 347 TikTok DDPs. That means that when there is an activity duration of more than 105 seconds, we count that as a breakpoint between two sessions of consecutive content consumption.

We operationalize passive engagement with a user having watched a video longer than their median watch time of a video. The watch time has been derived following previous studies via the timestamps for each video, and the last video in each session was removed from the dataset after deriving the watch time of the preceding video (Goetzen et al., 2023; Zannettou et al., 2023). Active engagement behavior encompasses all active actions that can be taken by a user concerning a video: liking, sharing, commenting, and favoring. For the sake of simplicity, we aggregate those actions as active engagements but acknowledge that this step depends on the research question – a more granular analysis is possible should the research question desire this.

During the preprocessing of the DDPs on the participant's local devices, an unstable sorting algorithm was used, which does not allow the above-described analysis for sessions with duplicate timestamps. Regarding two activities with the same timestamp, we do not know which came first. Therefore, it is impossible to know which video has been watched for x seconds, which has been directly skipped, or to which video a follow-up engagement action relates. Therefore, we excluded all sessions with duplicate timestamps from the analysis. This renders 47.45% (n = 12,750) of the overall detected sessions with more than one activity unusable, leaving 14,117 sessions for analysis. The exclusion of those sessions does not allow for empirical findings beyond within-session effects. Since the present study is meant to be solely a proof-of-concept, we nevertheless exemplarily measure cross-session effects.

Self-reported information exposure measures

To measure the participants' self-perception of information consumption, we asked participants to assess on a 5-point Likert scale how much they agreed (1 strongly disagree to 5 strongly agree) with the following four statements: a) *TikTok is important for me to stay up to date with current affairs (politics, economics, etc.). (M = 2.44, SD = 1.34) ; b) TikTok is important for me to stay up to date with general affairs (celebrities, sports, etc.). (M = 3.167, SD = 1.38)* ; c) *TikTok is important for me to learn new things (DIY, cooking, etc.). (M = 3.61, SD = 1.09) ; and d) TikTok is showing me primarily informative content (M = 2.344, SD = 1.15).*

The statements are based on past research on news use of young German adults on social media and cover the broader news categories of hard news (current affairs) and soft news (general affairs) and summarize the remaining content[12] under learning and general information (Anter & Kümpel, 2023).

Training a classifier for aDDPs

To retrieve a pre-labeled dataset for model training, we scraped all videos on the TikTok explore page mentioned earlier from the 31st of July 2023 until the 4th of August 2023. While we retrieved around 200 unique videos per day – removing duplicates that occurred through videos being listed under one category for several days – due to the several dates of data collection, the initial training data set consisted of 473 videos labeled as informative and 4,664 videos tagged as a different category. An overview of the video overlap throughout the five days of scraping can be found in Appendix I.

To ease the unbalanced nature of the data, we decided to add the informative labeled videos from an earlier data collection (on the 4th, 12th, 13th, and 17th of July), resulting in 955 informative videos in total. Given the overall diversity of included categories, this training dataset of 5,619 unique videos can be assumed to represent a higher variation of videos compared to, e.g., keyword sampling methods that only include an often smaller number of videos from one specific domain while holding a meaningful number of instances of the target class. The metadata collection was facilitated via the *4CAT Toolkit (Peeters & Hagen, 2022)* and the *Zeeschumier (Peeters, 2023)* browser extension.

For classification, we tested a Support Vector Machine (SVM) as a traditional classifier for binary classification and a simple, fully connected Neural Network (NN) architecture with six hidden layers (see Appendix II). The target variable was the binary classification decision between *informative* and *other*. As model inputs, we tested uni- and multimodal representations based on the retrieved feature embeddings for three modalities of a video post (video, audio, text).

The critical design choice of a multimodal classifier is its fusion-mechanic (Sleeman et al., 2021). Fusion describes how the different modes are fused into one multimodal representation before (*early fusion*), during (*intermediate fusion*), or after (*late fusion*) the classification. For the case of TikTok, *early fusion* is sufficient since we can expect all modalities to be present (Choi & Lee, 2019). In early fusion, we concatenate the three calculated embeddings before we feed them into the tested classifiers to one embedding vector (e.g., multimodal representation of the respective video). Besides being easily implemented, early fusion also affords without effort the exploitation of cross-modality correlations (Zeppelzauer & Schopfhauser, 2016).

For the neural networks, each fully connected layer is followed by a dropout layer to avoid co-adaption within the network (Hinton et al., 2012). The hyperpa-

[12] Tips and inspirations; Service; Consumption and welfare; Trivia, Activism; Comedy and fun

rameters for all neural networks were set at 50 epochs, a batch size of 40, a learning rate of 0.0001, and a dropout chance of 0.2 after hyperparameter tuning.

For both types of classifiers, we oversampled the minority class (informative) during training to be represented equally often compared to the majority class (other). Min-max normalization has been applied to each single-mode embedding vector based on all respective embeddings from the test, train, and inference corpus. Training and validation have been facilitated via 5-fold cross-validation. We report the average results over all five folds for precision, recall, F1-Score, and accuracy. The results show that all tested supervised machine learning techniques perform better than uniform random guessing – validating that all models pick up decisive features within the data to outperform an uninformed classification (see Table 2).

The common characteristic of the best-performing models (SVM^{T+A+V}, NN^{T+A+V}, SVM^{T}, NN^{T+A}) is that they include the text mode. The best-performing model is the single-mode SVM^{T}, closely followed by the trimodal models SVM^{T+A+V} and NN^{T+A+V}. The predictions of the NN^{T+A+V} and the SVM^{T} have a high variation in performance across the folds compared to the SVM^{T+A+V} (see Figure 3).

The SVM^{T} model classifies, on average, across all five-folds, 86.9% of the actual informative videos correctly, and 85.9% of the informative classified videos are indeed informative. Other tested models afford a higher recall, but the trade-off in terms of reduced precision always results in an overall reduced F1 score (see Table 2).

The learned classification is based on a classification by TikTok, and we are likely to reproduce an algorithmic error (see Boeschoten et al., 2020) that is part of TikTok's classification. Hence, future research needs to conduct robust (manual) training and validation data labeling. Tested models should be validated on a labeled sample from aDDPs videos to assess a model's performance appropriately on the set of videos in the aDDPs. Based on the used test and training data, this work shows that unimodal SVMs might be sufficient depending on the classification scheme and underlying data. Nevertheless, the within NN comparison also indicates that for NN classifiers, multimodality improves the classification significantly – supporting the assumption that they can exploit correlations between the modes. Given the recall and precision measures of the SVM^{T} model, we can assume that it misses ~15% of informative videos and misclassifies ~15% of them as "informative".

Table 2 Performance comparison of the tested classifiers. Maximum in bold. The first data row represents no trained model but shows the performance of uniform random guessing as a baseline. Performance values are reported, with "informative" being the target class. The model name reflects the classifier type (SVM = Support Vector Machine; NN = Neural Network) and the included modalities (T = Text; A = Audio; V = Video).

Modalities used	Model name (Base Classifier$^{\text{Initials of the modalities}}$)	Recall	Precision	F1-Score	Accuracy
-	Uniform random guess	.5	.11	.18	.493
Text	SVM^{T}	.869	.859	.864	.953
	NN^{T}	.881	.635	.730	.887
Audio	SVM^{A}	.751	.380	.505	.749
	NN^{A}	.781	.738	.758	.915
Video	SVM^{V}	.775	.597	.674	.873
	NN^{V}	.778	.819	.797	.933
Text + Audio	SVM^{T+A}	.915	.412	.568	.763
	NN^{T+A}	.909	.695	.781	.909
Text + Video	SVM^{T+V}	.813	.720	.763	.914
	NN^{T+V}	.916	.793	.844	.939
Video + Audio	SVM^{V+A}	.839	.801	.819	.937
	NN^{V+A}	.817	.814	.815	.937
Text + Audio + Video	SVM^{T+A+V}	.910	.804	.853	.947
	NN^{T+A+V}	.854	.856	.852	.949

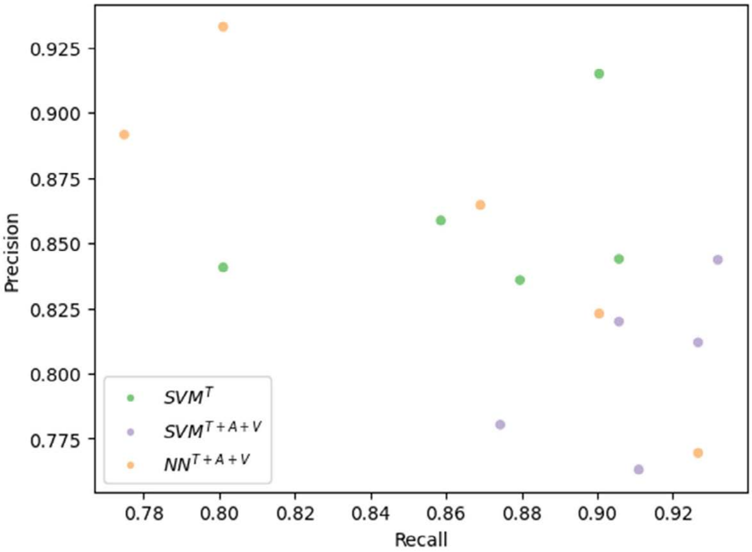


Figure 3 Precision and recall for each fold of the three best models by F1-score: the text only SVM (SVM^{T}), the trimodal support vector machine (SVM^{T+A+V}) and the trimodal neural network (NN^{T+A+V}).

Analysis & Results

Research question 1 asked if users who show engagement behavior on informative videos receive more of such videos in future sessions/ within sessions. For the present example, we first investigated the data on an aggregated level descriptively (see Table 3). We included the most recent 100 sessions, if available for each user, resulting in 11,475 videos over 1,242 sessions in our analysis. The SVM^{T} model labels 542 posts as “informative” and 10,933 as “other”. The informative labeled videos make up, on average, 5% of the videos watched by our participants. The average profits here from one outlier – user 17, with 12% of their videos being informative. Regarding engagement behavior, our participants clearly show less engagement behavior (active and passive) towards the informative content in their feeds than the engagement behavior towards other content (2% vs. 47% for passive and 0% vs. 4% for active). We conclude that passive engagement behavior for videos labeled as informative and active engagement behavior, in general, is sparse among our participants.

Table 3 Fraction of informative videos and engagement behavior aggregated per user.

		passive engagement		active engagement			
user	informative videos	info	other	info	other	#sessions	#videos
1	.03	.01	.18	.00	.02	100	1482
2	.02	.01	.66	.00	.04	31	140
3	.05	.01	.05	.00	.00	100	2002
4	.05	.03	.67	.00	.09	57	286
5	.07	.03	.43	.01	.07	97	674
6	.04	.02	.62	.00	.01	94	642
7	.06	.05	.75	.00	.01	16	93
8	.00	.00	.90	.00	.00	2	10
9	.08	.02	.70	.00	.01	23	120
10	.03	.02	.71	.01	.40	83	373

11	.02	.00	.09	.00	.00	100	1349
12	.03	.00	.20	.00	.00	96	1050
13	.05	.02	.54	.00	.01	97	650
14	.06	.02	.56	.00	.02	34	349
15	.05	.02	.40	.00	.05	98	733
16	.03	.00	.36	.00	.00	88	452
17	.12	.01	.13	.00	.01	58	911
18	.07	.03	.54	.00	.00	9	159
mean	.05	.02	.47	0	.04	65.72	637.50
total	-	-	-	-	-	461	3626

We then applied a binomial regression model with the respective engagement behaviors as an independent variable (IV) and the fraction of informative videos as the dependent variable (DV). We investigate two possible correlations: First, across sessions, the IV represents the fraction of engagement behavior on informative videos in sessions s_t, and the DV represents the fraction of informative videos in the following sessions s_{t+1}. Second, within sessions, the IV represents the fraction of engagement behavior on informative videos in the first half of session *s*, and the DV represents the fraction of informative videos in the second half of session *s*.

Given the sparsity of engagement behavior, we could not analyze all users for both engagement behaviors. For the users where the analysis could be conducted, we find no implications that the IV and DV correlate with three exceptions (user 4 and 6 passive and user 15 active) (see Table 4). This means that only in three cases is there an indication of a relationship between previous usage behaviors and the amount of future informative videos, suggesting that engagement behavior on a specific type of video will lead to users having more of such videos in their following sessions. RQ1, hence, cannot be answered affirmatively. Moreover, future research would need to apply time series analysis to investigate the causal direction of the relationship and, the time lag between engagements and possible effects on suggested videos.

Table 4 Binomial regression results.

		Across sessions		Within sessions	
		p	r-squared	p	r-squared
1	passive	.8	0.025765	.671	0.060967
	active	.837	-0.020926	.248	0.1648
2	passive	.481	-0.133834	-	-
	active	.636	0.089981	-	-
3	passive	.731	0.034955	.715	-0.050793

		Across sessions		Within sessions	
		p	r-squared	p	r-squared
4	passive	.041*	0.274038	.626	-0.255133
	active	.479	0.096461	.165	0.647343
5	passive	.258	0.116413	.843	-0.041792
	active	.096	0.170661	.799	0.053595
6	passive	.019*	0.241892	.683	0.094799
	active	.969	0.004048	.455	0.172234
7	passive	.537	-0.173152	-	-
	active	.8	0.071429	-	-
8	No sufficient engagement data on informative content.				
9	passive	.158	0.311532	-	-
	active	.598	0.119063	-	-
10	passive	.417	-0.090855	.527	0.32659
	active	.185	-0.147704	.792	-0.139792
11	passive	.198	0.130589	.803	0.038768
	active	.572	0.057455	-	-
12	passive	.225	-0.12557	.748	-0.050957
	active	.558	0.060883	.573	0.089542
13	passive	.698	-0.040057	.845	0.046675
	active	.942	-0.007459	.384	-0.205696
14	passive	.285	-0.191583	.128	-0.545731
	active	.246	-0.207755	.066	-0.634986
15	passive	.547	-0.061846	.361	0.210043
	active	.375	0.090977	.038*	0.455085
16	passive	.384	-0.094379	.427	-0.267155
17	passive	.522	0.086542	.886	-0.028408
	active	.425	-0.107593	.499	0.133362
18	passive	.133	0.579	.944	0.088475

Given the methodological nature of this paper, the analysis should not be taken as empirical evidence. The respective methodological pipeline is not grounded on a robust definition of *informative*. Nevertheless, with regards to the TikTok-defined term of informative videos for the majority of the participants, we do not find their engagement behavior impacting the fraction of informative content - neither within sessions - nor across sessions. The results for the cross-session comparison are unreliable, given the number of sessions that had to be excluded for the analysis because of duplicate timestamps.

Research question 2 asked for the relationship between self-reported content consumption and actual consumption of informative content on TikTok. Here, the full breadth of an augmented DDP can be used, as we rely on the survey data gathered from participants. Based on the self-reported data and the multimodal classification of the videos in a user watch list, we can test how closely users' self-perception comes to their digital behavior.

Self-reported information consumption was collected for *current affairs, general affairs*, *learning,* and *general information*. We again used the fraction of informative videos within each participant's 100 most recent sessions for the observed behavior. Analysis revealed a negligible correlation for *current affairs* (r = 0.268, p = 0.282), *learning* (r = 0.226, p = 0.366), and *general information* (r = -0.178, p = 0.478), a moderate correlation has been found for *general affairs* (r = 0.507, p = 0.032). Previous research (e.g., Araujo et al., 2017; Ohme et al., 2021; Parry et al., 2021) has shown that users' self-reports deviate from the observed digital behavior. Our pilot study suggests similar patterns for all dimensions of informative other than general affairs.

We note that we asked for qualitative assessments ("How much do you agree with …"), not for quantitative ("How often do you consume …") in terms of content consumption. The question items are less comparable with the cited studies – given that those explicitly asked for a quantification of content consumption.

Discussion

aDDPs present a promising future for digital trace data analysis. With open-source tools such as *Port* (Boeschoten et al., 2023), the means to collect such data is accessible to the broader research community. Using such tools also increases the transparency of the data collection. aDDPs are non-reactive and thereby come without the caveats of data collection methods that can compete otherwise (partially) with the collected data's granularity (e.g., tracking apps) or its modality (e.g., screenshot apps). The combination of granular information about user traces and the richness of publicly available video content data assessed through the initial DDPs make aDDPs an unprecedented database for critical social science research.

The paper presents a systematic approach to augmenting DDPs with multimodal data and using such data to answer substantial research questions. We do that specifically for TikTok, but this approach is flexible and adaptable to other data download packages. Augmenting DDPs of a multimodal nature presents a challenge to current research and has not been done before. This paper presents a unique approach with a clear pipeline on how to proceed with such an endeavor. It is a proof-of-concept on how content features of TikTok videos can be included in social science research, sampled via data donations.

Right now, aDDPs are especially helpful for vertical video platforms (VVPs) because researchers can collect the watch history retrospectively for half a year. The limit of half a year in the case of TikTok is a notable restriction, in line with the general unreliability and volatility of DDPs from different platforms (Carriére, 2023). It is not transparent whether the limitation comes from TikTok not saving the watch history for a user longer than half a year or if they only provide limited data.[13] Therefore, the *Digital Service Act* is a welcome prospect for improving the conditions for scientific work on user trace data – implementing an infrastructure that enforces transparency and scientific data access (Hase et al., 2023).

For TikTok DDPs, specifically session detection and the question of how a post was encountered (through the *for you feed* or else) are unsolved methodological questions. Here, it is similarly desirable that the platforms deliver even more detailed trace data. To detect session breakpoints, one could use the login timestamps. An initial attempt showed that they do not consistently mark the beginning of a session – users might stay logged in for a session break. While being reliable, login timestamps are not entirely sufficient.

Regarding the collection of DDPs, we must stay attentive to the difficulties and biases. Out of the 42 people who opened the survey invitation, only 18 donated their data. Future studies must carefully consider the reasons for the willingness to donate data for the platform of their interest (e.g., Pfiffner & Friemel, 2023). It remains a discussion within data donation studies to what degree classic representative samples are achievable. Nevertheless, for many research questions, answers coming from an in-depth analysis of online behavior coming from the digital traces of a specific subgroup may be a welcome complement to results from representative samples that are only able to rely on self-reports.

For 13.58 % of the analyzed subset of videos found in the data donations, we could not retrieve any metadata anymore and, thereby, for a similar fraction of videos as in previous studies on TikTok DDPs (Zannettou et al., 2023). Digital trace data from TikTok has the same limitations as trace data from other platforms. For the reproducibility of subsequent research, only the unique identifier of a video should be shared, not the content itself, to ensure the right to be forgotten on the video creator's side (General Data Protection Regulation, Regulation (EU) 2016/679, Art. 17; 2016). As conducted for other social media platforms, systematic research on the impact of no longer available content for TikTok is needed (e.g., Buehling, 2023; Zubiaga, 2018):

This paper is one of the first to compare uni- and multimodal classifications of TikTok videos, traditional machine learning, and deep learning approaches. Yet, we acknowledge that the classification is roughly cut, and more relevant content categories will need a robust definition on which basis a training and validation data set is manually labeled. Given the breadth of variation that multimodal representation with thousands of features proposes, we estimate that a minimum of 1000 videos for each class is desirable. However, further research is needed to explore the actual sample sizes.

The classification models have shown that an unimodal traditional machine-learning approach was sufficient. Looking only at the neural networks shows

[13] The suspicion originates especially from other activates such as following and liking being part of the DDP for the whole duration of the accounts existence.

that the trimodal neural network performs the best. Neural networks hold a high potential for improvement. Optimizations like a more sophisticated architecture (e.g., Shang et al., 2021; Tian et al., 2019) or better input data can lead to them superseding traditional machine learning classifiers for multimodal classification tasks. A juvenile indicator for that is that except for the text-only models, the neural network-based models performed generally better or were similarly suitable for all other test conditions.

We must also acknowledge that augmenting data introduces errors in the observed data. While self-reported user measures suffer from recall biases, augmented DDPs suffer from algorithmic errors that are an irreducible part of the pre-trained models employed to retrieve embeddings for each modality and missing data errors through DDPs only covering a fraction of an individual's media environment. We need to be aware that despite the great future of digital trace data, getting closer to a ground truth may be possible, but reaching it will remain a challenge.

While we showcase here the usefulness of an aDDP and the possibilities for substantial research, errors can be introduced in each part of the data collection, augmentation, and analysis. Future research should, therefore, apply the total error framework (Boeschoeten et al. 2022) when preparing an aDDP.

Augmentation needs resources, both from a human and a computational perspective. Doing this for a single research process is challenging, and we suggest working in greater collaborations, whereas 'seed DDPs' can be increasingly augmented - growing over time. Such consortiums could work together in larger data collection cycles to reach more significant and more complex datasets to answer multiple research questions (e.g., Ohme et al., 2023) or – assuming adequate privacy, ethical, and security measures – combine DDPs from different data collection cycles and automatically augment them. There remains a discussion as to under which conditions and how the data collected can be reused and shared – which would drastically increase the accessibility of the method to researchers unable to scrape or code. Examining this against EU, national, and institutional regulations would be the priority of such a consortium.

This study has shown that aDDPs open up new spheres of research. With such a procedure, researchers are not merely bound to the information the donations carry but can investigate a plethora of questions that rely on classifications that the platforms do not provide. aDDPs unite user-centric and content data collection. Embracing an aDDP allows research to expand questions on the distribution of anti-vax (Kim et al., 2023) or sexualized content (George & Surdeanu, 2023) with a user-centered perspective: *What do users actually see, and how do they react to it?* Vice-versa, do aDDPs allow studies that focus on user-centric data (e.g., survey, data donations) to cover more depth instead of relying purely on an existing data basis for the classifications of actors or domains or solely on the available metadata (Zannettou et al., 2023):

In a time when visual online platforms such as TikTok, YouTube Shorts, or Instagram have grown more prevalent – and with them an entirely new level of reliance on visual cues instead of textual description – it is as relevant as ever to explore the means to analyze such online content. Be it to explore the algorithmic curation of those new platforms, the harm they might do, or their impact on opinion formation. Consequently, this paper introduces a novel methodological framework to enhance the study of visual online platforms, enabling social science researchers to address previously inaccessible research questions.

Bibliography

Akyurt, E. (2024). *Flat-lay Photogtaphy of Variety of Vegtables*. Pexels. https://www.pexels.com/photo/flat-lay-photography-of-variety-of-vegetables-1435904/

Anter, L., & Kümpel, A. S. (2023). Young Adults' Information Needs, Use, and Understanding in the Context of Instagram: A Multi-Method Study. *Digital Journalism*, *0*(0), 1–19. https://doi.org/10.1080/21670811.2023.2211635

Araujo, T., Ausloos, J., Atteveldt, W. van, Loecherbach, F., Moeller, J., Ohme, J., Trilling, D., Velde, B. van de, Vreese, C. de, & Welbers, K. (2022). OSD2F: An Open-Source Data Donation Framework. *Computational Communication Research*, *4*(2), 372–387. https://doi.org/10.5117/CCR2022.2.001.ARAU

Araujo, T., Wonneberger, A., Neijens, P., & de Vreese, C. (2017). How Much Time Do You Spend Online? Understanding and Improving the Accuracy of Self-Reported Measures of Internet Use. *Communication Methods and Measures*, *11*(3), 173–190. https://doi.org/10.1080/19312458.2017.1317337

Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443. https://doi.org/10.1109/tpami.2018.2798607

Basch, C. H., Hillyer, Grace C., & Jaime, Chistie. (2020). *COVID-19 on TikTok: Harnessing an emerging social media platform to convey important public health messages*. https://doi.org/10.1515/ijamh-2020-0111

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Boeschoten, L., Ausloos, J., Moeller, J., Araujo, T., & Oberski, D. L. (2020). *A framework for digital trace data collection through data donation* (No. arXiv:2011.09851). arXiv. https://doi.org/10.48550/arXiv.2011.09851

Boeschoten, L., Schipper, N. C. de, Mendrik, A. M., Veen, E. van der, Struminskaya, B., Janssen, H., & Araujo, T. (2023). Port: A software tool for digital data donation. *Journal of Open Source Software*, *8*(90), 5596. https://doi.org/10.21105/joss.05596

Boeschoten, L., Voorvaart, R., Van Den Goorbergh, R., Kaandorp, C., & De Vos, M. (2021). Automatic de-identification of data download packages. *Data Science*, *4*(2), 101–120. https://doi.org/10.3233/DS-210035

Buehling, K. (2023). Message Deletion on Telegram: Affected Data Types and Implications for Computational Analysis. *Communication Methods and Measures*, *0*(0), 1–23. https://doi.org/10.1080/19312458.2023.2183188

Carriére, T. (2023, December 9). *Volatility of Data Download Packages*. Data Donation Symposium, Zurich. https://datadonation.uzh.ch/en/symposium-2023/

Chiatti, A., Davaasuren, D., Ram, N., Mitra, P., Reeves, B., & Robinson, T. (2019). *Guess What's on my Screen? Clustering Smartphone Screenshots with Active Learning*. http://arxiv.org/pdf/1901.02701v2

Choi, J.-H., & Lee, J.-S. (2019). EmbraceNet: A robust deep learning architecture for multimodal classification. *Information Fusion*, *51*, 259–270. https://doi.org/10.1016/j.inffus.2019.02.010

DataSkop. (2023). *Wie tickt TikTok?* DataSkop. https://dataskop.net

Driel, I. I. van, Giachanou, A., Pouwels, J. L., Boeschoten, L., Beyens, I., & Valkenburg, P. M. (2022). Promises and Pitfalls of Social Media Data Donations. *Communication Methods and Measures*. https://doi.org/10.1080/19312458.2022.2109608

DSA: Very Large Online Platforms and Search Engines. (2023). [Text]. European Commission - European Commission. https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413

Egger, R., Kroner, M., & Stöckl, A. (2022). Web Scraping. In R. Egger (Ed.), *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications* (pp. 67–82). Springer International Publishing. https://doi.org/10.1007/978-3-030-88389-8_5

Fiallos, A., Fiallos, C., & Figueroa, S. (2021). Tiktok and Education: Discovering Knowledge through Learning Videos. *2021 Eighth International Conference on eDemocracy & eGovernment (ICEDEG)*, 172–176. https://doi.org/10.1109/ICEDEG52154.2021.9530988

Freelon, D. (2023). *Dfreelon/pyktok* [Python]. https://github.com/dfreelon/pyktok (Original work published 2022)

Friemel, T. N., & Pfiffner, N. (2023). *The Data Donation Module*. https://datadonation.uzh.ch/en/infrastructure/

General Data Protection Regulation, Regulation (EU) 2016/679, Art. 17, § Uncategorized (2016). https://gdpr.eu/article-17-right-to-be-forgotten/

George, E., & Surdeanu, M. (2023). *It is not Sexually Suggestive, It is Educative. Separating Sex Education from Suggestive Content on TikTok Videos* (No. arXiv:2307.03274). arXiv. https://doi.org/10.48550/arXiv.2307.03274

Goetzen, A., Wang, R., Redmiles, E. M., Zannettou, S., & Ayalon, O. (2023). *Likes and Fragments: Examining Perceptions of Time Spent on TikTok* (No. arXiv:2303.02041). arXiv. https://doi.org/10.48550/arXiv.2303.02041

Haim, M., Leiner, D., & Hase, V. (2023). Integrating Data Donations into Online Surveys. *Software Review*.

Hase, V. (2023, November 9). *Fulfilling data access obligations: Platforms need to increase their compliance to enable data donation studies*. Data Donation Symposium, Zurich. https://datadonation.uzh.ch/en/symposium-2023/

Hase, V., Boczek, K., & Scharkow, M. (2022). Adapting to Affordances and Audiences? A Cross-Platform, Multi-Modal Analysis of the Platformization of News on Facebook, Instagram, TikTok, and Twitter. *Digital Journalism*. https://doi.org/10.1088/1751-8113/42/3/035201

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. (2017). *CNN Architectures for Large-Scale Audio Classification* (No. arXiv:1609.09430). arXiv. https://doi.org/10.48550/arXiv.1609.09430

Hiippala, T. (2017). The Multimodality of Digital Longform Journalism. *Digital Journalism, 5*(4), 420–442. https://doi.org/10.1080/21670811.2016.1169197

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). *Improving neural networks by preventing co-adaptation of feature detectors* (No. arXiv:1207.0580). arXiv. http://arxiv.org/abs/1207.0580

Huddar, M. G., Sannakki, S. S., & Rajpurohit, V. S. (2020). Multi-level feature optimization and multimodal contextual fusion for sentiment analysis and emotion classification. *Computational Intelligence, 36*(2), 861–881. https://doi.org/10.1111/coin.12274

Ibañez, M., Sapinit, R., Reyes, L. A., Hussien, M., Imperial, J. M., & Rodriguez, R. (2021). Audio-Based Hate Speech Classification from Online Short-Form Videos. *2021 International Conference on Asian Language Processing (IALP)*, 72–77. https://doi.org/10.1109/IALP54817.2021.9675250

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). *Bag of Tricks for Efficient Text Classification* (No. arXiv:1607.01759). arXiv. https://doi.org/10.48550/arXiv.1607.01759

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., & Zisserman, A. (2017). *The Kinetics Human Action Video Dataset* (No. arXiv:1705.06950). arXiv. https://doi.org/10.48550/arXiv.1705.06950

Kim, S. J., Villanueva, I. I., & Chen, K. (2023). Going Beyond Affective Polarization: How Emotions and Identities are Used in Anti-Vaccination TikTok Videos. *Political Communication, 0*(0), 1–20. https://doi.org/10.1080/10584609.2023.2243852

Lepa, S., & Suphan, A. (2019). *Der Elefant im Wohnzimmer der Kommunikationswissenschaft: Die rechnergestützte Analyse nonverbaler digitaler Kommunikation*. https://doi.org/10.25598/JKM/2019-10.6

Li, L., & Kang, K. (2023). *Exploring the Relationships between Cultural Content and Viewers' Watching Interest: A Study of Tiktok Videos Produced by Chinese Ethnic Minority Groups*. 37–46. https://www.scitepress.org/Link.aspx?doi=10.5220/0010610900370046

Meßmer, A.-K., Degeling, M., & Jaursch. (2023). *Response to the European Commission's call for evidence on a planned Delegated Regulation on data access provided for in the Digital Services Act (DSA)*. Stiftung Neue Verantowortung.

Ming, S., Han, J., Li, M., Liu, Y., Xie, K., & Lei, B. (2023). TikTok and adolescent vision health: Content and information quality assessment of the top short videos related to myopia. *Frontiers in Public Health, 10*. https://www.frontiersin.org/articles/10.3389/fpubh.2022.1068582

Mordecai, C. (2023). #anxiety: A multimodal discourse analysis of narrations of anxiety on TikTok. *Computers and Composition, 67,* 102763. https://doi.org/10.1016/j.compcom.2023.102763

Newman, N., Fletcher, R., Eddy, K., Robertson, C. T., & Nielsen, R. K. (2023). *Reuters Institute Digital News Report 2023*.

Ng, R., & Indran, N. (2023). Videos about older adults on TikTok. *PLOS ONE, 18*(8), e0285987. https://doi.org/10.1371/journal.pone.0285987

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B., & Robinson, T. N. (2023). *Digital Trace Data Collection for Social Media Effects Research: APIs, Data Donation, and (Screen) Tracking*. https://www.tandfonline.com/doi/full/10.1080/19312458.2023.2181319

Ohme, J., Araujo, T., Vreese, C. H., & Piotrowski, J. T. (2021). Mobile data donations: Assessing self-report accuracy and sample biases with the iOS Screen Time function. *Mobile Media & Communication, 9*(2), 293–313. https://doi.org/10.1177/2050157920959106

Ohme, J., & Mothes, C. (2020). What Affects First- and Second-Level Selective Exposure to Journalistic News? A Social Media Online Experiment. *Journalism Studies, 21*(9), 1220–1242. https://doi.org/10.1080/1461670X.2020.1735490

Pandeya, Y. R., & Lee, J. (2021). Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications, 80*(2), 2887–2905. https://doi.org/10.1007/s11042-020-08836-3

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, 5*(11), Article 11. https://doi.org/10.1038/s41562-021-01117-5

Peeters, S. (2023). *Zeeschuimer* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.8399900

Peeters, S., & Hagen, S. (2022). The 4CAT Capture and Analysis Toolkit: A Modular Tool for Transparent and Traceable Social Media Research. *Computational Communication Research, 4*(2), 571–589. https://doi.org/10.5117/CCR2022.2.007.HAGE

Pfiffner, N., & Friemel, Thomas. N. (2023). Leveraging Data Donations for Communication Research: Exploring Drivers Behind the Willingness to Donate. *Communication Methods and Measures, 17*(3), 227–249. https://doi.org/10.1080/19312458.2023.2176474

Pfiffner, N., Witlox, P., & Friemel, T. N. (2022). *Data Donation Module (Version 1.0.0)* [Computer software]. https://github.com/uzh/ddm

Primig, F., Szabó, H. D., & Lacasa, P. (2023). Remixing war: An analysis of the reimagination of the Russian–Ukraine war on TikTok. *Frontiers in Political Science, 5*. https://www.frontiersin.org/articles/10.3389/fpos.2023.1085149

Qi, P., Bu, Y., Cao, J., Ji, W., Shui, R., Xiao, J., Wang, D., & Chua, T.-S. (2023). FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(12), Article 12. https://doi.org/10.1609/aaai.v37i12.26689

Ram, N., Yang, X., Cho, M.-J., Brinberg, M., Muirhead, F., Reeves, B., & Robinson, T. N. (2020). Screenomics: A New Approach for Observing and Studying Individuals' Digital Lives. *Journal of Adolescent Research, 35*(1), 16–50. https://doi.org/10.1177/0743558419883362

Reeves, B., Ram, N., Robinson, T. N., Cummings, J. J., Giles, C. L., Pan, J., Chiatti, A., Cho, M. J., Roehrick, K., Yang, X., Gagneja, A., Brinberg, M., Muise, D., Lu, Y., Luo, M., Fitzgerald, A., & Yeykelis, L. (2021). Screenomics: A Framework to Capture and Analyze Personal Life Experiences and the Ways that Technology Shapes Them. *Human-Computer Interaction, 36*(2), 150–201. https://doi.org/10.1080/07370024.2019.1578652

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (No. arXiv:1908.10084). arXiv. http://arxiv.org/abs/1908.10084

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM, 63*(12), 54–63. https://doi.org/10.1145/3381831

Selva, J., Johansen, A. S., Escalera, S., Nasrollahi, K., Moeslund, T. B., & Clapés, A. (2023). Video Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2023.3243465

Shang, L., Kou, Z., Zhang, Y., & Wang, D. (2021). A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok. *2021 IEEE International Conference on Big Data (Big Data)*, 899–908. https://doi.org/10.1109/BigData52589.2021.9671928

Sleeman, W. C., Kapoor, R., & Ghosh, P. (2021). Multimodal Classification: Current Landscape, Taxonomy and Future Directions. *arXiv: Learning*.

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field. *Social Science Computer Review, 38*(5), 503–516. https://doi.org/10.1177/0894439319843669

Syed, M. S. S., Pirogova, E., & Lech, M. (2021). Prediction of Public Trust in Politicians Using a Multimodal Fusion Approach. *Electronics, 10*(11), Article 11. https://doi.org/10.3390/electronics10111259

Teather, D. (2023). *TikTokAPI* (Version 6.1.1) [Python]. https://github.com/davidteather/tiktok-api (Original work published 2019)

Tian, H., Tao, Y., Pouyanfar, S., Chen, S.-C., & Shyu, M.-L. (2019). Multimodal deep representation learning for video classification. *World Wide Web, 22*. https://doi.org/10.1007/s11280-018-0548-3

TikTok. (2019, August 16). *How TikTok recommends videos #ForYou*. Newsroom | TikTok. https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you

TiKTok. (2023). *TikTok Research API Terms of Service*. https://www.tiktok.com/legal/page/global/terms-of-service-research-api/en

TikTok. (2023a). *TikTok's DSA Transparency Report 2023*.

TikTok. (2023b, January 30). *Requesting your data | TikTok Help Center*. https://support.tiktok.com/en/account-and-privacy/personalized-ads-and-data/requesting-your-data

Valkenburg, P. M. (2022). Theoretical Foundations of Social Media Uses and Effects. In *Handbook of Adolescent Digital Media Use and Mental Health* (pp. 39–60). Cambridge University Press. https://doi.org/10.1017/9781108976237.004

Wedel, L. (2023). *A categorized multimodal TikTok dataset*. https://www.weizenbaum-library.de/handle/id/420

Wedel, L. (2024). *Augmented TikTok Data Donation Packages Repository* (Version 0.1) [Computer software]. https://github.com/lionwedel/augmented_tiktok_DDP/blob/main/README.md

Yeung, A., Ng, E., & Abi-Jaoude, E. (2022). TikTok and Attention-Deficit/Hyperactivity Disorder: A Cross-Sectional Study of Social Media Content Quality. *The Canadian Journal of Psychiatry, 67*(12), 899–906. https://doi.org/10.1177/07067437221082854

Zannettou, S., Nemeth, O.-N., Ayalon, O., Goetzen, A., Gummadi, K. P., Redmiles, E. M., & Roesner, F. (2023). *Leveraging Rights of Data Subjects for Social Media Analysis: Studying TikTok via Data Donations* (No. arXiv:2301.04945). arXiv. https://doi.org/10.48550/arXiv.2301.04945

Zeppelzauer, M., & Schopfhauser, D. (2016). Multimodal classification of events in social media. *Image and Vision Computing, 53*, 45–56. https://doi.org/10.1016/j.imavis.2015.12.004

Zhang, H., & Peng, Y. (2022). Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research. *Sociological Methods & Research*, 004912412210826. https://doi.org/10.1177/00491241221082603

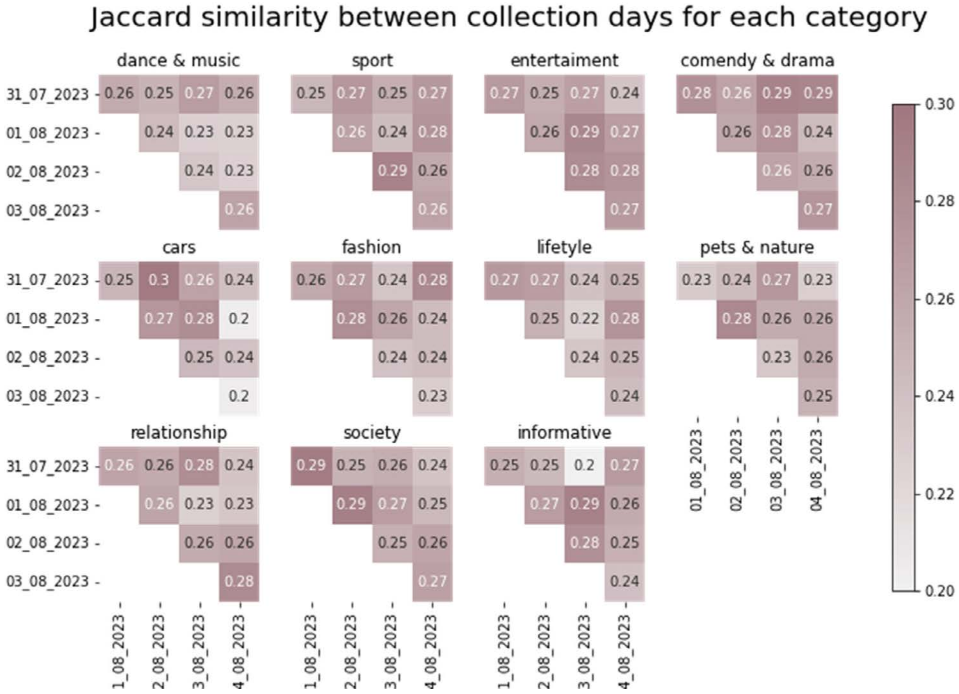
Zhou Ting. (2021). The Media Images of Old Influencers on TikTok: A Multimodal Critical Discourse Analysis. *Journal of Literature and Art Studies, 11*(10). https://doi.org/10.17265/2159-5836/2021.10.013

Zubiaga, A. (2018). A longitudinal assessment of the persistence of twitter datasets. *Journal of the Association for Information Science and Technology, 69*(8), 974–984. https://doi.org/10.1002/asi.24026

Zulko. (2023). *MoviePy* [Python]. https://github.com/Zulko/moviepy (Original work published 2013)

Appendix

I – Overlap measured by Jaccard similarity in unique videos between the five consecutive days of data collection for each category.



II – Neural Network Architecture

We used a Neural Network with six fully connected layers, with ReLu activation functions and five dropout layers for all input combinations. Below, we report the architecture as constructed in *PyTorch*. The layer size varies depending on the size of the input vector (number of input modalities). These in- and out-feature sizes adapted accordingly and always aimed to give the network a funnel shape.

```
six_layer(
  (classifier): Sequential(
    (0): Linear(in_features=6912, out_features=4096, bias=True)
    (1): ReLU(inplace=True)
    (2): Dropout(p=0.2, inplace=False)
    (3): Linear(in_features=4096, out_features=2048, bias=True)
    (4): ReLU(inplace=True)
    (5): Dropout(p=0.2, inplace=False)
    (6): Linear(in_features=2048, out_features=1024, bias=True)
    (7): ReLU(inplace=True)
    (8): Dropout(p=0.2, inplace=False)
    (9): Linear(in_features=1024, out_features=512, bias=True)
    (10): ReLU(inplace=True)
    (11): Dropout(p=0.2, inplace=False)
    (12): Linear(in_features=512, out_features=256, bias=True)
    (13): ReLU(inplace=True)
    (14): Dropout(p=0.2, inplace=False)
    (15): Linear(in_features=256, out_features=1, bias=True)
  )
  (sigmoid): Sigmoid()
)
```

Reflective Appendix

We encountered four main problems during the research process, which we reflect on in the following section. The first issue is that we could not freely decide which platform to look at. When conducting this study, TikTok was the only platform for which the data for augmented data download packages was available. Since we are especially interested in social media platforms for which a user's watch history is available and the users' reactions to the watched videos (watch time, like, share, etc.), we can link viewing behavior with other user-centric behavior and reactions. At the time of data collection, TikTok data download packages (DDPs) were the only DDPs that provided this additional fine-grained behavioral data. Facebook and Instagram lack sufficient watch history (watch time cannot be derived, or only the past seven days are included), and YouTube has a watch history but lacks user reactions. The data donation process would not pose increased hurdles compared to TikTok – all platforms allow users to download their data within reasonable steps.

Additionally, the meta-data access for other platforms has been recently uncertain, while TikTok seemingly shines with providing their official Research API. However, it can be expected that upcoming EU legislation (Digital Services Act) will enforce similar data cases for all platforms. Nevertheless, our proposed methods and our learnings can also be adapted to the analysis of audio-visual content from other platforms (e.g., YouTube (Shorts) and Instagram (Reels)). But especially the adaption to longer video content (e.g., YouTube) holds another set of challenges: the amount of data that would need to be processed would demand a more robust infrastructure, and a simple equal frame sampling of 30 frames per video might be insufficient to grasp the visual essence of each video – here scene detection would come into play as a method to detect the relevant scenes in the video which than would allow for a frame sampling stratified by the distinct scenes.

The second problem was the small sample size of retrieved data donations. While data donations are a promising way to retrieve fine-grained user trace data – this data collection method is prone to small sample sizes like ours (*N* = 18). We aimed at a larger sample. However, recruiting via convenience sampling (distributing the onboarding survey through university courses) proved very slow. Forty-five participants signed up for the study over one month, out of which only 18 went through with the donation (41.8%). To gather larger samples, a more large-scale recruiting method (e.g., via a panel provider) could improve sample sizes.

It is worth considering the differences and challenges users might encounter when requesting and downloading their data download packages – which at least partially can decrease the conversion rate. In the case of TikTok, an issue that participants have reported was that once you download the data, TikTok switches from the app to a browser window and requests the users to verify. While on our test devices, the verification ran flawlessly due to being connected to a Google Account on the respective Android Device, participants with a different verification method set in place had a less seamless experience. Sometimes, the page did not load due to an assumed issue on TikTok's side, or the requested verification type was a surprise, and participants did not understand what they were supposed to fill in. A handful of participants reached out to us over those. However, given that not every participant who successfully requested the data download package (DDP) donated their data, we must assume that this issue led to a decrease in successful data donations. Researchers should explore the donation process in different environments (OS x type of verification) to prepare instructions on this issue in case participants encounter it. Nevertheless, the process of requesting, downloading, and donating a DDP is extensive and remains a hurdle.

Additionally, people might not feel comfortable sharing such sensitive data in the first place – despite it being anonymized. Therefore, even larger data donation samples will likely be biased (self-selection, not privacy-aware, pro-research). Hence, working with data donations will often mean focusing on the behavior of particular groups that might be sufficiently represented by the sample (e.g., young males from urban areas) or concentrating on phenomena that can be assumed to be sufficiently independent of the sample biases (e.g., algorithmic curation of TikTok). Future data collection efforts have yet to show whether a representative sample is possible.

The third issue that became clear throughout the research process was that our pre-labelled dataset does not hold much value for empirical research. We utilized a pre-labelled dataset provided by TikTok, categorizing videos as "informative" or "other." While this dataset served as a sufficient foundation for our methodological proof-of-concept, it is essential to acknowledge its limitations for empirical research. Ultimately, the choice for the pre-labelled data set and the formulation of the survey questions were not aligned appropriately because, for this paper, the effort to label our data just for a methodological proof-of-concept was not justifiable. The TikTok-defined concept of "informative" lacks a clear operational definition, hindering our ability to interpret the results within established academic frameworks. Despite this limitation, the dataset proved suitable for our exploratory study. However, we manually checked the videos in the "informative" category during the classification procedure to confirm their distinctiveness from other categories. This was not planned – but it was doable with less than 500 videos to check. The primary requirement for our supervised machine learning pipeline was that the subsets of categories are sufficiently different from each other to inform the classification process. The precise meaning of the labels was secondary to this goal. However, it is crucial to note that while practical for machine learning, these labels do not sufficiently align with academic conceptualizations of "informativeness." Future research may need to consider time and budget for labelling. An alternative would be to adapt research questions to TikTok's labels (e.g., explore page or diversification labels). Here, research needs a content analysis to understand the label's meaning.

The final issue we encountered was our computing infrastructure. We initially underestimated the restrictions in scraping, downloading, and processing time our proposed pipeline demands. We suggest to set up the data collection and processing pipeline on a server from the get-go. While collecting and processing vertical videos in the four digits locally is still possible, it can be highly restrictive depending on your local hardware and internet connection. While setting up a server infrastructure takes time initially, it allows for scalability later. Setting up such infrastructure would have allowed us to analyse all videos within the data download packages, not just the most recent 100 sessions per user. Since we had to rely on our locally set-up computing infrastructure, we had to make this cut to keep the scraping and analysis within a reasonable timeframe.

methods, data, analyses | Vol. 19(2), 2025, pp. 46-86 DOI: 10.12758/mda.2024.07

Preferences, Participation, and Evaluation of Answering Questions About the Books Participants Have at Home Through Conventional and Image-Based Formats

Patricia A. Iglesias

Research and Expertise Centre for Survey Methodology (RECSM), Universitat Pompeu Fabra

Abstract

The collection of photos through online surveys has emerged as a valuable research tool given the growing use of smartphones, which have facilitated the capture and share of photos. However, gaps persist in understanding respondents' involvement in these tasks when asked to perform them in an online survey. Existing literature lacks insights into participants' preferences, their assessment of questions asking for photos, and how their characteristics might impact their participation in such queries. This paper addresses these gaps, while also comparing how image-based formats compare to conventional ones. Conducted among 1,270 parents living with children in primary school of an opt-in panel in Spain, the mobile online survey implemented in this study revealed a preference for conventional questions, and higher participation in that format than in the image-based one. Respondents able to choose their response format and preferring images presented higher participation rates than those without a choice. While both formats were perceived as equally easy, participants using conventional formats liked the questions better than those answering through photos. Finally, age, being female, having a tertiary education degree, and using the camera at least once a week positively impacted the participation in image-based questions, whereas comfort with new technologies increased the likelihood of liking this format. This study not only fills critical gaps in the literature but also sheds light on the complexities of asking for photos in online surveys.

Keywords: image collection, mobile online surveys, books at home, participation, respondent's evaluation, preferred response format

Although surveys are one of the most used methods to collect data, they suffer from measurement errors (Saris & Gallhofer, 2014). The increasing use of smartphones provides new measurement opportunities that could help reducing such errors (Revilla, 2022). In particular, smartphones have sensors, like the GPS, that allows tracking respondents' location and offer in-the-moment surveys (Ochoa, 2022) or the microphone, that can be used to record voice answers (Höhne & Gavras, 2022). Photos captured with the smartphone camera have also gained attention as a possible new data source, since they are expected to alleviate respondents' burden, enhance data accuracy and quality, and provide insights beyond what conventional response formats can offer (Revilla, 2022).

Research on the feasibility of requesting for photos in online surveys has focused mainly on the respondents' willingness to capture and share photos. Further, studies assessing participation in such questions have asked mostly for general photos (e.g., of the respondents' surroundings, see Bosch et al., 2019), but there is little evidence for more demanding tasks (e.g., submitting multiple photos or capturing items in various locations). Furthermore, scant attention has been devoted to investigating respondents' preferences, or to exploring their evaluation of such response formats. This lack of knowledge does not allow practitioners to make informed decisions regarding the use of visual data within the frame of online surveys: even though images could increase data quality, their relevance diminishes if respondents do not provide such photos.

Additionally, certain participants might be more inclined to participate in image-based response formats than others, potentially introducing biases into who submits photos and who does not. Thus, in this paper I provide new evidence about the respondents' preferences, levels of participation, and evaluation of questions asking to share visual data focusing on a more demanding task: capturing photos of all the books respondents have at home. Since inquiries about the number of books at home have been a recurring feature in numerous surveys within this field, this case study can both enhance our understanding of the efficacy of collecting visual data and facilitate substantive analyses in the realm of social sciences.

Traditionally, the number of books at home has been used to measure cultural capital and/or socioeconomic status (see Heppt et al., 2022; Sieben & Lechner, 2019). Further, analyses on this question show that the number of books impacts dimensions such as parent health literacy (Sanders et al., 2004), socio-emotional skills (Brunello et al., 2012), reading scores (Güre et al., 2023; McNally et al., 2023), and students' academic language comprehension (Heppt et al., 2022).

Direct correspondence to

Patricia A. Iglesias, Research and Expertise Centre for Survey Methodology (RECSM),
Department of Political and Social Sciences, Universitat Pompeu Fabra, C/ de Ramon
Trias Fargas, 25-27, 08005, Barcelona, Spain
E-mail: patricia.iglesias@upf.edu

However, this question exhibits limitations. First, respondents might answer the survey from a place different than home, preventing them from looking at the books they have, leading to inaccurate answers. Second, even if respondents are at home during the survey, counting each book can be arduous, especially if they have many. Thus, this question would mostly rely on estimates, potentially compromising accuracy. Third, the previous issues seem to have been addressed by presenting response intervals corresponding to the number of books. However, the large intervals affect the granularity. For instance, Gil Flores (2011) used the categories 0-10, 11-25, 26-100, 101-200, and more than 200 books. These categories do not provide detailed information on the exact number of books; having 26 books is very different from having 100 books. Fourth, using intervals could increase social desirability bias (having more books is well-regarded), since respondents might avoid selecting categories that fall in the lower range, which can cause an overestimation of the number of books at home.

One might argue that both the lack of granularity and possible social desirability bias could be mitigated by proposing an open-ended question (wherein respondents should type in the exact number of books). This approach might enhance the level of detail in the information and prevent respondents from inferring what is considered a low or high number of books. However, little research has asked for the number of books in this manner, and comparisons of the quality of both estimates have not been conducted to the best of my knowledge, let alone with recent data.

Moreover, count variables also have limitations. For instance, we can expect rounding errors, estimation errors due to satisficing and/or to low mathematical/spatial abilities, and a tendency to over-report the number of books. These problems could be reduced by measuring the number of books in a different way: through photos of the books sent by respondents. Photos of the books could transcend the wide intervals commonly found in questionnaires by offering a count directly observable in the images. Furthermore, photos can yield novel insights, such as the title of some of the books, or the methods employed for storage. Surveys including the question on the number of books usually do not explore such aspects. Knowledge on this could enrich the understanding of cultural and/or economic capital, since researchers might achieve better characterizations of their subjects. For instance, 40 books of English literature might relate differently to cultural capital than 40 cooking books. The language in which books are written could also indicate that participants are familiar with languages different than their native tongue, possibly also expanding their cultural capital. Further, storing books inside boxes or closets might affect how individuals in the dwelling engage with those books, by making them harder to reach. Thus, besides the number of books at home, I also ask for the languages of the books and their storage methods.

This paper investigates survey participants' preferences for, participation in, and evaluation of answering questions in a mobile online survey about the books they have at their main residence using two response formats: a conventional one (radio buttons and textboxes) and an image-based one (taking and sharing photos of the books). In addition, it examines the impact of respondent characteristics on their participation in and evaluation of each format. Thus, the core aims are two-fold: first, to shed light on the feasibility of using conventional versus image-based formats for collecting information; and second, to identify which individuals participate and positively evaluate each format when a more complex task is involved: providing information about the books they have access to at home. The analysis is conducted using data from the Netquest opt-in online panel in Spain collected in June 2023 among parents that had at least one child in primary school. While the quality of the data provided is a further consideration relevant to the evaluation of the image-based format, it falls beyond the scope of this paper.[1]

Background

The research presented in this section regards visual data produced during the survey, in line with the type of data asked in this paper. While there are also studies about sharing visual data already captured (i.e., produced before the survey), they are not explored in this section.

Respondents' Preferences, Participation, and Evaluation

Previous research has studied the feasibility of collecting visual data through online surveys by investigating two main aspects: the respondents' stated willingness to share visual data, and their actual sharing.

Regarding the willingness of respondents to share photos or videos, different results have been reached depending on the type of visual data file and the topic covered. For instance, 56% of participants in an online survey performed in the Netquest panel in Spain would accept to answer questions by taking and sharing photos of products with their smartphone (Revilla et al., 2019), whereas 65% of participants in wave 9 of the Understanding Society Innovation Panel in the UK would use their smartphone camera to take photos or scan barcodes for a survey (Wenz et al., 2019). Thus, the willingness is moderately high (over 50%) when it comes to photos of objects within the household or receipts. However, willing-

[1] Due to the complexity of evaluating data quality, especially in the case of images (this aspect requires a lot of additional analyses and explanations), it will be discussed in a separate paper.

ness seems to be lower for inquiries that might be considered as more intrusive: 18% of respondents in a survey using the online probability-based LISS panel in the Netherlands would be willing to share a selfie (Struminskaya, Toepoel, et al., 2021), while 14% of participants in a survey from Statistics Netherland stated willingness for the same task (Struminskaya, Lugtig, et al., 2021). Regarding a photo of their house, 38% in the first survey would be willing and only 12% in the second survey. Finally, 24% and 16%, respectively, would send a video of their surroundings.

Iglesias and Revilla (2024) asked for the willingness to capture and share a photo with the smartphone among participants of the Netquest opt-in online panel in Spain. 62% of respondents expressed willingness, while 31% stated it depended on the photo they are asked. This hints that the willingness for capturing and sharing photos is high: if put together, almost 93% of respondents would be willing to send photos. This study isolated the potential effect of skills and availability of producing visual data during the survey by asking respondents to not consider those aspects when answering. Moreover, the authors found that 99% of participants stated knowing how to capture photos with the camera in their mobile device, and that 83% would be able to take a photo of something in their house. By analyzing these three dimensions together, the authors estimated the expected participation, and found that 54% of respondents answering from home would participate in an online survey question asking them to capture and share a photo of something at their dwelling.

Additionally, a second set of studies have asked respondents to answer questions by actually sharing images captured during the survey. Some of these studies asked for screenshots: in particular, Ohme et al. (2021) and Sewall et al. (2022) requested screenshots of the iOS Screen Time function. The participation rate was 12% for the first study, and 78% in the last wave of the second study. They used different samples: Ohme et al. (2021) recruited a sample with an opinion research company in the Netherlands, while Sewall et al. (2022) used participants from the Prolific online panel in the United States with a task-approval of 95% or higher.

A study among Millennials from the Netquest panel in Spain and Mexico asked for a photo of the surroundings. 49% of respondents in Spain and 57% in Mexico sent such photos, whereas 24% and 17% skipped the question, respectively (Bosch et al., 2019). This proves to be higher than the willingness stated by survey respondents in the Netherlands regarding the surroundings, which might be explained by them being asked for a video rather than photos. Certainly, capturing a photo allows more control than a video (a photo can be quickly recaptured, while a video could be more burdensome).

Further, Bosch et al. (2022) used the Respondi panel in Germany to ask for photos captured during the survey (with a smartphone) related to given topics, and their equivalent in conventional format. The authors found that breakoff was higher for image requests in almost all cases, and that, among those not breaking off, participation was significantly greater in the group answering in conventional ways (over 99%) than in the groups doing so through image-based response formats (49% to 67%). When asked for their evaluation, respondents reported liking it better and finding the conventional response format easier than the image-based one.

Ilic et al. (2022) carried out an experiment with participants of the LISS panel, comparing conventional and image-based response formats. In addition to assigning participants to one of these two formats, a third group was introduced, in which respondents could choose their preferred response format for providing information. The authors asked conventional questions or photos of the respondents' favorite place in their home, an outdoor space of their dwelling (e.g., balcony), and their heating system. 57% of participants who had the option to choose their response format preferred answering through photos. Compliance rates[2] varied between the response formats, from 58% to 99% for conventional response formats and from 27% to 78% for image-based ones. Among participants with a choice, those who opted for image-based responses demonstrated higher compliance rates (from 50% to 78%) than those who were solely instructed to send photos (27% to 39%). Differences between the two groups answering through conventional formats were less pronounced (with a maximum difference of 9 percentage points). Non-complying respondents answering questions about their favorite place and heating system with images were asked for their reasons to not capture and submit photos. Half of participants stated privacy concerns regarding the question for their favorite place, and 10% reported technical problems. As for the heating system, 49% stated it was unreachable and 15% mentioned privacy concerns.

The results from these studies show that both willingness and participation in survey questions asking for photos not only varies among, but also within studies when asked for different types of images. This could be associated with the content of the photo that is being asked (e.g., more or less personal/sensitive, more or less difficult/burdensome to capture it), but also with the exact way in which it was asked, different levels of incentives, differences regarding the type of panel or the target populations (e.g., different countries, age cohorts, etc.), as well as temporal changes. These differences in the findings make it necessary to continue exploring other topics and types of tasks. More research is needed to understand the extent to which previous results can be generalized to different situations.

Further, participation in image-based response formats is only contrasted with their equivalent in conventional response formats in few studies, leading to limited evidence about the performance of both formats in similar settings.

[2] Measured as providing an answer for the conventional response format, and submitting a photo for the image-based response format.

Finally, respondents' preferences and evaluations have been even less studied so far: to the best of my knowledge, they have been reported by only one study each.

Impact of the Respondents' Characteristics

Previous research has studied the effects of the respondents' sociodemographic characteristics, experience as panelist, and their use of mobile devices on their willingness to share images in the frame of surveys, as well as their expected and actual participation when proposed image-based response formats. Summary tables of the variables impacting these dimensions are available in the supplementary material 1.

As for willingness, Struminskaya, Toepoel, et al. (2021) found no effects for gender and education, but the frequency of taking photos, trust in the anonymity of the answers, and being older than 65 years old positively impacted the likelihood of being willing to share sensor-collected data (including, among others, capturing a photo of the house, of self, and a video of the surroundings). Conversely, Iglesias and Revilla (2024) found a negative relation between age and willingness, but similarly to the previous study, they found no effects for gender and education. Finally, higher participation in previous surveys positively impacted the willingness to participate in questions asking for photos.

Concerning the expected participation, Iglesias and Revilla (2024) found that it is less likely for older respondents to take and share photos of something inside their dwelling. In contrast, the higher the number of surveys completed in the three months prior to the study, the more respondents are expected to share such photos. Gender and education did not impact the expected participation in this study.

Finally, Struminskaya, Lugtig et al. (2021), focusing on actual participation,[3] found that the frequency of use of the camera did not have an impact, and neither did the level of education or participation in previous surveys. Moreover, age impacted positively sharing a photo of the house or of the respondents, and females were more likely to share photos of receipts.

Overall, there is consensus on the lack of impact of education, while either no or positive effects are found for gender, frequency of taking photos, and participation in previous surveys. The only variable with opposed effects is age, which impact varies from positive to negative in different studies. This could be related to the samples varying in their concentration in different ages, as well as their different locations: the two studies in the Netherlands found positive effects, while the one in Spain found a negative impact. Further, the studies in the Neth-

[3] In this study, the authors focused on willingness and participation. All of those stating willingness participated in sharing photos.

erlands used probability-based panels, while the study in Spain used an opt-in panel. Indeed, more research is needed to understand such differences.

Research Questions and Hypotheses

This paper reports the results of an experiment conducted among online opt-in panel respondents of a mobile survey gathering data on the books they have at their main residence through conventional or image-based response formats. The image-based format involved requesting photos of the books, while the conventional format asked to answer questions related to the books by typing numbers or clicking a radio button. I compare three groups: 1) Conventional format: respondents are asked to answer 11 questions related to the books they have at home by typing in numbers of clicking radio buttons; 2) Image-based format: respondents are requested to send photos of the books at their home; 3) Choice: respondents can decide either to answer the 11 conventional questions or to share photos.

While most of previous research focused on relatively straightforward tasks, involving a single answer or photo, this paper explores the implications of employing conventional versus image-based response formats when interested in more complex tasks. On the one hand, the conventional task is much more demanding than what has been tried in previous studies comparing both response formats. Indeed, it requires participants to answer 11 cognitively demanding questions, that require estimating numbers (e.g., number of books of different categories) and percentages (e.g., proportion of books in Spanish). On the other hand, the image-based response format is also more demanding than in previous research: instead of requesting a single photo (e.g., of the heating system as in Ilic et al. 2022, or of the surroundings as in Bosch et al. 2019), participants were asked to provide photos of all the books in their residence. Thus, respondents might need several photos to capture all books. Further, respondents might need to move through different spaces/rooms within their household, since books might be dispersed (e.g., children have books in their rooms, or books that are being currently read are on night tables). If other people are using some of the rooms (e.g., children sleeping in their room), it might also not be possible to take the photos immediately. Finally, respondents were instructed to remove items such as decorative elements, to ensure clear visibility of the books. This can represent quite some work for participants to remove and put back such items, and can generate high burden if they have books in several places and have to do it several times. Consequently, this study presented respondents with a challenging task, expected to be more time-consuming and effort-intensive than previous examples studied in image-based data collection. Given these complexities, this research could provide novel insights into aspects previously explored such as the respondents' preferences, participation, and evaluation, but in a slightly different context.

This study first delves into the respondents' preferences concerning response formats when given the choice between conventional and image-based response ones, especially for addressing demanding tasks. Thus, the first research question is:

RQ1: *Do respondents prefer to provide the information about the books in their dwelling through images or answering questions in conventional ways?*

The only study on respondents' preferences between conventional and image-based response formats (Ilic et al., 2022) found that participants mostly choose images. However, due to the specificity of the task studied in this paper, I expect respondents to be more reluctant to capture and share photos since the books could be in many different places, respondents might need to tidy each area before capturing the photos, may not be at home when answering the survey (thus unable to capture the photos), or could have privacy concerns (e.g., perceive the task as more intrusive than in the case of photos of the heating system). Thus, my first hypothesis posits that, when respondents are in a position of choosing a method, the preference for the conventional response format will surpass that of the image-based one (*H1*).

Second, this study assesses the levels of participation, i.e., participants actually answering the 11 conventional questions (for the conventional format) or sending at least one photo displaying their books (for the image-based format). Further, it investigates whether being able to choose one response format affects participation. Thus, the second research question is:

RQ2: *Does the participation vary between a) image-based versus conventional answer formats, and b) respondents choosing their preferred format versus respondents being only proposed one format?*

Based on previous literature, lower participation is expected from respondents (with or without a choice) for image-based inquiries compared to conventional answer formats (*H2a*). Moreover, participants with a choice are expected to participate more than those without a choice in the case of the image-based response format (*H2b*), while participation levels in conventional formats is expected to be similar across groups (*H2c*), as in Ilic et al. (2022).

Respondents participating in the image-based response format might still dislike it or find it difficult, which might potentially affect the participation in future surveys. Thus, the next research question:

RQ3: *How does the evaluation of respondents about the book-related questions vary between a) image-based versus conventional response formats, and b) respondents choosing their preferred format versus being only proposed one?*

Considering the cognitive effort required to provide accurate answers in the conventional format, which involves tasks like estimating numbers and proportions, in contrast to the familiarity of capturing photos with smartphones or tablets, a task for which most respondents in the same panel declared having the skills (see Iglesias & Revilla, 2024), I anticipate that respondents utilizing the image-based response format will perceive the task as easier than those employing the conventional one (*H3a*). However, respondents may like the conventional format more than the image-based one (*H3b*), given the respondents' familiarity with conventional questions, and that capturing photos, although not difficult, might present practical challenges in the case of the books at home, potentially leading to a more time-consuming and tiresome experience. Moreover, I expect that participants choosing their response format present better evaluations than those unable to choose, regarding both the perception of easiness (*H3c*) and the extent to which they like the tasks (*H3d*).

Finally, certain respondents' characteristics could influence their participation and evaluation of image-based response formats.[4] Thus, my last research question is:

RQ4: *How do the respondents' sociodemographic characteristics, experience as panelist, comfort with new technologies, trust in the confidentially of the answers, and use of mobile devices influence their participation and evaluation in image-based versus conventional response formats?*

To the best of my knowledge, no research has studied the influence of these factors on the evaluation of an image-based format. Therefore, I do not formulate hypotheses in this case but follow an exploratory approach. In contrast, I propose the hypotheses bellow regarding the impact of the different factors on participation, since there is some research on this aspect.

Since the survey implemented in this study did not target old population[5] (most respondents were under 50 years old), age is not expected to significantly impact participation in image-based response formats (*H4a*). Similarly, no effect is expected for gender (*H4b*) and education (*H4c*), in line with previous literature. Conversely, familiarity with the camera included in the mobile device (*H4d*) and sharing photos (*H4e*) are expected to have a positive impact in participation in the image-based format. Although there are mixed findings in the previous literature regarding this aspect, I anticipate that individuals accustomed to using cameras in smartphones and sharing photos will be more inclined to participate: since those using smartphones more often are more familiar with them,

[4] This study aimed to investigate how these factors influenced respondents' preferences for one format over the other. However, such analysis could not be conducted due to the low number of respondents opting for the image-based response format (n=12).

[5] According to data from the Economically Active Population Survey of the Statistics Office of Spain, 99% of children attending primary school have parents of maximum 54 years old. See the section "Data collection" for more details.

and those frequently sharing photos might have fewer privacy concerns, these two aspects could increase participation. Similarly, although not studied in the previous literature, a higher level of comfort with new technologies is expected to boost participation in image-based response formats (*H4f*). Moreover, trust in the confidentiality of the answers is expected to impact positively the sharing of photos (*H4g*). Further, it is expected that households with more children will present lower participation (*H4h*), as it might translate into having more books and eventually in more places, making the task of capturing photos more tiresome. Finally, previous experience as a panelist is expected to negatively impact the participation in image-based requests, as respondents might be more accustomed to conventional formats than innovative ones (*H4i*). Table 8, available in the conclusions, summarizes the hypotheses.

Addressing these research questions, this paper contributes to the existing literature by presenting results on the request for images within the context of a mobile survey, exploring a relevant topic in social sciences and focusing on more complex questions and tasks than what has been studied previously. Further, this study focuses on a specific demographic group, namely parents living with children who attend the first, third, or fifth year of primary school in Spain. This introduces practical challenges, such as limited response time due to parental duties, difficulties in capturing photos amid childcare responsibilities, and potentially less organized living spaces, especially concerning children's books. Moreover, this is the first study collecting images with the ***WebdataVisual*** tool (Revilla et al., 2022), which was developed with the goal of having a more user friendly tool. Finally, the relation with smartphones moves forward swiftly, and technology is more accessible each day to smartphone users. Thus, this study complements the previous literature by contributing a contemporary perspective, recognizing the changing landscape in smartphone usage.

Data and Methodology

To address the research questions, an experimental design was implemented. This experiment is part of a bigger study. In this section, only the relevant elements for this paper will be presented. For a depth review on the overall study design, readers can consult the full study protocol (Iglesias et al., 2023).

Experimental Design and Groups

The experiment aimed to collect information about the books present in respondents' main residences using conventional and image-based response formats.[6]

For the conventional response format, 11 questions grouped in the following three dimensions were asked:

- *Number of books*: four open-ended questions about 1) the total number of books at home, and the number of books 2) for toddlers and children who do not know how to read, 3) for literate children and teenagers, and 4) aimed at a general audience.
- *Language*: three open-ended questions asking for the percentage of books 1) in Spanish, 2) in one of the three co-official languages in Spain (Catalan, Galician, and Euskera), and 3) in other languages.
- *Storage*: four radio-button questions asking whether books are stored 1) on shelves, 2) inside closets or drawers, 3) on center, coffee, or night tables or over a desk, and 4) in other places.

For the image-based response format, respondents were only asked to provide photos of their books, under the assumption that the aforementioned information could be extracted through image classification, i.e., the process of extracting and labeling the information contained in an image (Bandyopadhyay, 2021). Both conventional and image-based questions regarding the 11 items will be referred to as "test questions" in this paper.

Three experimental groups are considered: *Text*,[7] *Images*, and *Choice*. For the sake of simplicity, the names of the two first groups reflect their respective assigned answering format. In the third group (*Choice*), participants could select between the conventional or image-based formats. Throughout this paper, respondents choosing the conventional format are referred to as members of *TextChoice*, while those preferring images are named *ImagesChoice*. Respondents stating no preference were assigned to the image-based format (thus, are considered members of *ImagesChoice*). Table 1 presents a summary of the groups and response formats compared in this paper.

[6] This collection will help answering substantive questions regarding children's academic performance in relation to the number of books. Since previous literature has found no impact of e-Books in children's academic performance (Heppt et al., 2022), information on them was not collected.

[7] The design of the full experiment considers two different methods within the conventional response formats: *Text* and *TextPlus*. The only difference between both methods is that in the latter an illustration was provided to respondents to help them estimate the number of books. This is used in a different paper to study whether such illustration can help improve the quality of the answers in conventional formats. Since this does not affect the response format, for the analytical purposes of this paper, respondents in *TextPlus* as well as those in *Text* are all included in the *Text* group.

Table 1 Groups and Response Formats

Group	Response format for the test questions
Text *TextChoice*	11 conventional questions.
Images *ImagesChoice*	Capturing and sending photos of the books at home.

Questionnaire

The questionnaire consisted of up to 65 questions, extending beyond the test inquiries and covering topics such as the sociodemographics of respondents, the characterization of (one of) their child in primary school, activities related to literature engagement, usage of camera-related functions on their mobile devices, comfort with new technologies, and self-assessment of their spatial, mathematical, and verbal abilities. Further, respondents were asked to evaluate their experience when answering conventional or image-based response formats, and to provide additional information such as whether they had technical problems while uploading their photos. For more details, the full questionnaire (in Spanish and in English) is available in the supplementary material 2.

Since photos of the books at home could have been potentially asked to any respondents, a message at the beginning of the questionnaire requested them to answer from home. However, this could not be verified, as respondents did not share geolocation data. Thus, respondents could continue with the survey even if they were not at home.

Moreover, respondents had to complete the survey on smartphones or tablets. This restriction was imposed because taking photos of all books with computers (even laptops) was deemed too inconvenient. Further, the ***WebdataVisual*** tool used to collect the photos only allows capturing them during the survey when using mobile devices. Respondents entering the survey via computers were asked to switch to a smartphone or tablet and were unable to continue if they did not do so.

Sample and Data Collection

The target population consisted of parents of children enrolled in the first, third, or fifth year of primary education in Spain at the moment of the survey. These specific years were selected because changes in Spain's evaluation system (shifting from quantitative to qualitative evaluation) were implemented in those courses at the moment of the survey. Thus, including the second, fourth, and sixth years in the same survey might have impacted the substantive objectives to be fulfilled with the collected data, as the questionnaire asked for grades in Spanish and mathematics (for details, see the study protocol by Iglesias et al., 2023). Quotas for age, gender and educational level of respondents were used to get a sample similar on these variables to adults with children between 6 to 12 years (the average ages of children attending primary school in Spain). These estimates were derived from the Economically Active Population Survey of the Statistics Office of Spain.[8]

Data were collected in June 2023 through the Netquest opt-in panel in Spain (www.netquest.com), which invites panelists to participate in surveys, and rewards them with points determined by the questionnaire's length (for more information about the kind of surveys and rewards in this panel, see Revilla, 2017).

Out of 4,854 individuals invited to participate, 2,443 started the survey. 899 were filtered out due to security checks or survey requirements not being met (e.g., not providing consent to participate or not having a child in the first, third, or fifth year of primary education), while 72 individuals were excluded because demographic quotas had already been fulfilled. 202 entered to the survey but broke off before the first test question (i.e., first question about the books at home), leading to 1,270 individuals arriving to the test questions: 53% were female, the mean age was 42 years and 92% of participants were 30 to 50 years old. 45% possessed a higher education degree. Of all respondents arriving to the test questions, 636 were in the *Text* group, 305 in the *Choice* group (261 in *TextChoice* and 44 in *ImagesChoice*), and 329 in the *Images* group.

The allocation in a given group was performed right before the first test question, with respondents being assigned to the group with the least individuals at that moment. The group *Text* is two times larger than the others since it contains two groups (see footnote 7). Checks for balance were conducted (see supplementary material 3) on age, gender, and level of education, revealing no differences between the composition of the groups *ImagesChoice* and *Images*, and between *TextChoice* and *Text*. When comparing participants answering through either conventional or image-based format, differences are found for gender, with a significantly higher proportion of women in the image-based than in the conventional group (50% versus 59%). However, the difference between groups

[8] The public dataset (available at https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176918&menu=resultados&idp=1254735976595#tabs-1254736030639) displays the age of the members of the dwellings in ranges of 5 years. For the estimation aimed to calculate the quotas, dwellings with children between 5-9 years old during the first trimester of 2022 were considered, since those children will be 6-10 years old in the same trimester of 2023. Thus, our quotas are a proxy for dwellings with children between 6 and 12 years old, since it is not possible to know that exact range of ages from the publicly available data. Margins of +/-3 percentage points were used for the quotas, since the target population of this study is not exactly the same as the one used in the Economically Active Population Survey.

being of 9 percentage points, I do not expect it to influence the overall results presented in this paper.

On average, respondents reaching the test questions have been members of the Netquest panel for 6.9 years, and completed 13 surveys in the three months previous to this study. 99% of respondents used a smartphone to answer the survey (1% used a tablet). For those finishing the survey, the median completion time was 9.3 minutes (9.5 for the image-based response format groups, and 9.2 for those answering through conventional questions).

Analyses

R 4.2.3 was used to perform the analyses. The script is available in the project's repository (https://osf.io/7y3sq/).

Addressing ***RQ1*** (preferences for one response-format over the other), respondents in the group *Choice* were asked twice for their preferred response format and were offered three options: conventional, image-based, or no preference. The first question regarding preference was presented before the test questions so participants could answer those questions by using the response format they chose (those without a preference were assigned to answering with photos). The second preference question was presented after they answered the test questions with the chosen format, to assess whether they would still prefer it. The proportions of participants selecting each option within those who saw the questions are reported. Comparisons are made between the three categories to test *H1*.

Further, respondents not choosing images were asked for their reasons through a multiple-choice question with the following options: camera in the mobile device not working, privacy concerns, expected lack of skills, having too many books, and others (with the option of explaining further in a textbox). The main reasons are presented.

To study participation (***RQ2***), different dimensions were considered. Indeed, when facing a given question, participants have three main options: provide an answer (participation), skip the question but continue with the survey (item nonresponse) or abandon the survey (breakoff). In this study, the interest is in comparing a set of 11 questions with a request for photos. Thus, there are different ways to operationalize breakoff, item nonresponse and participation, in each response format. Consequently, I computed and report several indicators, which were estimated separately for respondents answering conventional and image-based formats.

As for the conventional one, five indicators are presented:

- *Breakoff*: Percentage of respondents, within those who saw the first question about books, that left the survey on one of the screens where the 11 test questions were displayed.

- *Minimum participation*: Percentage of respondents providing a substantive answer to at least one question.
- *Partial participation:* Percentage of respondents providing a substantive answer to at least six questions.
- *Full participation:* Percentage of respondents providing a substantive answer to all 11 questions.
- *Average number of substantive answers* out of the 11 possible ones.

In the last four indicators the calculations are computed out of all respondents seeing the 11 questions and continuing with the survey (i.e., not breaking off). Further, the option "I don't know" was presented to participants answering through conventional formats. Even though this might be a valid response, particularly when participants genuinely lacked the information, I excluded "I don't know" when studying participation, because this can be selected as a way to avoid any cognitive effort, and because there is no equivalent for the image-based response format. Conclusions reached in this paper do not change when considering "I don't know" as participation (see supplementary material 4).

For the three questions concerning the language(s) of the books (the percentages of books in Spanish, in one of the co-official languages in Spain, and in other languages), if the answered questions added up to 100, the three items were considered as complete. For example, if a respondent had all their books in Spanish, they might have typed "100" for Spanish and left the others blank. This is considered as a participation without item nonresponse. In any other case where there was a blank response without adding up to 100, the empty questions were considered as nonresponse.

As for the image-based format, three indicators were used:

- *Breakoff:* Percentage of respondents leaving the survey on the screen where the image request was presented over the number of respondents who saw this screen.
- *Minimum participation:* Percentage of respondents capturing and sharing at least one image. Since it is not possible to know if one photo captured all the books in the dwelling, sending at least one image was considered as "minimum participation". The denominator was the number of respondents required for images who did not breakoff in the test question.
- *Average number of photos* among participants sharing at least one image.

Comparisons were made at both the response format (conventional versus image-based) and group (*Text* versus *TextChoice, Images* versus *ImagesChoice*) levels. Regarding response format, the percentages of respondents breaking off were compared to test *H2a*. The rest of indicators used in both formats cannot be directly compared since these measures gauged different aspects. For instance, answering one test question in the conventional response format (minimum participation) is not equivalent to sending one photo.

At the group level, I compared the *Images* and *ImagesChoice* groups based on their percentages of breakoff, respondents sending at least one photo, and the average number of photos to test *H2b*. Similarly, the *Text* and *TextChoice* groups were compared regarding their percentages of breakoff, respondents providing at least one, six and 11 substantive answers, and the average number of such answers to test *H2c*.

Further, the reasons for not uploading images, asked to participants skipping the question requesting images, are reported. The same categories presented to those not choosing images in the preference question were offered, with an additional category for technical issues.

Regarding ***RQ3***, two aspects of the respondents' evaluation were considered: the extent to which they found the test questions easy/difficult, and how much they liked/disliked them. These aspects were originally measured through a scale from 0 ("Extremely difficult"/ "Totally disliked") to 4 ("Extremely easy"/ "Totally liked"), and were recategorized into "Difficult"/ "Dislike", "Not easy nor difficult"/ "Not like nor dislike", and "Easy"/ "Like". The proportions of respondents in each of these categories over those presented with these questions are compared among response formats and groups to test *H3a* to *H3d*.

Further, an open-narrative question among those disliking any of the two response formats was presented. The answers to these questions were coded and the frequency of the codes was estimated (n=20 for images, n=15 for conventional questions). Only codes mentioned more than once are presented. With such small groups conclusions cannot be reached, but the reasons still help understanding why respondents did not like the respective response formats.

Comparisons between categories of a variable within the same group (***RQ1***) and comparisons among groups and formats (***RQ2*** and ***RQ3***) were performed with Chi-squared tests, with significance at the 5% level.

Finally, regarding ***RQ4***, logistic regression analyses were performed to assess the extent to which participation and evaluation of the test questions were impacted by the respondents' characteristics. These characteristics included gender (1=female, 0=male), age (numerical), level of education (1=tertiary education, 0=secondary education or less), number of children (numerical), frequencies of camera use and images sharing (1=at least once a week, 0=less often), experience as a Netquest panelist (logarithm of the number of surveys completed in the last three months), comfort with new technologies (1=very or totally comfortable, 0=not at all to quite comfortable) and trust in the confidentiality of the answers (1=trust, 0=no trust).

For evaluation, liking the survey and finding it easy were used as dependent variables. Regarding participation, the dependent variable for the conventional format was the full participation, and for images the minimum participation.

The former was chosen as it presented the ideal scenario in the conventional format, i.e., answering all the 11 questions. The minimum participation was selected for the image-based format because even one photo has the potential to contain all the information of interest, making it the minimum standard for image submissions. These two regressions were employed to test *H4a* to *H4i*.

Results

Respondents' Preferences

To address ***RQ1***, Table 2 presents the preferences of respondents in the *Choice* group before the test questions and after.

Table 2 Preferences of Respondents in Choice Group Before and After Test Questions (in %)

	Preference for…		
Group	Conventional (a)	Image-based (b)	No preference (c)
Before seeing the test questions			
Choice (n=305)	85.6^{b,c}	3.9^{c}	10.5
After seeing the test questions and having stated a preference			
Initially preferred conventional (n=258)	91.1^{b,c}	1.6^{c}	7.4
Initially preferred image-based (n=12)	33.3	50.0	16.7
No initial preference (n=31)	16.1^{c}	12.9^{c}	71.0

Note: Letters in superscript specify the statistically significant differences between categories.

Among respondents who had the option to choose a response format before the test questions, a clear preference emerged: 85.6% favored the conventional format, while only 3.9% preferred the image-based one. These results support *H1*. Another 10.5% expressed no particular preference, which resulted in the *Images-Choice* group being composed of more respondents without a preference than of respondents actively choosing images.

When asking for the reasons for not choosing images to those who preferred conventional questions (261 cases), respondents mainly stated having an extensive book collection and being reluctant to photograph all of them (43.8%), and concerns related to privacy (39.2%).

Regarding preferences after seeing the test questions, 91.1% of participants using the conventional format would choose it again, while only half of those using images would do so. Still, the number of respondents choosing images was very small (n=12), which prevents reaching conclusions on this matter. Finally, 71.0% of those without a preference, who were assigned to answering with images, still did not state a preference after answering the test questions, and 12.9% would choose images after having used them to answer.

Participation

Concerning the participation of respondents in conventional and image-based response formats (*RQ2*), Table 3 presents the breakoff rates, while Table 4 displays the indicators of full, partial, and minimum participation, and the average number of answers responded and photos sent by participants.

Table 3 Breakoff Rate by Response Format and Group (%)

By…	Sample size	Breakoff rate
Response format		
Conventional	897	**0.6**
Image-based	373	7.5
Group		
Text	636	0.6
TextChoice	261	0.4
Images	329	8.2
ImagesChoice	44	2.3

Note: Bold notes statistically significant differences among formats or groups.

The percentage of breakoff is significantly lower among respondents using the conventional format (0.6%) compared to those asked to send images (7.5%). When comparing the groups, there are no significant differences between those with and without a choice, and there is an inclination for breakoff to be more distinct among the images groups: 8.2% in the *Images* group broke off, while 2.3% in the *ImagesChoice* group did so, but the difference is not statistically significant.

Table 4 Other Indicators of Participation by Response Format and Group

By…	Sample size	Type of participation (%) Minimum	Partial	Full	Average number of answers/photos
Response format					
Conventional	892	100	99.9	79.5	10.5
Image-based	345	39.7			2.9
Group					
Text	632	100	99.8	78.5	10.4
TextChoice	260	100	100	81.9	10.6
Images	302	**37.7**			2.9
ImagesChoice	43	**53.5**			2.7

Note: Bold notes statistically significant differences among formats or groups answering through the same format.

As per the different levels of participation in the conventional format, almost all respondents answered at least half of the questions (partial participation), with no major differences among the two conventional groups, and 79.5% of respondents answered the 11 questions (full participation). Although there are no significant differences among groups, the *TextChoice* group has a slightly higher proportion of respondents providing all answers (3.4 pp). Finally, the average number of answered questions is 10.5, with no significant differences among groups.

Regarding the image-based format, 39.7% sent at least one image (minimum participation). In this case, there are statistically significant differences among groups, with 53.5% of those in *ImagesChoice* providing images, against 37.7% among those without a choice in the group *Images*. Finally, the average number of photos per respondent among those actually sending photos (i.e., excluding the 60.3% who did not send the photos when required) is 2.9 photos, ranging from 1 to 16 photos per respondent, and without statistically significant differences between the two image-based groups. Considering all respondents asked for photos (also those who did not send any), the mean number of images drops to 1.2. In all cases, these photos might or might not cover all the books at the residence.

As for the reasons for not sending photos, respondents mentioned privacy concerns (43.0%), technical issues when uploading the photos (13.5%), camera in the device not working (10.6%), and lack of skills (10.1%). In the open-ended question, 11.1% of respondents explained that they were not at home.

Overall, these results confirm *H2a* (lower participation in the image-based format compared to the conventional one). Furthermore, *H2b* and *H2c* are also supported, as significantly higher proportions of participants provided images in the *ImagesChoice* than in the *Images* group, whereas results were more similar between *Text* and *TextChoice*.

Evaluation of the Test Questions

To address ***RQ3,*** respondents' evaluations are presented in Table 5.

Table 5 Easiness/Difficulty and Like/Dislike by Response Format and Group (in %)

	Response formats		Groups			
Categories	Conventional (n=891)	Image-based (n=135)	*Text* (n=632)	*TextChoice* (n=259)	*Images* (n=112)	*ImagesChoice* (n=23)
Easy	64.4	66.7	**61.4**	**71.8**	63.4	82.6
Not easy nor difficult	27.2	26.7	29.0	22.8	28.6	17.4
Difficult	8.4	6.7	**9.7**	**5.4**	8.0	0
Like	**53.5**	**24.4**	52.2	56.8	21.4	39.1
Not like nor dislike	**44.8**	**60.7**	45.9	42.1	61.6	56.5
Dislike	**1.7**	**14.8**	1.9	1.2	17.0	4.3

Note: Bold notes statistically significant differences among formats or groups answering through the same format.

For the easiness/difficulty to answer the test questions, no significant differences are observed between response formats, with most respondents finding it easy to answer the 11 questions (64.4%) as well as capture and send photos (66.7%). These results do not support *H3a*.

Regarding groups, respondents in both *TextChoice* and *ImagesChoice* tend to perceive both formats as easier compared to participants in the equivalent non-choice groups, indicating that offering the option to choose leads to a more positive perception on the ease of both response formats. However, results are significant only for the groups using the conventional format, providing partial support for hypothesis *H3c*. Additionally, the group with the highest prevalence in the category "Easy" is *ImagesChoice*. Further, the perception of easiness of participants in groups without a choice is very similar (63.4% for *Images* and 61.4% for *Text*).

Stronger differences are found when examining the extent to which respondents (dis)liked answering these questions. 53.5% of the respondents liked answering the conventional questions and 1.7% disliked it. In contrast, only 24.4% of those answering through images liked it and 14.8% expressed dislike. Further, there were high levels of indifference (60.7% of "not like nor dislike") among the image-based response format. All the categories present statistically significant differences in favor of conventional questions, providing support for *H3b*.

As per the groups, those who could choose presented higher levels of liking compared to those without a choice, although the differences are not statistically significant. The variation between groups was particularly pronounced for those answering through images: liking among *ImagesChoice* respondents (39.1%) was 18 percentage points higher than in the *Images* group (21.4%). The lack of statistically significant findings does not support H3d.

Of those disliking capturing and sending photos (n=20), 12 individuals reported privacy concerns. As for the conventional format, five of the 15 respondents expressing dislike mentioned they chose "dislike" by mistake, and three found it too difficult, burdensome or time consuming to answer all the questions regarding books. Thus, the reasons for not liking both formats vary.

Impact of Respondents' Characteristics

Impact on Participation

Regarding the impact of participants' characteristics (***RQ4***), Table 6 presents the results of logistic regressions explaining the full participation for the conventional response format (i.e., answering the 11 questions) and minimum participation for the image-based format (i.e., sending at least one photo).

Table 6 Logistic Regressions for Participation

	Participation (=1)	
	Conventional	Image-based
Female	-0.157	0.653**
Age	0.008	0.046**
Tertiary education	0.500***	0.530**
Number children	-0.067	-0.213
Using camera at least once a week	0.559**	0.578*
Sharing images at least once a week	-0.188	-0.045
Number surveys last three months	0.455	0.109
Comfortable with new technologies	0.418**	0.380
Trust confidentiality	0.400**	0.368
Constant	-0.070	-3.438***
n	854	341
Log Likelihood	-412.151	-213.047

Note: *p<0.1; **p<0.05; ***p<0.01

First, having a higher education degree, and using the camera at least once a week have significant and positive effects on participation in both conventional and image-based response formats. These results contradict *H4c* (stating no influence of education) but support *H4d*.

Second, other variables have significant effects for only one response format. For example, while identifying as female affects positively and significantly the participation in image-based response formats, the effect is not significant in the conventional one. Therefore, female participants were more inclined to capture and share images, but gender did not play a role when it came to answering the conventional questions. Since gender influenced at least one of the formats, hypothesis *H4b* is not supported. Similarly, age is a significant factor only for the image-based response format: an older age is associated with higher participation in the questions asking for photos, contradicting *H4a*.

Higher levels of comfort with new technologies and trust in the confidentiality of the answers significantly impact the participation in the conventional response format, but not in the image-based format, not supporting *H4f* and *H4g*.

Finally, the frequency of sharing photos, the number of children, and the number of Netquest surveys answered before this study do not influence participation in either format, thus *H4e, H4h,* and *H4i* are not supported.

Impact on Evaluation

As per the evaluation of the test questions, Table 7 shows the results of the logistic regressions for finding them easy and liking them.

First, education, number of children and frequency of sharing images did not significantly influence the perception of easiness and liking, neither in conventional nor image-based response formats.

Second, some variables influence the perception of easiness or liking but just of one response format. Using the camera at least once a week negatively impacted the liking of the conventional format. Further, being female only influenced (in a positive way) the perception of easiness of the image-based format.

Other variables, as age and trust in confidentiality, affected both dimensions of only one format. While being older decreased the likelihood of liking and finding the questions in the conventional format easy, trusting in the confidentiality of the answers made it more likely.

Finally, some variables impacted both response formats. Feeling comfortable with new technologies increased the likelihood of liking the two formats, and finding the conventional format easy. Further, the number of Netquest surveys completed in the three months prior to this study had significant positive effects on the respondents' perceived ease of the test questions, and in the likability of the conventional format.

Table 7 Logistic Regressions for Easy and Like

	Easy (=1)		Like (=1)	
	Conventional	Image-based	Conventional	Image-based
Female	-0.189	0.742*	-0.003	0.617
Age	-0.026*	-0.004	-0.041***	-0.067
Tertiary education	-0.100	-0.407	0.087	-0.071
Number children	0.004	-0.219	-0.078	0.395
Using camera at least once a week	0.031	0.614	-0.460**	-0.171
Sharing images at least once a week	-0.172	0.235	0.138	0.793
Number surveys last three months	0.644***	1.698**	0.723***	1.037
Comfortable with new technologies	0.636***	0.557	0.444***	1.064**
Trust confidentiality	0.414***	0.393	0.456***	0.722
Constant	0.764	-1.776	1.071	-2.103
n	854	135	854	135
Log Likelihood	-530.533	-77.286	-563.897	-61.894

Note: *p<0.1; **p<0.05; ***p<0.01

Conclusions

Summary of Main Results

In this paper, the focus was on the respondents' preferences, participation, and evaluation of questions answered though conventional or image-based format, and the impact of respondents' characteristics on their participation and evaluation of both formats. Table 8 presents a summary of the hypotheses and their support based on the findings of this study.

First, a clear preference among respondents for the conventional format over the image-based one was found (*RQ1*). These results were very conclusive as only 4% of participants opted for sending photos. These findings contradict the only study investigating respondents' preferences, by Ilic et al., 2022, where most respondents opted for images. This could be due to the task in this paper being more demanding than the one conducted by Ilic et al., (2022), who asked for one photo of three places within the household.

Table 8 Summary of Hypotheses and Their Support

Hypotheses	Result
H1: Higher preference for conventional response format.	Supported.
H2a: Lower participation in image-based format.	Supported.
H2b: Higher participation in the image-based response format when possible to choose.	Supported.
H2c: Participation in conventional format not affected by having a choice.	Supported.
*H3*a: Respondents using the image-based format perceive the test questions as easier.	Not supported
H3b: Respondents like the conventional format better than the image-based one.	Supported.
H3c: Participants with a choice find the test questions easier than those without a choice.	Partly supported (statistically significant results only for the conventional format groups).
H3d: Participants with a choice like the test questions better than those without a choice.	Not supported
No effect on participation in image-based response format of: *H4a*: age, *H4b*: gender, *H4c*: education.	Not supported (positive effect on participation for the image-based format).
Positive effect on participation in image-based format of: *H4d*: familiarity with the device camera, *H4e*: sharing photos with the device, *H4f*: being comfortable with new technologies, *H4g*: trusting the confidentiality of the answers.	*H4d* supported. *H4e*, *H4f* and *H4g* not supported
Negative effect on participation in image-based response format of: *H4h*: more children in the household, *H4i*: higher participation in previous surveys.	Not supported

Second, participation (*RQ2*) was lower among image-based format respondents with only 40% sending photos, compared to 80% of participants in the conventional format answering all questions. Participation in questions asking for images was lower than in some previous studies (55% in Bosch et al., 2019, 49-67% in Bosch et al., 2022), but in line with what was found by Ilic et al. (2022), where less than 40% of respondents in three out of six groups complied with the task. Additionally, participation was below the expected participation when it comes to sending photos of something in the house (54% in Iglesias & Revilla, 2024, also using the Netquest panel in Spain). The variance between the actual and expected participation rates might stem from the different target populations and the increased complexity of the tasks assigned in this study.

Moreover, the option to choose significantly influenced participation rates in questions requesting photos, with 54% of respondents in the *ImagesChoice* group sending at least one (compared to 38% among those in the group *Images,* without a choice), although the sample size of the group abstaining from conventional questions was small. The participation rates were similar to the study by Ilic et al. (2022) (between 50%-78% for those choosing images, and 27% and 39% for those automatically assigned to send photos). Being able to choose did not result in significant differences in participation for conventional questions, also similar to Ilic et al. (2022). The reasons for refraining from sharing photos were predominantly linked to privacy concerns, technical challenges, and participants not being at home during the survey.

Third, the perception of easiness of the test questions was similar between the two formats (***RQ3***), but the conventional format was liked better. Further, respondents in *TextChoice* found the questions about books easier compared to those in the *Text* group. *ImagesChoice* respondents also found the question easier than those without a choice in *Images*, but these results were not significant.

Finally, concerning the factors influencing the participation (***RQ4***), age, being female, and using the camera in the mobile device at least once a week increased the likelihood of participating in the image-based response format questions. Unlike previous literature (Iglesias & Revilla, 2024; Struminskaya, Lugtig, et al., 2021), counting with tertiary education also had a positive effect on participation in the image-based format. However, the number of children in the household, frequency of sharing images, number of Netquest surveys completed, comfort with new technologies, and trust in the confidentiality of the answers did not demonstrate any significant impact.

Regarding the evaluation, being female and completing more Netquest surveys made it more likely to find the image-based format easy, while feeling comfortable with technology favored the liking of this type of questions. Age, number of children in the household, education, trust in the confidentiality of the answers, and frequency of capturing and sharing images did not impact the evaluation of image-based questions.

Limitations

The study has some limitations. First, data were collected in an opt-in panel in Spain. Different results could be obtained in other types of panels or places. Further, respondents in this panel are used to answering questions in conventional formats, which could partly explain their higher participation in that format rather than in the image-based one.

Second, the targeted population was very specific: parents of children attending the first, third, or fifth year of primary school. Having such a specific population might create additional challenges (e.g., respondents less able to capture photos since they are taking care of their children), thus findings from this paper should not be generalized to other topics (i.e., photos of things other than books at home) or to other populations without carefully considering the similarities and differences with the target population and topic of this study. Still, researchers can use these results as a starting point to plan the collection of photos in other settings.

Third, the quality of the information collected through the images and the conventional questions was not assessed in this study. Respondents in the conventional format might have provided approximate answers, not have considered the books of all the members, or invented answers to finish the survey more quickly. Participants sending photos might have photographed only part of the books or sent off-topic photos. Thus, analyses of data quality are needed.

Finally, a significant number of respondents completed the survey from locations other than their home, even if a message asking to answer from home was presented at the beginning of the survey. This predominantly impacted participants using the image-based format, as they could not capture and send real-time photos while answering the survey. Similarly, conventional format respondents willing to count or refer to the books to provide a more accurate answer could not do so.

Practical Implications

The outcomes of this study provide valuable insights for guiding future research endeavors involving photo collection through online surveys. First, participation in questions asking for images is likely to be lower compared to conventional questions. Thus, researchers should balance whether the content of the images obtained outweighs the potentially lower participation.

Second, if researchers are interested in continuing with photo collection, they should consider strategies to improve participation. One approach could be mentioning the reward for sharing photos to participants when presenting the question, allowing them to assess the potential benefits of capturing and sending images. Further, when possible, participants might be informed before the survey about the photos they will be asked for, enabling them to answer the questionnaire when they are able to capture such photos.

Third, although a small part of respondents preferred sending images over the conventional format, having a choice made a difference in terms of participation. This could lead researchers to present respondents with the opportunity to choose to increase participation, but the implications of requiring images (e.g., longer implementation or fees associated to storing images) should be considered before making such a decision involving a potentially limited part of their sample. Given the non-negligible risk of low participation and the great efforts to implement the collection of photos in surveys, maybe we as researchers are not yet ready to replace conventional questions with photos, but we might combine both formats to have additional/complementary information. In this case, it would mean asking for the characteristics of books in conventional ways and additionally ask for the photos of the books.

Fourth, respondents liked the conventional format better than the image-based one, which reinforces the idea of carefully deciding when it is worth it to ask for images (weighting the benefits and disadvantages). Researchers should consider options to make the overall experience more likable, such as giving the option of capturing the photos whenever the respondent wishes to do so (e.g., letting them know before the survey).

Finally, the high number of responses per participant in the conventional format (10.5 out of 11) poses an optimistic scenario, suggesting that respondents will answer questions, even if they are cognitively demanding. However, such answers might be (deliberately or not) incorrect, given the difficulty associated to answer with accuracy the number of books per category and the percentages of books written in certain languages.

Overall, researchers should carefully consider when and how to ask for images in a survey, balancing the benefits of this format (i.e., potential better quality and types of insights) with its disadvantages (lower participation, investment in resources and time). These factors should also be compared with the expected outcomes of conventional questions, in order to decide which type of questions work better, or even consider using both formats to compensate their drawbacks and promote their benefits: combining the two formats might lead to higher participation rates, since respondents would answer the conventional questions and potentially also send photos, which could allow gaining details regarding the number of books, extracting other information of interest, and also assessing the accuracy and quality of the answers provided in conventional ways.

Data availability statement

The script and supplementary materials are available in the project's repository (https://osf.io/7y3sq/). The dataset will be available in the same repository soon after the main papers related to this project are accepted.

Acknowledgements

I am very grateful to Melanie Revilla for her valuable and constant feedback in the development of this paper, and to her, Birgit Heppt, Anna Volodina, and Clemens Lechner for their collaboration on the design of this research. Also, I would like to thank Carlos Ochoa, Oriol J. Bosch and Agustín González for their review and comments on the questionnaire, and María Paula Acuña and Ksenija Ivanovic for their help in checking the programming of the survey.

Funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 849165), the Agencia Nacional de Investigación y Desarrollo (ANID) under the "Becas Chile" Doctoral Fellowship programme (grant No. 72220301 to Patricia A. Iglesias), and GESIS - Leibniz-Institut für Sozialwissenschaften.

Conflict of interest

The author declared no potential conflicts of interest regarding the research, authorship, and/or publication of this article.

Ethical approval

The WEB DATA OPP project, from which this is study is part, was reviewed and approved by the Institutional Committee for Ethical Review of Projects from the Universitat Pompeu Fabra.

Informed consent

All participants were presented with an information sheet before starting and only those providing informed consent could participate in the survey.

References

Bandyopadhyay, H. (2021, July 6). *Image Classification Explained: An Introduction*. V7. https://www.v7labs.com/blog/image-classification-guide

Bosch, O., Revilla, M., & Paura, E. (2019). Answering mobile surveys with images: An exploration using a computer vision API. *Social Science Computer Review, 37*(5), 669–683. https://doi.org/10.1177/0894439318791515

Bosch, O., Revilla, M., Qureshi, D., & Höhne, J. K. (2022). A new experiment on the use of images to answer web survey questions. *Journal of the Royal Statistical Society, 185*(3), 955–980. https://doi.org/10.1111/rssa.12856

Brunello, G., Weber, G., & Weiss, C. T. (2012). Books are forever: Early life conditions, education and lifetime earnings in Europe. *ISER Discussion Paper, No. 842, Osaka University, Institute of Social and Economic Research (ISER).* https://www.econstor.eu/bitstream/10419/92617/1/715521357.pdf

Gil Flores, J. (2011). Medición del nivel socioeconómico familiar en el alumnado de Educación Primaria. *Revista de Educación, 362*. https://doi.org/10.4438/1988-592X-RE-2011-362-162

Güre, Ö. B., Şevgi̇n, H., & Kayri̇, M. (2023). Reviewing the Factors Affecting PISA Reading Skills by Using Random Forest and MARS Methods. *International Journal of Contemporary Educational Research, 10*(1), 181–196. https://doi.org/10.33200/ijcer.1192590

Heppt, B., Olczyk, M., & Volodina, A. (2022). Number of books at home as an indicator of socioeconomic status: Examining its extensions and their incremental validity for academic achievement. *Social Psychology of Education, 25*(4), 903–928. https://doi.org/10.1007/s11218-022-09704-8

Höhne, J. K., & Gavras, K. (2022). *Typing or Speaking? Comparing Text and Voice Answers to Open Questions on Sensitive Topics in Smartphone Surveys* (SSRN Scholarly Paper 4239015). https://doi.org/10.2139/ssrn.4239015

Iglesias, P. A., Ochoa, C., & Revilla, M. (2024). A practical guide to (successfully) collect and process images through online surveys. *Social Sciences & Humanities Open, 9*, 100792. https://doi.org/10.1016/j.ssaho.2023.100792

Iglesias, P. A., & Revilla, M. (2024). Skills, availability, willingness, expected participation and burden of sharing visual data within the frame of web surveys. *Quality & Quantity, 58*(2), 1071–1092. https://doi.org/10.1007/s11135-023-01670-3

Iglesias, P. A., Revilla, M., Heppt, B., Volodina, A., & Lechner, C. (2023). Protocol for a web survey experiment studying the feasibility of asking respondents to capture and submit photos of the books they have at home and the resulting data quality. *Open Research Europe, 3*(202), 1–14. https://doi.org/10.12688/openreseurope.16507.1

Ilic, G., Lugtig, P., Schouten, B., Streefkerk, M., Mulder, J., Kumar, P., & Höcük, S. (2022). Pictures instead of survey questions: An experimental investigation of the feasibility of using pictures in a housing survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 185*(Supplement_2), S437–S460. https://doi.org/10.1111/rssa.12960

McNally, S., Leech, K. A., Corriveau, K. H., & Daly, M. (2023). Indirect Effects of Early Shared Reading and Access to Books on Reading Vocabulary in Middle Childhood. *Scientific Studies of Reading*, 1–18. https://doi.org/10.1080/10888438.2023.2220846

Ochoa, C. (2022). Willingness to participate in geolocation-based research. *PLoS ONE, 17*(12). https://doi.org/10.1371/journal.pone.0278416

Ohme, J., Araujo, T., de Vreese, C. H., & Piotrowski, J. T. (2021). Mobile data donations: Assessing self-report accuracy and sample biases with the iOS Screen Time function. *Mobile Media & Communication, 9*(2), 293–313. https://doi.org/10.1177/2050157920959106

Revilla, M. (2017). Analyzing Survey Characteristics, Participation, and Evaluation Across 186 Surveys in an Online Opt-In Panel in Spain. *Methods, Data, Analyses, 11*(2), 135–162. https://doi.org/10.12758/mda.2017.02

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. (2021). Improving the Use of Voice Recording in a Smartphone Survey. *Social Science Computer Review, 39*(6), 1159–1178. https://doi.org/10.1177/0894439319888708

Revilla, M., Couper, M. P., & Ochoa, C. (2019). Willingness of online panelists to perform additional tasks. *Methods, Data, Analyses, 13*(2), 223–252. https://doi.org/10.12758/mda.2018.01

Revilla, M., Iglesias, P. A., Ochoa, C., & Antón, D. (2022). *WebdataVisual: A tool to collect visual data within the frame of web surveys* [Computer software]. OSF. https://doi.org/10.17605/OSF.IO/R7CAX

Sanders, L. M., Zacur, G., Haecker, T., & Klass, P. (2004). Number of Children's Books in the Home: An Indicator of Parent Health Literacy. *Ambulatory Pediatrics, 4*(5), 424–428. https://doi.org/10.1367/A04-003R.1

Saris, W., & Gallhofer, I. (2014). *Design, evaluation, and analysis of questionnaires for survey research* (Second edition). Wiley.

Sewall, C. J. R., Goldstein, T. R., Wright, A. G. C., & Rosen, D. (2022). Does Objectively Measured Social-Media or Smartphone Use Predict Depression, Anxiety, or Social Isolation Among Young Adults? *Clinical Psychological Science, 10*(5), 997–1014. https://doi.org/10.1177/21677026221078309

Sieben, S., & Lechner, C. M. (2019). Measuring cultural capital through the number of books in the household. *Measurement Instruments for the Social Sciences, 1*(1), 1–6. https://doi.org/10.1186/s42409-018-0006-0

Struminskaya, B., Lugtig, P., Toepoel, V., Schouten, B., Giesen, D., & Dolmans, R. (2021). Sharing Data Collected with Smartphone Sensors. *Public Opinion Quarterly, 85*(S1), 423–462. https://doi.org/10.1093/poq/nfab025

Struminskaya, B., Toepoel, V., Lugtig, P., Haan, M., Luiten, A., & Schouten, B. (2021). Understanding willingness to share smartphone-sensor data. *Public Opinion Quarterly, 84*(3), 725–759. https://doi.org/10.1093/poq/nfaa044

Wenz, A., Jäckle, A., & Couper, M. P. (2019). Willingness to use mobile technologies for data collection in a probability household panel. *Survey Research Methods, 13*(1), 1–22. https://doi.org/10.18148/srm/2019.v1i1.7298

Appendix

Supplementary Material 1: Variables Affecting the Willingness to Share Visual Data

Table 1 Summary of the Effect of Respondents' Characteristics on Willingness to Share Sensor-Collected Data

Variables	No effect	Positive effect	Negative effect
Age		Struminskaya, Lugtig, et al., 2021 Struminskaya, Toepoel, et al., 2021 (65+ y/o)	Iglesias and Revilla, 2024
Female	Iglesias and Revilla, 2024 Struminskaya, Toepoel, et al., 2021	Struminskaya, Lugtig, et al., 2021	
Education	Iglesias and Revilla, 2024 Struminskaya, Lugtig, et al., 2021 Struminskaya, Toepoel, et al., 2021		
Trust in the anonymity of answers		Struminskaya, Toepoel, et al., 2021	
Frequency of taking photos	Struminskaya, Lugtig, et al., 2021	Struminskaya, Toepoel, et al., 2021	
Participation in previous surveys	Struminskaya, Lugtig, et al., 2021	Iglesias and Revilla, 2024	

Note: Empty boxes mean no findings in the previous literature. The results by Struminskaya, Lugtig, et al. (2021) can be applied to actual participation, since all participants who were willing to participate also shared photos.

Table 2 Summary of the Effect of Respondents' Characteristics on the Expected Participation to Share Sensor-Collected Data

Variables	No effect	Positive effect	Negative effect
Age			Iglesias and Revilla, 2024
Female	Iglesias and Revilla, 2024		
Education	Iglesias and Revilla, 2024		
Trust in the anonymity of answers			
Frequency of taking photos			
Participation in previous surveys		Iglesias and Revilla, 2024	

Note: Empty boxes mean no findings in the previous literature.

Supplementary material 2: Fieldwork Document

Download fieldwork document

Supplementary Material 3: Balance Checks for the Experimental Groups

Table 1 Proportion of Categories in Sociodemographic Variables for Participants Assigned to a Group, per Format (in %)

Category	Conventional format (*n=897*) (a)	Images-based format (*n=373*) (b)
Female (vs. male)	50[b]	59
40 y/o or more (vs. 18-39 y/o)	67	62
Tertiary education (vs. low and middle education)	46	44

Note: letters in superscript specify the statistically significant differences.

Table 2 Proportion of Categories in Sociodemographic Variables for Participants Assigned to a Group, per Group Combining Format and Preference (in %)

Category	*Text* *(n=636)* (a)	*TextChoice* *(n=261)* (b)	*Images* *(n=329)* (c)	*ImagesChoice* *(n=44)* (d)
Female (vs. male)	51	49	59	57
40 y/o or more (vs. 18-39 y/o)	66	70	61	68
Tertiary education (vs. low and middle education)	44	49	44	46

Note: Since the analyses are performed between groups of the same format, statistical comparisons were not conducted among groups of different methods. In this table, no statistically significant differences were observed.

Supplementary Material 4: Participation Indicators Considering “Don’t Know” as Participation

Table 1 Breakoff Rate by Response Format and Group, Considering “Don’t Know” as Participation

By…	Sample size	Breakoff rate
Response format		
Conventional	897	**0.6**
Image-based	373	7.5
Group		
Text	636	0.6
TextChoice	261	0.4
Images	329	8.2
ImagesChoice	44	2.3

Note: bold notes statistically significant differences among formats. No statically significant differences were found among groups.

Table 2 Other Indicators of Participation by Response Format and Group, Considering “Don’t Know” as Participation

	Sample	Type of participation (%)			Average number of
By…	size	Minimum	Partial	Full	answers/photos
Response format					
Conventional	892	100	100	98.9	11
Image-based	345	39.7			2.9
Group					
Text	632	100	100	98.6	11
TextChoice	260	100	100	99.6	11
Images	302	**37.7**			2.9
ImagesChoice	43	**53.5**			2.7

Note: bold notes statistically significant differences among formats, and among format-corresponding groups.

Reflective Appendix

When collecting photos of books at home from participants in an online panel, who are predominantly accustomed to answering conventional survey questions, several challenges arose. First, there were difficulties in study design, including decisions about which information to request, what to exclude, whom to target, and how to organize the classification process. Some challenges were related to the substantive questions (e.g., asking for children's grades) and others to the collection of images (e.g., whether to ask for screenshots).

Second, limitations emerged during the data collection and processing stages: only a small number of respondents preferred photos over the conventional format, and more than half did not submit photos when requested due to factors such as privacy concerns, technical issues, or not being at home. Furthermore, some of the photos submitted lacked key information, and the manual classification process led to inconsistencies across researchers, delaying the data analysis.

Challenges in the design phase were anticipated and addressed in advance. However, those that emerged during the data collection and processing stages proved more complex. Notably, fewer participants than expected chose for and submitted photos, which limited the ability to conduct certain supplementary analyses. While these additional analyses were not central to the study, they would have provided valuable insights. As a result, the primary analyses presented in the main paper were completed as planned, but greater participation and preference for photo submissions would have allowed for a more comprehensive exploration of the data.

These issues are likely to persist as long as photos remain an emerging and unfamiliar data type for survey respondents. Consequently, researchers should anticipate facing similar challenges in future studies. However, these obstacles can potentially be mitigated by implementing the recommendations outlined in the final section of this appendix.

Design Difficulties

The first difficulty faced when designing this survey was the definition of the sample. Initially, since this project was designed in collaboration with substantive researchers who wanted to study the link between the books at home and the children's grades at school, the target population of interest were parents of children in primary school. However, changes in the regulation in Spain regarding the evaluation system (see section "Sample and data collection" in the paper) made it necessary to adjust the target population to parents of children in first, third, or fifth year of primary school. Without this adjustment to the sample, respondents might not be able to provide a grade depending on the year of primary school attended. This adjustment led to a lower quantity of Netquest panelists that matched the required survey profile. Further, a 3% error margin was added for the quotas to consider the lack of exact information about the target population and ensure their fulfillment.

Second, respondents in the conventional format were asked about the total number of books and their distribution among the following categories: books for toddlers and children who do not know how to read, books for literate children and teenagers, and books aimed at a general audience. While the need for including categories was clear, it was difficult to decide which exact categories to use and how to classify books within these categories (e.g., books containing text and drawings could be intended for toddlers, read to them by others, or for literate children capable of reading on their own).

Third, one initial concern regarded eBooks and the potential difficulties to capture this information. If asked for photos, respondents might have had to photograph the screen of the device, which might or not have contained the cover, titles, or author of the books. Similarly, if they were to answer the survey from the same device used for reading, respondents would have had to quit the survey, take screenshots, return to the survey, and then upload them. Ultimately, the collection of eBooks was discarded given results in previous literature indicating their lack of impact on children's academic achievement and cultural capital (Heppt et al., 2022; Otte, 2023; Pagel & Heppt, 2016).

Fourth, photos were first classified manually for two main reasons: 1) technical limitations, as discussions with computer vision experts revealed that algorithms needed improvement to accurately count books in photos when arranged in various ways, thus having a training dataset would be helpful, and 2) enhancing the algorithms was beyond the team's skills. Therefore, manual classification was the most fitting option to start for the project. However, the research team always considered that implementing an automatic classification, and comparing it with the manual one, would also be of high interest. Thus, collaborations were established to further investigate this option.

Main Limitations During Data Collection and Processing

One of the limitations was the inability to control whether respondents were answering from home, which was relevant for both response formats: being at home could help provide a better estimate for conventional questions, and it was essential for sending photos of the books at home. Initially, the inclusion of a question asking where respondents were answering the survey was considered (e.g., Revilla and Couper, 2021, did that when asking for voice answers), but the location could not be confirmed as geolocated data were not collected. Consequently, such a question would not distinguish between participants who were actually at home from those merely stating they were. Instead, a reminder was included at the survey's outset, urging participants to respond from their home locations (if they were not at home, they could leave and re-access the survey with the link included in the invitation e-mail or a link available in the fieldwork company's app). Regrettably, this instruction seemed to go unnoticed or was disregarded by some respondents when accessing the survey. Indeed, 8% of respondents who did not submit photos when asked, stated that they were not at home. However, this number could be higher since "not being at home" was not an option in the radio-button question that asked for reasons for not uploading photos. Instead, respondents wrote this reason in the "Another reason" category, where they could type their own explanations. I expect that not being at home might have impacted the quantity of photos submitted, although it is uncertain whether those respondents would have submitted the photos even if they had been at home.

Additionally, fewer people than expected chose the images-based format (only 4%). Although H1 stated that respondents would prefer the conventional format over the images-based one due to the complexity of the task, I did not expect the difference to be so pronounced as previous research, based on a simpler task, found a higher preference for photos (Ilic et al., 2022). Moreover, around 40% of respondents actually asked for photos shared at least one. The low resulting proportion in the preference for photos and participation when asked for them, prevented the execution of some of the planned analyses, as assessing the impact of the respondents' characteristics, behaviors, and opinions on their preference for one response format over the other. Further, there are less images remaining for the analyses on data quality.

Moreover, photos were not always clear. Each dwelling is different, and while some books are well organized and clear in the photos, others are arranged in ways that is not possible to read the title or discern features that would allow classification in the three categories. This is critical especially for children's and teenagers' books, since these books are often stored in ways that make it more difficult to extract the information. This difficulty was anticipated when designing the survey, and thus specific instructions (which are available in the protocol by Iglesias et al., 2023) were designed to try to minimize the problem. These instructions aimed to include all the relevant information while also being as concise as possible. They explained how the photos should be taken in terms of lightning, distance to the books, and exclusion of distracting items. However, it was not expected that respondents will follow all the instructions since a lot of efforts on their side was needed to do so (e.g., removing several personal items from in front of the books). Moreover, the instructions and visual examples shown were related to adults' books in shelves: instructions for children's books or other types of storage were not presented, even if they might have been relevant.

Further, since classification was conducted manually, inconsistencies between researchers due to the complexity of the task were identified, especially in assigning books to categories. This was addressed by constantly reviewing differences between classifiers and re-classifying photos (thus extending the duration of the classification step), but inconsistencies were not completely eliminated. Fewer challenges arose regarding consistency in storage and languages, but errors might still have occurred due to the coexistence of different languages in Spain and the fact that researchers were not fluent in all of them. In any case, the photos and manual classification outcomes are intended to be used by computer vision experts to improve existing algorithms. These improvements are expected to facilitate at least an accurate total count of books, which would be a significant contribution for research involving the collection of photos of books. The potential of such improvements will depend on both the resources available to the computer vision experts and the accuracy of the initial manual classification, since errors in manual classification of the training photos (i.e., photos of books collected and classified in this study) might lead to inaccurate results of the algorithms.

Recommendations for Future Research

Based on the experience in this study, the recommendations for researchers are:

1. To plan ahead when considering the collection of photos. Many extra steps are needed, starting with the programming of a tool allowing the collection of photos in a survey, which needs to be tested in different moments, operating systems, and browsers. These tests might lead to time-consuming improvements. Furthermore, images need to be stored in safe folders with enough capacity as to contain all the photos, which can be of several megabytes. Further, the servers storing such folders need to operate quickly to ensure that the experience of respondents uploading the photos is positive. Servers offering this service are for payment, thus funding for this needs to be assigned when planning the project.

2. Clear operationalization of the items to be observed in the photos is crucial from the outset, as this could be the first step enabling researchers to discern whether images are the best fit for their study. In this regard, researchers must clearly define the items they want to extract from the images, and establish the method for extracting the information. For more information on this matter and the other steps to be considered before, during, and after image collection, see Iglesias et al. (2024).

3. The definition of the items should also be conveyed to respondents when it does not interfere with the project's objective, so they can easily identify if the items of interest (e.g., books) are clearly visible in the photos.

4. Finally, the classification of the images is a critical issue. In any research, the method of classification (manual and/or automatic) needs to be defined, and the necessary resources must be allocated accordingly. In the case of this study, classification was manual. For this purpose, guidelines and examples were created to train the classifiers. However, it is important to note that manual classification is time-intensive, demanding meticulous attention to details and potential problems with the photos.

Researchers interested in collecting photos through surveys should be aware that there are numerous practical challenges involved in the design, collection, and analysis stages, more than with conventional questions alone. Therefore, it is crucial for researchers to have a well-defined plan for the entire process to ensure that photos are collected and analyzed successfully, making them valuable in addressing the research questions.

To achieve this, the guidelines provided by Iglesias et al. (2024) could be particularly helpful, as they offer a comprehensive overview of the entire process, from operationalization to analysis. However, given that challenges may still arise, researchers should remain flexible and be prepared to adapt their approach as necessary. For example, they should anticipate potential low participation rates and have a contingency plan in place, such as supplementing photo collection with conventional survey questions for respondents who do not provide photos.

methods, data, analyses | Vol. 19(2), 2025, pp. 87-124 DOI: 10.12758/mda.2025.08

Researching the Moment of Truth: An Experiment Comparing In-the-Moment and Conventional Web Surveys to Investigate Online Job Applications

Carlos Ochoa

Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra

Abstract

Understanding how people seek and apply for jobs online is crucial for addressing social inequality, discrimination, and aiding companies in attracting suitable candidates. Conventional surveys struggle to capture the nuances of online job searches that, as many online events, are characterized by repetition, low distinctiveness, and limited emotional impact. These characteristics can lead to memory-related errors, becoming more likely as the time between the event and the survey increases. Passively collected data, such as metered data provided by online panel members who install tracking software on their browsing devices, offer an alternative. While these data provide objective insights into online job searches, they suffer other types of errors, and cannot capture subjective information and all potential objective data of interest. This paper explores an alternative approach: sending surveys to individuals in a metered panel shortly after an event of interest is detected through metered data. These "in-the-moment" surveys aim to fill in missing information not obtainable through passive data collection while reducing memory-related errors that affect conventional surveys. To assess the feasibility and benefits of this method, an experiment comparing in-the-moment surveys triggered by online job applications with conventional surveys was conducted in an opt-in online panel in Spain to research how people apply for a job online. The results reveal that metered panelists accept well in-the-moment surveys, displaying high participation levels and positive evaluations regarding effort and satisfaction, without perceiving an increased privacy risk. Moreover, the data indicate positive impacts on data quality, with longer and more detailed responses to open-ended questions. However, not all aspects saw substantial improvements, with the reduction of non-recall being weaker than expected, possibly due to participants' overconfidence in their memories. The significant disparities observed in substantive results between both types of surveys also suggest that participants are not fully aware of what they do not remember.

Keywords: in-the-moment surveys, metered data, passive data, web surveys, digital traces, job applications.

The job market consistently ranks among citizens' top policy priorities in most countries,[1] as employment is vital for economic growth (Boltho & Glyn, 1995) and a fundamental aspect of mental health at the individual level (Ezzy, 1993). Within this field, job search is receiving increasing attention from researchers. Over the past decades, the internet has transformed job searching. By 2015, 54% of U.S. adults had researched jobs online (twice as many as in 2005), and 45% had applied for jobs online (Smith, 2015). Since then, the internet has become a crucial employment resource globally, but to an unequal extent for different population groups, such as older and less educated individuals. Therefore, understanding online job search and application behaviors is crucial, particularly for defining policies against social inequality and discrimination (Karaoglu & Hargittai, 2022) and for helping companies attract suitable candidates (Mansouri et al., 2018).

Several aspects of online job search have been investigated: effectiveness to escape unemployment (Kuhn & Mansour, 2014), impact of online reviews on job seekers (Faiz, 2020), use of different online platforms and their outcomes (Dillahunt, 2021), digital inequalities (Karaoglu & Hargittai, 2022) and gender differences (Fluchtmann et al., 2021).

However, research on job search is limited by the lack and/or inadequacy of available data. Most studies rely on surveys where participants report past job searches, which can be significantly affected by memory limitations. Job search involves a series of repetitive events (i.e., finding, reading, and applying for job offers) that are low in distinctiveness and emotional impact, involve little rehearsal (i.e., minimal time spent thinking/talking about each event), and are of short duration. These factors, combined with the passage of time, increase the likelihood of memory errors (Tourangeau, 2000) and recall bias (Walker & Skowronski, 2009). These issues are also prevalent in other online activities studied by researchers, such as housing searches, online purchases, and media consumption.

[1] https://www.undp.org/sustainable-development-goals, https://www.ipsos.com/en/global-public-ranks-ending-hunger-and-poverty-and-ensuring-healthy-lives-top-priorities-among-un

Acknowledgements

I am very grateful to Melanie Revilla for her valuable and constant feedback and guidance in the development of this paper, and to her, Eva Fortes, Giampiero Passaretta, María José Gonzalez and Clara Cortina for their collaboration on the design of this research. Also, I would like to thank Patricia Iglesias for their review and comments on the questionnaire. 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). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Direct correspondence to

Carlos Ochoa, Research and Expertise Centre for Survey Methodology,
Universitat Pompeu Fabra, Barcelona, Spain
E-mail: carlos.ochoa@upf.edu

Passively collected data, which do not require active participation in the data gathering from the observed individuals (Link et al., 2014), have also been used to study job seeking. For example, the professional social network Linkedin.com regularly releases job market reports based on data from job seekers and employers using their services. Similarly, opt-in online panels like Netquest and Yougov have requested some of their members who regularly participate in surveys to install tracking software on their browsing devices (a “meter”) to gather information on online activities, such as visited URLs, search terms, and app usage. Researchers can use these “metered” panels (Revilla et al., 2021) to investigate online job search behaviors.

In contrast to survey data, metered data (a type of digital trace data) are not subject to memory errors (Revilla, 2022), making them well-suited for collecting objective information, such as the number of job offers accessed per day or the time spent per offer. However, metered data can be affected by other errors (see “Background” section) and cannot capture subjective information, like why someone decides to apply for a job. Furthermore, meters cannot capture all objective information, such as whether the candidate secured the job.

This paper explores a third option: sending a survey to a sample of individuals from a metered panel when an event of interest is detected using metered data. This method can add missing information that cannot be collected passively, while reducing memory errors that affect retrospective surveys by shortening the time gap between the event and the data collection. However, some doubts arise about its applicability and effectiveness: (1) Will panelists agree to participate in such “in-the-moment surveys”? (2) How will respondents evaluate their experience? and (3) To what extent can these surveys provide better or new data compared to retrospective surveys?

This paper addresses these questions by reporting the results of an experiment comparing an in-the-moment survey triggered by online job applications detected through metered data with a conventional survey, i.e., a retrospective web survey sent to members of an opt-in online panel asking whether they applied for a job in the last six months. Both surveys requested additional information about one job application, along with sociodemographic and personality trait questions.

Background

Metered Data

Metered data offer substantial advantages over surveys for measuring online behaviors, such as greater granularity and robustness against memory errors, but are not error-free.

First, metered panels are usually formed from a subset of opt-in online panel members, which may introduce self-selection bias due to their non-probability-based recruitment (Baker et al., 2010). This bias can be exacerbated when panelists are asked to install a meter. Revilla et al. (2021), examining Netquest panels in nine countries, found that panelists who installed the meter when offered differed from those who did not in terms of gender, education, income, age, and panel loyalty. This self-selection may limit the capacity of metered panels to produce precise population estimates. Nonetheless, despite these limitations, opt-in online panels remain prominent in online research,[2] with metered data becoming increasingly popular in media, political, and social research (Revilla, 2022).

Besides representativeness issues, metered data suffer from other errors often overlooked by researchers, as discussed in Bosch and Revilla's (2022) Total Error framework for digital traces collected with Meters (TEM). When using metered data to trigger in-the-moment surveys, researchers must consider that these errors may cause the non-detection of events that should trigger surveys ("false negatives") and the detection of events that should not ("false positives;" Bosch et al., 2025).

False negatives and false positives can arise from various scenarios. False negatives include pausing the meter during job applications, technological limitations, inability to track mobile apps events, and overlooking relevant URLs. False positives can result from shared metered devices (Revilla et al., 2017), leading to incorrect event attribution. A key issue that can cause both false negatives and false positives, depending on the researcher's decisions, is when websites use the same URL for multiple events.

In summary, using metered data to detect events is fallible, potentially resulting in a sample that is not representative of all job applications. Additionally, non-detection of events extends the fieldwork time needed to reach the target sample size. Conversely, false positives require longer questionnaires with screening questions to exclude mistakenly detected participants, increasing data collection costs.

In-the-Moment Surveys Triggered by Metered Data

Participation

Previous research shows that metered panelists exhibit an overall high willingness to participate in in-the-moment surveys triggered by metered data, ranging from 69% to 95%, depending on the conditions offered to participants (Ochoa & Revilla, 2022). However, stated willingness may not always translate to actual

[2] https://shop.esomar.org/knowledge-center/library-2021/Global-Market-Research-2020-pub2942

participation due to practical issues, such as not receiving or seeing the survey invitation in time (Ochoa & Revilla, 2022).

There has been little experimental research on in-the-moment surveys triggered by metered data, with one notable exception by Revilla and Ochoa (2018). In this study, a pop-up invitation was used to invite metered panelists of the Netquest panel in Spain to take part in a survey when a flight purchase was detected, but only 18 individuals completed it. The authors cited technological issues and short fieldwork times as possible reasons for the low participation. Overall, the limited evidence suggests that obtaining participants in the moment is a significant challenge.

Reduction of Recall Errors

The use of in-the-moment surveys aims to minimize the time gap between the event of interest and data collection, thus reducing errors from memory limitations. Tourangeau (2000) identifies four classes of memory problems: encoding issues (experiences not properly recorded), storage problems (corrupted memories), retrieval failures (inaccessible memories), and reconstruction errors (partially retrieved memories are inaccurately reconstructed).

When asking participants for subjective evaluations instead of factual data, such as reflections on feelings during an event, memory issues can worsen. These evaluations might not have been formed at the time, leaving nothing to remember, an extreme form of encoding problems. When asked later, reconstruction errors can occur if evaluations are made a posteriori, combining factual memories and present circumstances. This can be related to the discrepancy between the experiencing and remembering self (Kahneman & Riis, 2005).

The longer the time between an event and its recall, the greater the chance of retrieval and reconstruction failure. This applies to all types of events, from hospital stays to consumer purchases (Jobe et al., 1993). Most theories attribute the decline in accessibility over time to the interfering effects of later experiences, making online events (frequent and repetitive) particularly susceptible to rapid forgetting.

Some research methods leverage this fact to improve data quality. Ecological Momentary Assessment (EMA), for instance, prompts participants to report their current experiences via alarms sent at a predetermined or random schedule (Shiffman et al., 2008; van Berkel et al., 2017). This method avoids retrospective reporting but is impractical for studying specific events, as it would require frequent surveys to capture individuals experiencing the event of interest by chance. Coincidental surveys (Lamas, 2005) tried this approach in the early 20th century for measuring radio audiences but were found to be costly and operationally difficult.

In-the-moment surveys are similar to EMA, but target only those detected experiencing an event of interest. Besides metered data, other passive data sources can be used for detecting events and triggering surveys: GPS data (known as geofencing, e.g., Haas et al., 2020), smartphone accelerometer data (e.g., Hardeman, 2019), and Bluetooth beacons (e.g., Allurwar, 2016). All these methods combine self-reports with passively collected data (Keusch & Conrad, 2022).

However, even with these methods, a time gap between the event and the response may remain. While smaller than in retrospective surveys, this gap can still affect data quality.

Several empirical models have been proposed to quantify information loss over time (Rubin & Wenzel, 1996). They all predict that forgetting occurs rapidly at first and then slows down, supporting the benefits of in-the-moment surveys. However, since these "retention functions" may vary by individual and context, it remains unclear how close to the event surveys should be conducted to achieve a positive effect.

In this regard, Revilla and Ochoa (2018) compared survey responses collected up to 48 hours after the event of interest (probably too long to be considered "in the moment") with up to two months later, finding no significant differences in answers.

Research Questions, Hypotheses, and Contribution

Based on the literature on in-the-moment research and its limitations, the following research questions and hypotheses are proposed.

RQ1. What are the levels of participation for in-the-moment surveys triggered by metered data among metered panelists compared to an equivalent conventional web survey?

Participation is expected to be slightly lower for in-the-moment surveys (*H1*). The technology used in this study, developed to detect online events and invite participants shortly after (see section "Software" in "Data and Methods"), aims to bridge the gap between the high willingness to participate reported in the literature and the low participation observed in the single previous study. However, in-the-moment surveys may interrupt participants, raise privacy concerns by highlighting the implications of sharing metered data, and be perceived as intrusive (Ochoa & Revilla, 2022), potentially decreasing participation.

RQ2. How do participants evaluate in-the-moment surveys compared to conventional web surveys?

In-the-moment surveys might be easier for participants, as they ask about fresh experiences and may seem more relevant. However, including questions unrelated to the event, such as sociodemographic ones, could dilute this positive effect. Additionally, the same issues affecting participation (see RQ1) may also impact evaluations. Thus, overall evaluations are expected to be similar to those of conventional surveys (*H2*).

RQ3. Is data quality higher from in-the-moment surveys compared to conventional web ones?

I expect in-the-moment surveys to produce better data quality (*H3*) due to reduced memory errors, such as fewer "don't remember/don't know"[3] responses (i.e., explicit non-recall). Additionally, lower effort required to answer event-related questions and higher respondent interest may reduce satisficing. This could lead to higher data quality, with longer and more meaningful responses to open-ended questions and fewer invalid and inconsistent answers.

Finally, some information (e.g., exact submission time of the job application) can be obtained from metered data for in-the-moment surveys, while conventional surveys require direct questions. This may lead to differences in non-response rates and answer precision.

RQ4. Do in-the-moment surveys lead to different substantive results compared to conventional surveys?

Both methods may yield different responses to the same questions due to varying sources of error and selection bias. Consequently, I expect different results for comparable questions related to the event of interest (*H4*). Assuming no other differences between the two methods except the time elapsed since the event (i.e., controlling for selection bias), responses given in the moment should have higher credibility. Therefore, discrepancies in substantive answers may indicate distorted recall (i.e., memory alterations of which respondents are unaware).

By addressing these research questions, this paper contributes to the existing knowledge in several ways. First, it explores the feasibility and potential benefits of in-the-moment surveys triggered by metered data, a topic not yet researched. Second, it tests a new approach to overcoming the technical issues that hindered the only previous academic attempt to develop such surveys. Finally, it evaluates in-the-moment surveys for studying real-world issues like job searching, which may inspire practical applications and provide insights into participants' perceptions of risks (e.g., privacy) and benefits in a real-world setting.

[3] While these answers do not convey the exact same meaning, in practice, they are often indistinguishable. Participants unable to provide an answer may not be aware if they did not see the requested information or if they forgot it. Both types of answers are reported separately in SOM4.

Data and Methods

Data

The data were collected from the Netquest opt-in online panel (www.netquest.com) in Spain. Netquest panel members regularly participate in surveys and earn points proportional to the length of the surveys, which can be redeemed for gifts (Revilla, 2017). In addition, some panelists are offered the possibility to install the meter in exchange for two to 12 additional points per week, depending on the number of devices where the meter is installed. The panelists invited to join the metered panel are not randomly selected from the survey panel. Instead, Netquest selects them based on their likelihood to accept the meter installation, determined by an internal predictive algorithm, and the need for participants for different research projects. The average installation rate is between 20% and 42%, depending on the country (Revilla et al., 2021).

Data for the in-the-moment surveys were collected from the 10th of March to the 3rd of October 2023 (207 days) using the metered panel. Data for the conventional survey were collected from the 30th of May to the 4th of June 2023 (five days) using the opt-in online panel, which includes the members of the metered panel.

In this study, the objective was to compare two samples of around 200 panelists who had applied for a job. A detailed description of how both samples were produced can be found in the “Methods” section, as this process is a fundamental part of this research.

Participants in the in-the-moment survey who did not confirm having applied for a job (105) and those who responded 48 hours after the application (21) were discarded. This decision was based on the results of Revilla and Ochoa (2018), who did not find relevant differences between responses collected 48 hours after the event and those collected up to two months later. Consequently, the final number of valid participants in this survey was 177, all of them metered panelists. Among them, 46.9% responded through a mobile device (smartphone or tablet). Their average age is 41.7 years, with 55.4% being women. 49.7% are mid-educated and 44.6% are highly educated. Their median number of participations in surveys in the last three months is 32.

As for the conventional survey, the number of valid participants was 201, out of which 56 were metered panelists, and the remaining 145 were participants in regular surveys only. 71.6% responded through a mobile device. The average age in this group is 38.6 years, with 61.2% being women. 47.3% are mid-educated and 44.8% are highly educated. Their median number of participations in surveys in the last three months is 25.

Both samples present significant differences in age, number of participations in surveys in the last three months, being metered and the device used to participate (see Appendix 1).

Software

Given past technical issues reported in the literature that made the implementation of in-the-moment surveys difficult, this study used WebdataNow (Revilla et al., 2022), software specifically designed for conducting in-the-moment surveys triggered by metered (or geolocation) data.

WebdataNow performs three main functions: (1) receiving metered (or geolocation) data from a panel, (2) identifying events of interest in the data, and (3) triggering survey invitations to the relevant panelists. The events of interest are defined by a list of regular expressions[4] that match the URLs intended to trigger the survey. Additionally, WebdataNow allows researchers to set a notification delay (the time between event detection and survey invitation) and a maximum time limit for participants to access the survey after the invitation is sent. For this study, the notification delay was set to five minutes.

Methods

To address the research questions raised in this study, the same topic (how individuals decide to apply for a job) was investigated using an in-the-moment and a conventional survey. These methods differ essentially in how candidate participants are selected and invited to participate. In addition, the questionnaire had to be adapted to each method. The following sections cover such differences.

Sample Selection

When selecting candidate panelists for the two samples, both metered and non-metered panelists were eligible for the conventional survey, while only metered panelists were eligible for the in-the-moment survey.

In this research, the opt-in online panel had fewer metered than non-metered panelists, risking the failure to reach the target sample size for the in-the-moment group. Prioritizing metered panelists for the in-the-moment group would have meant that the conventional survey sample consisted only of non-metered panelists. This would have led to two issues: first, it would not have provided a realistic sample for the conventional group, as the Netquest panel typically includes metered panelists in regular survey samples; second, it would have created a perfect correlation between the method (in-the-moment versus conventional) and the type of panelists (metered versus non-metered), hindering the identification of method-specific effects, which is the primary focus of this study.

[4] A regular expression is a sequence of characters that specifies a search pattern in text. See Appendix 2 for examples.

Therefore, I proceeded as follows. First, the in-the-moment survey was activated for a randomly selected half of the metered panelists, meaning that panelists who applied for jobs on one of the pre-identified websites (see subsection “In-the-moment survey”) using a device with the meter installed would receive an invitation to participate.

Second, once more than half of the in-the-moment target sample was achieved (30th of June 2023), non-metered and metered panelists who were not activated for the in-the-moment survey were randomly selected to form the sample for the conventional survey. The number of invitations sent was determined based on the target sample size (200) and the estimated proportion of panelists seeking a job (15%). This estimation was based on the number of visits to job search websites observed in the whole Netquest metered panel for Spain over a six months period.

Finally, after reaching the sample target for the conventional survey, the remaining non-invited metered panelists were activated for the in-the-moment survey. The detection of job applications stopped when the target sample size was reached (3rd of October 2023).

Following this process ensured that the conventional group included some metered panelists and that panelists were invited to participate in only one of the surveys.

In-the-Moment Survey

Panelists in the in-the-moment group received an invitation to participate in a survey five minutes after applying for a job on one of the listed websites (see Appendix 2). This invitation was sent only once during the project, the first time they were detected. This list covers the most popular job search platforms in Spain for which it was possible to identify a unique URL shown when a visitor applies for a job. Since the meter used in this research did not allow for the detection of activity occurring within apps, applications from apps could not be detected either.

The inability to detect all the participants’ job applications, together with other sources of error affecting metered data mentioned in the “Background” section, led to high levels of false negatives and false positives. Although not directly measurable, such levels are estimated to be around 85% and 34%, respectively (see SOM1).

All the detected panelists received the invitation by email. Additionally, panelists using the panel app also saw a push notification on their smartphones and/or tablets. The panel app, which facilitates various aspects of panel membership (e.g., invitation, redemption of points for incentives), can be installed voluntarily by the Netquest panelists, but is mandatory for those who want to install the meter on a mobile device (Revilla et al., 2021). As a result, approximately 90% of the metered panelists have the app.

Since the invitation to participate was sent via both emails and push notifications, participants did not necessarily take the survey on the same device where the job application was detected.

The message included in the invitation emphasized that participation time was limited, but without specifying a clear limit. However, the potential impact of this mention on participation was expected to be low since it could only be seen after opening the email and/or clicking on the app notification. We introduced a time limit message to encourage participants to provide responses promptly. Nevertheless, we allowed participants to complete the survey after this time limit to explore whether individuals would still participate, enabling us to potentially compare them with respondents who answered shortly after receiving the invitation. However, due to the limited sample size, a conclusive analysis comparing these two groups proved unfeasible.

Twenty-five percent of respondents completed the survey within 15 minutes after the job application and 50% within 72 minutes. However, the distribution of the delay in participating is strongly skewed to the right: 25% of respondents took more than 8 hours to participate, and 10% more than 2.3 days. For the analyses, we excluded responses submitted more than 48 hours after the invitation, as explained in the “Data” section.

Due to the possibility of shared devices, the questionnaire was designed to confirm that the panelist was indeed the one who applied for the job, without disclosing private information in case it had been obtained from a third party. To achieve this, participants were asked, after obtaining informed consent, whether they had engaged in four different online activities within the last 48 hours. One of these was “reading job offers”. Only those who responded affirmatively to this question were allowed to proceed with the questionnaire. By adding this step, the risk of revealing third-party private information to the participant and causing any harm was considered to be extremely low.

Then, the questionnaire explicitly informed participants that they were invited to participate because they were detected looking at a job description, with the specific website and approximate time of detection provided. Approximately half of the sample was informed that the survey was sent close to the event of interest to enhance the quality of the data for researchers, while the other half was told that the purpose was to help them recalling their answers more easily. This message aimed to assess if the communicated benefit of the method had an impact on the results.[5]

Participants were then asked to confirm whether they had visited the job offer and whether they had finally applied for the job. After this section, the questions used for both the substantive and methodological research were presented. An

[5] The results of this experiment indicated a slight and non-significant effect in favor of communicating that the main benefit is for the respondent in terms of breakoff and survey evaluation. For more details, see SOM4.

English translation of the full questionnaire and screenshots are available in the supplementary online material (SOM2).

The questionnaire aimed to assess potential differences in online job application behavior among various demographic groups (especially males and females), including whether participants met all job requirements, their self-reported likelihood of being hired, and whether the job position met their expectations. Additionally, the questionnaire included questions about participants' sociodemographic background and personality traits (19 items in two batteries of questions). One of the key substantive hypotheses that this survey aimed to confirm was whether females applied for a job offer less frequently than males when they did not meet all the requirements of the job offer (Ochoa et al., 2023).

Five questions to assess participant's evaluation of the survey were also asked in the questionnaire, before the personality trait questions.

The full questionnaire included up to 69 questions and was optimized for mobile devices. The average time to complete it was 10.2 minutes and the median 9.2 minutes. Respondents could continue without answering the questions, except those used to filter other questions. A warning message was shown to 7.9% of participants who tried to skip a question when multiple questions were presented on the same page. Following the panel's usual practice, going back was not allowed.

Conventional Survey

The in-the-moment questionnaire was adapted to be used in a conventional survey. Since in a conventional survey it is not possible to refer to a concrete job application detected using metered data, participants were asked about their most recent job application in the last six months. Questions such as "Why did you apply for this job offer?" were rephrased as "Think about the last job application you submitted online for a job offer. Why did you apply to this job offer?"

Besides reformulating job application related questions, other changes were made:

- The initial section designed to verify that the panelist was the one who made the job application was removed.
- Two questions were added to gather when and in which website the application took place. In the in-the-moment survey this information was gathered using metered data.
- A question was added asking participants to what extent they were confident (0 to 100%) that the job application they reported was actually the last one they did.

The final conventional questionnaire (see SOM2) included up to 69 questions. The average time to complete the questionnaire was 9.6 minutes and the median 8.6 minutes. All the remaining features of the questionnaire (e.g., possibility to skip questions, etc.) were the same as in the in-the-moment questionnaire. The warning message presented when trying to skip a question in pages with multiple questions was shown to 11.4% of participants.

Analyses

Comparisons Between Groups

The analyses were performed using R version 4.2.3. Various metrics (participation metrics, survey evaluations, quality indicators, and substantive answers) are calculated for participants in the in-the-moment group and compared with those in the conventional group.

When the calculated metrics represent proportions (e.g., proportion of participants evaluating the survey as "easy"), Fisher exact tests are used for group comparisons. For metrics representing means (e.g., mean number of characters in answers to an open-ended question), *t*-tests are used for comparisons. Metrics representing means of reported percentages or probabilities (e.g., estimated probability of being hired) are compared also using *t*-tests. In all these cases the resulting *p*-values are reported.

As described in the "Data" section, the sample selection method did not guarantee equal sample compositions in both groups. To account for these differences, logistic regressions are conducted for dichotomous variables, and linear regressions for continuous numerical variables, while controlling for three sociodemographic variables: gender (two groups), age (numeric), and education level (two groups). Additionally, two panel variables are used as controls: the number of participations in Netquest surveys in the three months before this study[6], and being a metered panelist. The inclusion of this last variable is crucial because all participants in the in-the-moment group are metered, whereas only 28% of the conventional group participants are. This factor could potentially confound both the method and sample composition effects in a direct comparison. Similarly, the type of device used to complete the survey, which may influence data quality (Lambert & Miller, 2015), was included as a control variable because the proportion of PCs is significantly larger in the in-the-moment group (see Appendix 1).

Similar to the direct comparisons, *p*-values are reported for the regression analyses, using a significance level of 5% in both cases. However, due to the limited sample size, detecting significant effects with all these covariates poses challenges, especially for questions presented to only a subset of respondents due to the questionnaire's routing conditions.

[6] I also attempted including the total number of participations in panel surveys and the log transformation of both variables as covariates, which yielded comparable results. Details of these analyses can be found in the SOM4.

Open Questions

Several open-ended questions were used to gather both objective (e.g., name of the employer) and subjective (e.g., main reason to switch jobs) information about the job applications. Answers to such questions required coding to capture their substantive meaning while also assessing data quality, including non-recall, off-topic answers, or overall answer coherence.

Answers were coded by two native speakers. Initially, the main coder created a codebook. Then, a secondary coder used the same codebook to repeat the process. The intercoder reliability was 96%. The reported results are those produced by the main coder, after reviewing those of the secondary coder.

Participation

To compare participation levels between the two groups, three primary metrics are used: (1) Start rate, which indicates the proportion of invited panelists who initiate the survey (starts) compared to the total number of invited panelists (invites). (2) Breakoff rate, which represents the percentage of panelists who abandon the survey (breakoffs) divided by the number of panelists who start the survey (starts). (3) Incidence rate, calculated as the number of valid surveys (completed surveys not discarded due to screening questions) over the total number of completed surveys (completes).

The start rate helps assessing whether inviting panelists while they are actively engaged in the activity of interest (job search) leads to a higher likelihood of them disregarding the survey invitation. Conversely, the dropout rate provides insights into whether inquiring about a recent meter-detected activity prompts participants to abandon the survey less frequently without completing it. Lastly, the incidence rate, which measures fieldwork efficiency (Ochoa & Porcar, 2018), assesses the potential benefits of contacting people in the moment in terms of sample utilization.

Survey Evaluations

Five questions are used to evaluate participants' perceptions of both surveys: self-reported effort to participate, satisfaction, trust in survey anonymity, perceived intrusiveness, and willingness to participate again in a similar survey.

The first four questions utilized scales with five levels, consisting of two negative options (e.g., very difficult, quite difficult), one neutral option (e.g., neither difficult nor easy), and two positive options (e.g., quite easy, very easy). For each of these questions, the proportion of positive answers, combining the two positive levels, is compared.

The question regarding willingness to participate again involved three response options (yes, no, and not sure). The proportion of affirmative answers between the two surveys is compared.

Data Quality

To assess differences in data quality between the groups, five commonly used indicators are employed. The full details of the variables used for each indicator can be found in Appendix 3. The indicators are:

1. *(Explicit) non-recall*: This indicator measures the proportion of respondents unable to recall requested information in a question, attributed to the effect of time and/or the lack of effort (Groves, 1989). I focus on explicit non-recall, where participants overtly declare their inability to provide the requested information. This evaluation spans across 22 different questions, including open-ended questions and questions with “Don’t know” and/or “Don’t remember” options (both considered as non-recall). Two of the open-ended questions for the conventional group were not asked to the in-the-moment group, as the same information was obtained through metered data.
2. *Invalid answer*: Invalid answers, which serve as an indicator of low data quality (Revilla & Ochoa, 2015), were identified through manual coding of responses to eight open-ended questions. A response is considered invalid if it fails to answer what was asked.
3. *Length of answers*: The mean number of characters in the answers to narrative open-ended questions, after discarding invalid answers, is sometimes used as a measure of data quality (Revilla & Ochoa, 2015). This indicator is calculated for three different open-ended questions. Two additional questions were excluded from the analysis due to a very limited number of responses (less than 20 per group).
4. *Straight-lining*: Straight-lining refers to selecting the same option in a set of consecutive questions sharing the same answer scale, even when it is not reasonable to expect identical responses (Green & Krosnick, 2001). This indicator is calculated for one set of four questions and another set of eight questions.
5. *Inconsistencies*: Inconsistencies are assessed by analyzing the proportion of answers to specific questions where participants do not adhere to the instructions or provide combinations of answers that do not logically align (DiLalla & Dollinger, 2006), considering three groups of cases:
   - Numerical answers out of bounds for four open-ended numerical questions. Inconsistencies are noted when participants provide answers outside the range of 0-100%.
   - Incoherent answers across three groups of related questions, where the answer to one question should logically align with the answer to another question (e.g., the number of applications without meeting requirements should be below the total number of applications).

- Selecting more than the maximum allowed in a multiple-answer question.

Certain potential indicators, such as survey duration, were discarded due to their unclear relationship with quality, especially in online surveys where respondents may keep the survey open but inactive while engaging in other activities. Moreover, technical limitations hindered the utilization of some indicators, such as assessing the external validity of answers by comparing them to the actual job description seen by participants. Future versions of the meter may address this limitation.

Differences in Substantive Results

The potential effect of the survey type on substantive answers is assessed by comparing results derived from six questions requesting objective information (e.g., the percentage of met requirements) and two requesting subjective information (e.g., the expected probability of being interviewed).

Additionally, as control measures, two substantive results completely unrelated to the event of interest (personality traits that should not present differences between groups) are explored. The full detail on which variables are used for the substantive results can also be found in Appendix 3.

Results

Participation (RQ1)

Table 1 provides a summary of participation levels in both surveys. The percentages in the table are calculated relative to the preceding category, as indicated by the indentation of these categories in the first column. For instance, the percentage of starts for the in-the-moment survey (88.2%) is derived from the number of invited panelists. Similarly, the percentage of breakoffs (1.3%) is based on the number of starts, and so forth.

The ratio of participants who initiated the survey over the total number of invited panelists is significantly higher for the in-the-moment group (88.2%) compared to the conventional group (62.5%). When accounting for the 283 panelists from the conventional group who attempted to start the survey but found it closed due to reaching the target sample size (“Survey closed” in the table, 13.6%), the overall figure increases to 76.1%, which is still significantly lower than that of the in-the-moment group. Similarly, the percentage of breakoffs is significantly lower in the in-the-moment group (1.3%) compared to the conventional one (5.0%).

Table 1 Participation in in-the-moment (ITM) and conventional (Conv) surveys

	ITM		Conv	
	n	%	*n*	%
Invited	356		2,080	
Non-starts	42	11.8	498	23.9
Starts	314	88.2	1,299	62.5
Breakoffs	4	1.3	65	5.0
Non-consent	5	1.6	58	4.5
Screened-out	107	34.1	975	75.1
Not searching in the last 48h / 6 months	30	28.0	791	81.1
Not confirming last search / –	24	22.4	-	-
Not applying to the detected job / any job	51	47.7	173	17.7
Other (e.g., bot detection)	2	1.9	11	1.1
Complete	177	56.4	201	15.5
Complete after 48h	21	6.7	-	-
Survey closed	-	-	283	13.6

It is worth noting that these differences can be attributed to both the survey type and the profile of participants in each group. Specifically, while all participants invited to the in-the-moment survey are metered panelists, only 467 out of the 2,080 who started the conventional survey (22.5%) fall into this category. Metered panelists, who may generally have a more positive attitude toward surveys (Revilla et al., 2021), could contribute to the higher level of participation and lower level of breakoff rates observed in the in-the-moment group.

To further explore these effects and disentangle the impact of the type of survey from the panelists' profile, logistic regression analyses are conducted with participation and breakoff as the dependent variables, and survey type as the main independent variable. The analyses also controlled for the variables detailed in the section "Comparisons Between Groups" (see Appendix 1 for a descriptive analyses per group).[7]

Given the strong correlation between survey type and metering status ($r = .75$), a multicollinearity analysis was conducted. The Variance Inflation Factor (VIF) values for these two variables indicate acceptable levels of multicollinearity (see SOM3). Therefore, the specification of the model using all seven aforementioned variables was retained.

[7] The type of device used to complete the survey was excluded from the participation analysis since this variable is only recorded once participants start the survey.

After controlling for these variables, the positive effect of the in-the-moment survey on participation remains significant ($p < .001$), but not on breakoff, where gender, age, and, specially, the number of past participations account for more explanatory power. However, these results lead us to reject the hypothesis that in-the-moment surveys have lower participation levels than conventional surveys (*H1*).

It is noteworthy also to discuss the variations in sample utilization between the two methods. To obtain 177 complete surveys for the in-the-moment group, a total of 315 panelists were invited, and among those who participated, 57.1% yielded valid participations (incidence rate). Conversely, for the conventional group, a significantly larger number of 2,080 panelists were invited, and only 16.3% of the participants ultimately provided valid responses.[8]

The difference in sample utilization can primarily be attributed to the need to ask participants in the conventional survey whether they applied for a job in the last six months, something that is known in advance for most of the in-the-moment survey participants. However, the improved sample utilization in the in-the-moment group comes at the cost of an extended fieldwork period (207 days vs. five days).

Survey Evaluations (RQ2)

Table 2 presents a comparison of survey evaluations made by participants in each group. Sample sizes are provided in Columns 2 and 3. Columns 4 to 7 display the proportions of positive answers for the in-the-moment group (ITM) and the conventional group (Conv), along with the difference (Diff = ITM-Conv) and the *p*-value resulting from a significance test. The last two columns present the impact of the in-the-moment group in a logistic regression model, incorporating the control variables described before.

Participants in the in-the-moment survey perceive the assigned task as significantly easier (+10.2 percentage points, pp) and more satisfactory (+11.6 pp) compared to the conventional survey. However, levels of trust in survey confidentiality and perceived intrusiveness are similar between the two surveys. This suggests that panelists do not perceive any additional risk in participating in the in-the-moment survey, despite the explicit mention of the invitation being triggered by activity detected using metered data. Moreover, the willingness to participate again is similarly high (94.4% and 93.5%).

[8] This incidence rate aligns with our initial estimations based on metered data, which further supports the notion that metered and non-metered panelists exhibit similar behavior, at least regarding online job applications.

Table 2 Survey evaluations by group

	N		Proportion %				Log. regression	
Question	ITM	Conv	ITM	Conv	Diff	*p*-value	Effect	*p*-value
Effort: easy	177	201	85.3	75.1	10.2*	.015	0.6	.121
Satisfaction: high	177	200	70.1	58.5	11.6*	.024	0.3	.345
Privacy: trust	177	200	74.4	70.0	4.4	.358	0.3	.474
Intrusiveness: low	177	201	50.3	52.2	-2.0	.757	-0.3	.431
Do it again: yes	177	200	94.4	93.5	0.9	.831	0.3	.582

Notes: Sample sizes for the in-the-moment (ITM) and conventional (Conv) groups. Proportion %: percentage of positive answers per group, difference in proportions (Diff) and significance (*p*-value). Log. regression: coefficient (Effect) and significance (*p*-value) of the in-the-moment group in a logistic regression controlling for gender, age, education level, metering status, number of participations (last three months) and device used to participate.

Once again, these differences appear to be a combined effect of the survey type and sample profile. The percentage of participants who rated their participation as easy was 74.5% for non-metered conventional participants, 76.8% for conventional metered panelists, and 85.3% for in-the-moment (metered) participants. Similarly, the percentages of participants who reported liking the participation experience in these three groups were 56.6%, 62.5%, and 70.1%, respectively. However, when conducting regression analyses controlling for being a metered panelist along with sociodemographic variables, the effect of the survey type is no longer statistically significant, which may be due to the limited statistical power resulting from splitting the sample into these groups.

In conclusion, despite the limitations posed by the smaller sample size, it can be inferred that in-the-moment surveys receive similar evaluations in terms of ease and satisfaction, compared to conventional surveys (support for *H2*).

Differences in Data Quality (RQ3)

The results of evaluating the 43 quality indicators described in Appendix 3 are presented in Table 3, following exactly the same structure as in Table 2.

Table 3 Quality indicators

	N		% of cases				Log. regression	
Non-recall indicators	ITM	Conv	ITM	Conv	Diff	*p*-value	Effect	*p*-value
Company name	177	201	16.4	25.9	-9.5*	.032	-0.6	.117
Job description	177	201	2.3	10.4	-8.2*	.001	-1.4*	.039
Salary	177	201	13.6	20.4	-6.8	.101	0.0	.952
Contract	177	201	17.5	25.4	-7.9	.080	0.1	.817
Experience	177	201	14.1	23.4	-9.3*	.026	-0.6	.173
Perks	177	201	14.7	20.9	-6.2	.140	-0.7	.058
% of met requirements	177	201	23.7	19.4	4.3	.318	0.2	.601
Specific not met req.	97	101	7.2	6.9	0.3	1	0.3	.735
% of fit	177	201	9.6	18.9	-9.3*	.013	-0.8	.075
Salary – Not fitting	134	133	7.5	3.8	3.7	.288	1.0	.346
Hours – Not fitting	134	133	1.5	0.8	0.7	1	18.1	.998
Flexibility – Not fitting	134	133	1.5	3	-1.5	.447	15.1	.993
Location – Not fitting	133	133	1.5	0.8	0.8	1	-0.6	.643
Tasks – Not fitting	134	133	0.0	0.0	0.0	1	0.0	1
Manager – Not fitting	134	133	2.2	2.3	0.0	1	15.6	.993
Company – Not fitting	134	133	2.2	3.8	-1.5	.500	15.4	.993
Contract – Not fitting	134	133	6.0	4.5	1.5	.785	16.6	.992
Applications in last 6m.	177	201	32.2	34.3	-2.1	.743	-0.2	.538
Apps. without req. in last 6m.	176	201	37.5	44.8	-7.3	.173	-0.5	.111
Apps. without fit in last 6m.	176	201	36.9	40.8	-3.9	.460	-0.6	.069
Job search website	177	201	-	2.5	-2.5*	-	-*	-
Last application date	177	201	-	49.8	-49.8*	-	-*	-

	N		% of cases				Log. regression	
Invalid answers	ITM	Conv	ITM	Conv	Diff	*p*-value	Effect	*p*-value
Company name	177	196	0.6	1.5	-1.0	.625	14.9	.995
Job description	177	195	1.1	3.6	-2.5	.178	-1.2	.234
Salary	51	54	0	13.0	-13.0*	.013	-19.1	.994
Contract	89	106	0	1.9	-1.9	.501	-0.3	1
Experience	106	103	0	2.9	-2.9	.118	0.8	1
Perks	17	12	17.6	16.7	1.0	1	95.4	1
Why applying without req.	77	91	5.3	11.0	-5.8	.263	-1.0	.220
Why applying without fit	134	132	16.7	21.4	-5.6	.350	-0.9*	.027

Table 3 (continued)

	N		Num. of characters.				Lin. regression	
Length of answers	ITM	Conv	ITM	Conv	Diff	*p*-value	Effect	*p*-value
Job description	175	188	41.3	28.0	13.3*	<.001	11.7*	.012
Why applying without req.	74	81	71.4	52.0	19.5*	.004	21.5*	.047
Why applying without fit	112	103	60.8	54.6	6.1	.325	11.2	.263
	N		% of cases				Log. regression	
Straight-lining	ITM	Conv	ITM	Conv	Diff	*p*-value	Effect	*p*-value
Job details (4 items)	177	201	10.7	15.4	-4.7	.224	-0.7	.103
Fit of features (8 items)	134	133	9.0	9.8	-0.8	.837	-0.4	.490
	N		% of cases				Log. regression	
Inconsistencies	ITM	Conv	ITM	Conv	Diff	*p*-value	Effect	*p*-value
% of req. out of limits	135	162	0	0	-	-	-	-
% of fit. out of limits	160	163	0	0	-	-	-	-
Probability of interview out of limits	176	200	0	0	-	-	-	-
Probability of hiring out of limits	176	200	0	0	-	-	-	-
% of met req. < 100 + meeting all req.	135	162	14.1	6.2	7.9*	.030	0.0	.965
> 3 options in motivation question	177	201	7.3	6.0	1.4	.680	-0.1	.848
Apps. without met req. > total apps.	104	105	1.9	1.9	0.0	1	-0.7	.571
Apps. without perfect fit > total apps.	103	112	3.9	1.8	2.1	.430	-0.8	.377

Notes: Sample sizes for the in-the-moment (ITM) and conventional (Conv) groups. Proportion %: percentage of positive answers per group, difference in proportions (Diff) and significance (*p*-value). Log./Lin. regression: coefficient (Effect) and significance (*p*-value) of the in-the-moment group in a regression (linear for means, logistic for proportions) controlling for gender, age, education level, metering status, number of participations (last three months) and device used to participate.

Out of the 22 non-recall indicators, 14 show better results for the in-the-moment group, indicating lower non-recall (negative effects). However, only six of these effects are significant. When adding the control variables, the number of favorable results for the in-the-moment groups decreases to 11, with three of them significant.

The largest favorable effect is observed for one of the two variables recorded using metered data instead of relying on a question for the in-the-moment group, with a decrease of 49.8 pp. Excluding this variable, the observed effects range from –9.5 pp to +4.3 pp. The median effect across all 22 variables is –2.3 pp, while the mean effect is –5.2 pp.

Among the six questions that exhibit higher levels of non-recall for the in-the-moment group, four belong to the same set of eight questions that asked participants whether each job feature matched what they were looking for, with an explicit option of "I don't remember." Interestingly, one of the other two questions showing a higher level of non-recall for the in-the-moment group is the one asking for the percentage of met requirements. This question included an input box for participants to write their answer and two radio buttons to indicate "I don't know" and "I don't remember" (see questionnaire in SOM2). In Table 3, both options are considered as non-recall. However, if only the option "I don't remember" is considered as non-recall, the in-the-moment group exhibits a lower level of non-recall (8.8% vs. 12.0%; see SOM4).

In terms of the percentage of invalid answers, seven out of eight indicators studied show lower levels for the in-the-moment group, but the effects are generally moderate, only one of them being significant. When controlling for the usual variables, the significant effect remains but two effects are reversed.

Regarding the length of answers to open narrative questions, all three questions studied show longer answers for the in-the-moment group, with relative effects ranging from +11.4% to +47.5%. Two of these effects are significant, also when controlling for the usual covariates.

The two straight-lining indicators favor the in-the-moment survey, although the effect is not statistically significant. However, this effect is only substantial (–4.7%) in the case of the set of four questions.

Finally, out of the eight consistency indicators, the four related to exceeding the limits of numerical questions do not show a single case in any of the two groups, while the remaining four exhibit very small and non-significant differences, with slightly worse results for the in-the-moment group.

In conclusion, these results do not clearly support the beneficial effects of in-the-moment surveying on data quality (*H3*), except for longer answers to open-ended questions. While statistically significant effects are lacking, 25 out of 44 indicators favor in-the-moment surveys, 12 favor conventional surveys, and 7 are neutral. This favorable trend for in-the-moment surveys needs further validation with larger samples. The positive impact on response length contrasts with the weaker-than-expected effect on non-recall, possibly due to participants' overconfidence in their memory accuracy.

To assess this potential overconfidence, participants in the conventional survey were asked to what extent they were confident that the information they reported actually corresponded to their last job application, as requested. Given that memories tend to fade over time, particularly in the initial stages, one might expect participants to report lower confidence levels for job applications made further in the past. However, as depicted in Figure 1, reported confidence levels remain relatively constant and consistently high across the 0 to 160-day range. This contradicts existing knowledge about memory decay, suggesting thus that overconfidence is occurring.

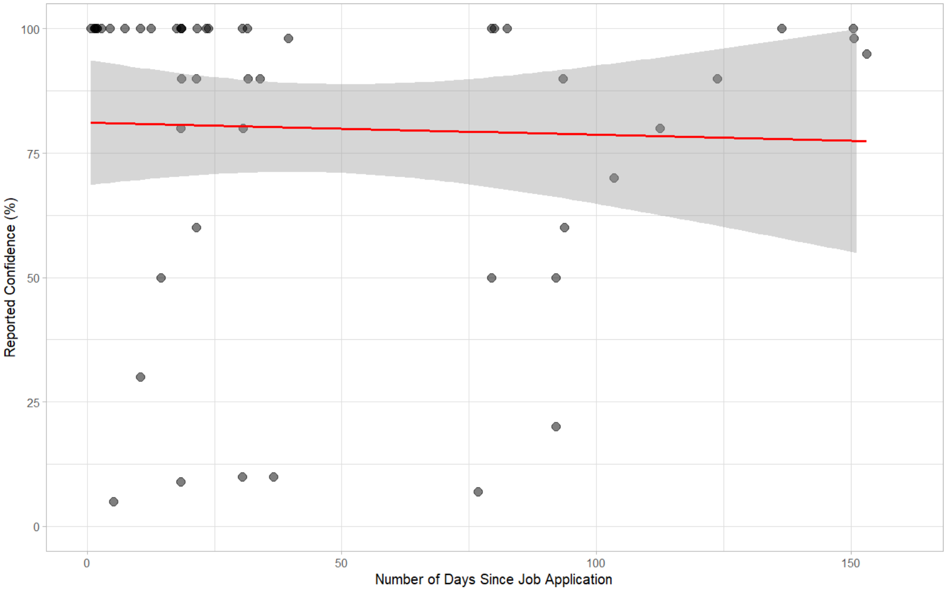


Figure 1 Confidence in reporting last job application over time by participants in the conventional survey

Differences in Substantive Results (RQ4)

Table 4 presents the substantive answers of interest for the two groups, maintaining the same structure as previous tables.

Despite the limited sample size, which particularly affects some questions asked only to part of the respondents, several significant disparities emerge in the substantive findings depending on the survey type. The estimated probabilities of being interviewed and hired, both subjective measures, are 7.9 pp and 8.7 pp lower, respectively, for the in-the-moment group, both significant. After adjusting for the usual covariates, the effects become 7.8 pp and 10.9 pp, with only thelatter remaining significant. As the questions in the conventional survey explicitly requested participants to report information “at the time of applying,” the observed differences can be attributed to distorted recall, meaning recall errors or alterations introduced in the answers without participants being aware.

Table 4 Substantive differences

	N		Value (%)				Regression	
	ITM	Conv	ITM	Conv	Diff	*p*-value	Effect	*p*-value
Met requirements	135	162	80.1	84.0	-3.9	.082	-4.6	.162
Non-compliants	135	162	71.9	62.3	9.6	.108	0.4	.297
Fit of features	160	163	72.3	76.9	-4.6*	.046	-3.4	.340
Non-fitters	160	163	83.8	81.6	2.2	.660	0.3	.568
% apps. without met req	88	92	52.8	46.3	6.5	.242	17.5*	.046
% apps. without perfect fit	88	92	48.9	48.4	0.5	.933	-0.9	.915
Prob. of interview	176	200	47.7	55.6	-7.9*	.006	-7.8	.074
Prob. of hiring	176	200	39.6	48.3	-8.7*	.003	-10.9*	.012
Conformity (control)	177	201	2.7	2.7	0.0	.838	-0.1	.133
Efficacy (control)	176	201	3.9	3.9	0.0	.680	0.0	.836

Notes: Sample sizes for the in-the-moment (ITM) and conventional (Conv) groups. Value: value of the substantive result (a proportion or an average numeric value) per group, difference of values (Diff) and significance (*p*-value). Regression: coefficient (Effect) and significance (*p*-value) of the in-the-moment group in a regression (linear for means, logistic for proportions) controlling for gender, age, education level, metering status, number of participations (last three months) and device used to participate.

The percentage of participants who admitted to applying without meeting all the job requirements, a key aspect that motivated this research, especially in terms of potential gender differences, differs between the two survey types: 71.9% for the in-the-moment survey versus 62.3% for the conventional survey (on average 23.6 days after the event of interest). However, this effect vanishes when controlling for the usual covariates.

In contrast, the two personality traits included as controls yield almost identical results in both groups (with and without controls), aligning with our expectations. Therefore, these findings support the hypothesis (*H4*) that the time elapsed since the occurrence of the event of interest can impact the substantive answers provided by survey participants.

Discussion

Summary of Main Results

The results from this study reveal that members from metered panels readily embrace in-the-moment surveys triggered by online events, displaying heightened levels of participation in comparison to an equivalent conventional survey (RQ1). Furthermore, participants evaluated this new method similarly to a conventional survey regarding the required effort and overall satisfaction, and did not perceive an increased risk concerning privacy or intrusiveness associated with this survey format (RQ2).

The responses provided by participants also revealed moderate positive impacts on data quality (RQ3). While some of these effects are substantial and statistically significant, such as the increase in the length of responses to open-ended questions, other observed effects are not conclusive. Particularly, the positive effect on explicit non-recall is weaker than expected, possibly due to participants' overconfidence in the accuracy of their memories. Moreover, significant disparities in substantive results emerged based on the data collection method employed, further supporting the notion that recall errors influence the gathered data more extensively than participants are aware (RQ4).

Limitations

There are several limitations affecting this study. Firstly, it relies on a sample from a single opt-in online panel (Netquest) in a single country (Spain). Different panels and countries, as well as other sampling methods (e.g., probability samples) may produce different results, underscoring the need for caution when attempting to generalize findings.

Secondly, the technical solutions chosen for the study might have influenced the outcomes. Utilizing alternative platforms could lead to different results. For instance, the way push notifications are presented to mobile panelists can significantly impact their participation behavior (e.g., shorter/longer delays). Similarly, the impossibility of detecting certain online events (i.e., events within mobile apps) may have affected some results.

Thirdly, the sample size for this research was constrained by the availability of metered panelists, preventing certain analysis (e.g., the effect of elapsed time since the job application on discrepancies in substantive results between both surveys; see SOM4) potentially limiting the ability to detect significant effects for certain observed differences. To validate the findings of this paper, further investigations with larger sample sizes are necessary, maybe focusing on high-prevalence events (e.g., online purchases) to address the current limitations stemming from the constrained size of existing metered panels.

Practical Implications

Surveying people in the moment using metered panels is a promising methodology, especially suited for researching repetitive, low-emotional, and hard-to-distinguish events. Despite the limited sample size in this study, the results indicate that conducting research close to the event of interest leads to slightly better data quality and reveals clear differences in substantive results. Such substantive differences suggest a significant reduction of distorted recall, wherein people inadvertently fail to report accurate information.

Nevertheless, this study has also highlighted several inconveniences associated with this type of surveys that researchers must carefully consider before deciding to use it. In-the-moment surveys require the use of specific technology, a complex set-up that involves the identification of specific URLs related to the events of interest, extended fieldwork times, and regular supervision, among other challenges (see Reflective Appendix). Additionally, the limited size of existing metered panels poses limitations on obtaining large samples for specific target populations.

This combination of pros and cons suggests that in-the-moment surveys are well-suited for high prevalence activities that occur frequently over time, but they may not be ideal for activities with an excessive number of repetitions in a short period. The latter scenario could lead to participants misidentifying the specific event of interest in the survey. Examples of suitable activities may include post-purchase satisfaction surveys for online purchases and opinion polls targeting audiences during live streaming media consumption.

Finally, it is essential to emphasize that in-the-moment surveys are not designed to replace conventional surveys but rather to serve as a valuable additional methodology in very specific cases. Their unique strengths make them particularly useful for certain research scenarios, at the cost of extended fieldwork times and increased complexity.

Data Availability and Supplementary Online Material (SOM)

The anonymized dataset, together with all the scripts used for the analyses and the supplementary online material of this paper can be found at: https://osf.io/67sgz

References

Allurwar, N., Nawale, B., & Patel, S. (2016). Beacon for proximity target marketing. *International Journal of Engineering and Computer Science, 15*(5), 16359–16364. https://ijecs.in/index.php/ijecs/article/view/1125

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., Groves, R. M., Kennedy, C., Krosnick, J., Lavrakas, P. J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R. K., & Zahs, D. (2010). Research synthesis: AAPOR report on online panels. *Public Opinion Quarterly, 74*(4), 711–781. https://doi.org/10.1093/poq/nfq048

Boltho, A., & Glyn, A. (1995). Can macroeconomic policies raise employment? *International Labour Review, 134*(4–5), 451–470.

Bosch, O. J., & Revilla, M. (2022). When survey science met web tracking: Presenting an error framework for metered data. *Journal of the Royal Statistical Society Series A: Statistics in Society, 185*(Supplement_2), S408–S436. https://doi.org/10.1111/rssa.12956

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

DiLalla, D. L., & Dollinger, S. J. (2006). Cleaning up data and running preliminary analyses. In F. T. L. Leong & J. T. Austin (Eds.), *The psychology research handbook: A guide for graduate students and research assistants* (2 ed., pp. 241–254). Sage Publications. http://doi.org/10.4135/9781412976626.n16

Dillahunt, T. R., Israni, A., Lu, A. J., Cai, M., & Hsiao, J. C. Y. (2021). Examining the use of online platforms for employment: A survey of US job seekers. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems,* Article 562, 1–23. https://doi.org/10.1145/3411764.3445350

Ezzy, D. (1993). Unemployment and mental health: A critical review. *Social Science & Medicine, 37*(1), 41–52. https://doi.org/10.1016/0277-9536(93)90316-V

Faiz A. (2020). Impact of online job search and job reviews on job decision. *Proceedings of the 13th International Conference on Web Search and Data Mining,* 909–910. https://doi.org/10.1145/3336191.3372184

Fluchtmann, J., Glenny, A., Harmon, N. A., & Maibom, J. (2022). *The gender application gap: Do men and women apply for the same jobs?* (IZA Discussion Paper No. 14906). SSRN. https://ssrn.com/abstract=4114410

Green, M. C., & Krosnick, J. A. (2001). Comparing telephone and face-to-face interviewing in terms of data quality: The 1982 national election studies method comparison project. In M. L. Cynamon & R. A. Kulka (Eds.), *Health survey research methods* (pp. 115–121). U.S. Department of Health and Human Services.

Groves, R. M. (1989). *Survey errors and survey costs*. John Wiley & Sons.

Haas, G.-C., Trappmann, M., Keusch, F., Bähr, S., & Kreuter, F. (2020). Using geofences to collect survey data: Lessons learned from the IAB-SMART study [Special issue: Advancements in online and mobile survey methods]. *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2020-00023

Hardeman, W., Houghton, J., Lane, K., Jones, A., & Naughton, F. (2019). A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. *International Journal of Behavioral Nutrition and Physical Activity, 16,* Article 31. https://doi.org/10.1186/s12966-019-0792-7

Jobe, J. B., Tourangeau, R., & Smith, A.F. (1993). Contributions of survey research to the understanding of memory. *Applied Cognitive Psychology, 7*(7), 567–584. https://doi.org/10.1002/acp.2350070703

Kahneman, D., & Riis, J. (2005). Living, and thinking about it: Two perspectives on life. In F. A. Huppert, N. Baylis, & B. Keverne (Eds.), *The science of well-being* (pp. 285–304). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198567523.003.0011

Karaoglu, G., Hargittai, E., & Nguyen, M. H. (2022). Inequality in online job searching in the age of social media. *Information, Communication & Society, 25*(12), 1826–1844. https://doi.org/10.1080/1369118X.2021.1897150

Keusch, F., & Conrad, F. G. (2022). Using smartphones to capture and combine self-reports and passively measured behavior in social research. *Journal of Survey Statistics and Methodology, 10*(4), 863–885. https://doi.org/10.1093/jssam/smab035

Kuhn, P., & Mansour, H. (2014). Is internet job search still ineffective? *The Economic Journal, 124*(581), 1213–1233. https://doi.org/10.1111/ecoj.12119

Lambert, A. D., & Miller, A. L. (2015). Living with smartphones: Does completion device affect survey responses? *Research in Higher Education, 56*, 166–177. https://doi.org/10.1007/s11162-014-9354-7

Lamas, C. (2005). *The value of coincidental surveys to monitor the validity of tv meter panel measurements* [Paper presentation]. EMRO Conference, Kwa Maritane, South Africa. https://www.aimc.es/a1mc-c0nt3nt/uploads/2010/10/emro2005.pdf

Link, M. W., Murphy, J., Schober, M. F., Buskirk, T. D., Hunter Childs, J., & Langer Tesfaye, C. (2014). Mobile technologies for conducting, augmenting and potentially replacing surveys: Executive summary of the AAPOR task force on emerging technologies in public opinion research. *Public Opinion Quarterly, 78*(4), 779–787. https://doi.org/10.1093/poq/nfu054

Mansouri, B., Zahedi, M., Campos, R., & Farhoodi, M. (2018). Online job search: Study of users' search behavior using search engine query. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. Association for Computing Machinery, 1185–1188. https://doi.org/10.1145/3209978.3210125

Ochoa, C., Cortina, C., & González, M. J. (2023, June, 27–29). *Gender differences in job application requirements: Do women demand more of themselves than men? A survey-based study of job-seeking behavior in Spain* [Conference presentation]. 29th International Conference of Europeanists, Reykiavik, Iceland. https://www.upf.edu/documents/244683118/246905697/CEC_2023+%281%29.pptx

Ochoa, C., & Porcar, J. M. (2018). Modeling the effect of quota sampling on online fieldwork efficiency: An analysis of the connection between uncertainty and sample usage. *International Journal of Market Research, 60*(5), 484–501. https://doi.org/10.1177/1470785318779545

Ochoa, C., & Revilla, M. (2022). Acceptance and coverage of fast invitation methods to in-the-moment surveys. *International Journal of Market Research, 64*(5), 565–574. https://doi.org/10.1177/14707853221085204

Ochoa, C., & Revilla, M. (2023). Willingness to participate in in-the-moment surveys triggered by online behaviors. *Behavior Research Methods, 55,* 127–1291. https://doi.org/10.3758/s13428-022-01872-x

Revilla, M. (2017). Analyzing survey characteristics, participation, and evaluation across 186 surveys in an online opt-in panel in Spain. *methods, data, analyses, 11*(2), 135–162. https://doi.org/10.12758/mda.2017.02

Revilla, M. (2022). How to enhance web survey data using metered, geolocation, visual and voice data? *Survey Research Methods, 16*(1). 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. (2015). What are the links in a web survey among response time, quality, and auto-evaluation of the efforts done? *Social Science Computer Review, 33*(1), 97–114. https://doi.org/10.1177/0894439314531214

Revilla, M., Ochoa, C., Iglesias, P., & Antón, D. (2022). *WebdataNow: A tool to send in-the-moment surveys triggered by passive data*. OSF. http://doi.org/10.17605/OSF.IO/G3MSC

Revilla, M., Paura, E., & Ochoa, C. (2021). Use of a research app in an online opt-in panel: The Netquest case. *Methodological Innovations, 14*(1). https://doi.org/10.1177/2059799120985373

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103*(4), 734–760. https://doi.org/10.1037/0033-295X.103.4.734

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4*, 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Smith, A. (2015). *Searching for work in the digital era*. Pew Research Center. https://www.pewresearch.org/internet/2015/11/19/searching-for-work-in-the-digital-era

Tourangeau, R. (2000). Remembering what happened: Memory errors and survey reports. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 29–47). Lawrence Erlbaum Associates Publishers.

van Berkel, N., Ferreira, D., & Kostakos, V. (2017). The experience sampling method on mobile devices. *ACM Computing Surveys, 50*(6), Article 93, 1–40. https://doi.org/10.1145/3123988

Walker, W. R., & Skowronski, J. J. (2009). The Fading affect bias: But what the hell is it for? *Applied Cognitive Psychology, 23*(8), 1122–1136. https://doi.org/10.1002/acp.1614

Appendix 1. Distribution of Control Variables

Table A1 presents the distribution of the control variables used in the regression analyses for each group.

Table A1 Distribution of control variables in in-the-moment (ITM) and conventional (Conv) surveys

	ITM	Conv
Age (mean)*	41.7	38.6
Gender		
Male	44.6%	38.8%
Female	55.4%	61.2%
Education		
Low-Mid	55.4%	55.2%
High	44.6%	44.8%
Survey participations – last 3 months (mean)*	33.1	27.2
Metered*	100.0%	27.9%
Survey device*		
PC	53.1%	28.4%
Mobile (smartphone and tablet)	46.9%	71.6%

Notes: *indicates a significant effect (5% level) between ITM and Conv.

Appendix 2. List of Job Search Websites

Table A2 lists the job search websites that triggered in-the-moment surveys and the regular expressions used to identify the URLs shown to applicants.

Table A2 List of job search websites

Website	Regular expression identifying a job application
infojobs.net/	infojobs\.net\/candidate\/application\/apply
ticjob.es/	ticjob\.es\/\S*submit-application ticjob.es\/esp\/\S*?status=applied ticjob.es\/esp\/\S*?applied
es.indeed.com/	indeed\.com\/\S*\/post-apply es\.indeed\.com\/pagead\/clk
es.jooble.org/	es.jooble.org\/away\/
infoempleo.com/	infoempleo\.com\/killerquestion\/ infoempleo\.com\/inscription
Tecnoempleo.com	tecnoempleo\.com\/\S*enviar\.php
monster.com	www\.monster\.es\/\S*apply
Randstad.es	randstad\.es\/\S*\/apply\/ randstad\.es\/candidatos\/ofertas-empleo\/\S*\/gracias
Adecco	4dec\.co\/\S*applyFinishOK
Trabajos.com	trabajos\.com\/\S*\/oferta-respondida
Jobatus.es	jobatus\.es\/oferta-trabajo\/\S*?jc=True

Appendix 3. List of Indicators

The following tables summarize the quality indicators, substantive measures used to assess survey differences, and the variables for estimating these indicators.

Table A3.1 Survey evaluation questions

Variables	Question wording and recoded categories
Effort	To what extent did you find it easy or difficult to respond to this survey? ☐ Very easy (recoded as “easy”) ☐ Quite easy (recoded as “easy”) ☐ Neither easy nor difficult ☐ Quite difficult ☐ Very difficult
Satisfaction	To what extent did you like or dislike responding to this survey? ☐ I liked it a lot (recoded as “high”) ☐ I liked it quite a bit (recoded as “high”) ☐ I neither liked nor disliked it ☐ I disliked it quite a bit ☐ I disliked it a lot
Privacy	To what extent do you trust or distrust that your responses to this survey are truly anonymous? ☐ I trust completely (recoded as “trust”) ☐ I trust quite a bit (recoded as “trust”) ☐ I neither trust nor distrust ☐ I distrust quite a bit ☐ I distrust completely
Intrusiveness	To what extent did you find this survey intrusive or not? ☐ Totally intrusive (recoded as “intrusive”) ☐ Very intrusive (recoded as “intrusive”) ☐ Moderately intrusive ☐ Slightly intrusive ☐ Not at all intrusive
Do it again	Would you participate in a survey like this again? ☐ Yes ☐ No ☐ Not sure [if gender = female]

Table A3.2 Quality indicators – Non-recall

Variables	Type	Calculated as …
☐ Company name ☐ Job description	Open-ended	% of answers declaring not remembering or giving non-specific answers
Information in the job description: ☐ Salary ☐ Type of contract ☐ Required experience ☐ Offered perks	Single-response	% of "don't remember" answers
☐ Percentage of met requirements ☐ Percentage of job features that did not fit expectations In the last 6 months: ☐ Number of applications ☐ Number of applications without meeting requirements ☐ Number of applications made without fitting expectations	Numerical open-ended	% of "don't know/ remember" answers
☐ Specific not-met requirements	Multiple-response	% of "don't know/remember" answers
List of non-fitting job features: ☐ Salary ☐ Hours ☐ Flexibility ☐ Location ☐ Tasks ☐ Manager ☐ Company ☐ Contract	Set of single-response questions	% of "don't remember" answers
☐ Job search website ☐ Last application date	In-the-moment: Metered data Conventional: Open-ended	In-the-moment: 100% informed Conventional: % of "don't know" answers

Table A3.3 Quality indicators – Invalid answers

Variables	Type	Calculated as …
□ Company name □ Job description Information in the job description: □ Salary (specify which) □ Type of contract (specify which) □ Required experience (specify which) □ Offered perks (specify which) Reasons for □ Applying without meeting requirements □ Applying without a perfect fit	Open-ended	% of invalid answers (not answering what was asked)

Table A3.4 Quality indicators – Length of answers

Variables	Type	Calculated as …
Reasons for □ Applying without meeting requirements □ Applying without a perfect fit	Open-ended	Mean number of characters

Table A3.5 Quality indicators – Straight-lining

Variables	Type	Calculated as …
□ Job details (4 questions sharing the same three answer categories) □ Fit of features (8 questions sharing the same four answer categories)	Set of single-response	% of respondents selecting the same answer option in all the questions withing the set

Table A3.6 Quality indicators – Inconsistencies

Variables	Type	Calculated as …
☐ Percentage of met requirements ☐ Percentage of job features that did not fit expectations ☐ Probability of being interviewed ☐ Probability of being hired	Numerical open-ended	% of answers outside the range of 0-100%
☐ Percentage of met requirements (< 100%) & requirements not met (= none) In the last 6 months: ☐ Number of applications without meeting requirements < number of applications ☐ Number of applications without a perfect fit < number of applications	Numerical open-ended (except requirements not met, that is multiple-response)	% of combined answers

Table A3.7 Substantive indicators

Variables	Type	Calculated as …
☐ Percentage of met requirements ☐ Proportion of non-compliant participants (applying without meeting all requirements) ☐ Percentage of job features that did not fit applicant's expectations ☐ Proportion of non-fitting participants (applying without a perfect fit) ☐ Probability of being interviewed ☐ Probability of being hired In the last six months ☐ Proportion of non-compliant participants (applying without meeting all requirements) ☐ Proportion of non-fitting participants (applying without a perfect fit)	Numerical open-ended	Differences in means or proportions
Control variables: ☐ Conformity (average score of 11 questions using 5-point scales) ☐ Efficacy (average score of eight questions using 5-point scales)	Sets of 5-point scale questions	Differences in means

Reflective Appendix

This appendix examines the methodological challenges during the experiment design and setup, the unforeseen issues encountered during setup and data collection, and the strategies used to address them.

Foreseen Challenges

In this project, we anticipated several methodological and practical challenges prior to fielding, some of which required adaptations to our original plan. These challenges primarily included (1) the limited representativeness of the metered panel, (2) difficulties in detecting job applications on certain webpages that do not provide unique URLs for such events, (3) the inability to detect in-app job applications on iOS or Android operating systems, (4) challenges in customizing in-the-moment surveys with specific job offer details, (5) difficulties in comparing survey responses with actual job data, and (6) the need to adapt questionnaires for in-the-moment administration by adding screening questions to protect the private information of non-panelists who might use the panelists' metered device to apply for a job.

These limitations were acknowledged and addressed in the main paper. Since these challenges were effectively managed using established strategies, this appendix focuses on the unforeseen challenges encountered during the project and the measures implemented to address them.

Unforeseen Challenges

Setup

The setup phase revealed new limitations of both the software and the method itself:

1. *Non-identifiable URLs:* Some websites did not display specific URLs when applying for a job, making these events undistinguishable from others. Additionally, some websites redirected to employers' sites without showing an identifiable URL for the event of interest. Consequently, four websites (linkedin.com/jobs, jobtoday.com, insertia.net, primerempleo.com) had to be excluded, reducing the ability to detect job applications. In other cases, related events (e.g., initiating the job application process) were used to trigger the survey rather than the actual job application event. In these cases, participants who did not progress to the event of interest were discarded in the questionnaire.

2. *Triggering URLs may change over time:* Websites deploying new versions resulted in changes to triggering URLs, necessitating a monthly repetition of the URL identification process that had not been initially planned.

3. *Job offer identification:* URLs displayed during job applications often lost reference to the job offer, preventing the planned prepopulation of surveys with specific job details such as company names, although job site and time could still be used. This also hindered the direct comparison between survey responses and job offer details that was initially planned, limiting the ability to assess the validity of the survey answers.

A contingency plan to record all web content viewed by participants was considered but not implemented due to the high sensitivity of the collected data and the need for additional approvals. This approach may be explored in a follow-up study.

Fieldwork Execution

In-the-moment fieldwork execution was expected to be much slower than for conventional surveys and affected by the seasonality of the events of interest. The reduced ability to detect applications, as discussed in the “Foreseen challenges” section, combined with the decrease in job applications during July and August due to the vacation period in Spain, required an extension of fieldwork into September to compensate.

Practical Recommendations and Future Developments

In addition to the study’s conclusions, researchers working with in-the-moment surveys and similar data types should consider the following recommendations:

1. Dedicate resources to the technology: In this project, enhancing the integration between in-the-moment surveys and metered data allowed us to assess the actual effect of time on participant’s responses. In general, effective use of new data types requires specialized technology or careful revisions to existing technologies.

2. Address metered data errors: Researchers should recognize and address errors in metered data (often overlooked), such as participants using non-metered devices or the non-detection of app events, as these issues can impact new methods built on such data and affect feasibility. The Total Error framework for digital traces (TEM) by Bosch and Revilla (2022) provides a comprehensive description of these errors.

3. Embrace technology and internet knowledge: Researchers should have a solid understanding of web and app technologies, as well as internet pro-

tocols. This knowledge is essential for making well-informed decisions and overcoming unexpected challenges during the project, such as realizing that websites with non-identifiable URLs had to be discarded, as was the case in this project.

4. Evaluate pros and cons: Assess when in-the-moment surveys are beneficial. This project illustrates the challenges faced: while they can provide better data than conventional surveys, in-the-moment surveys are time-consuming, require significant support (especially technological), and often result in smaller sample sizes. Therefore, careful feasibility assessment is crucial.

Finally, future research using metered data could benefit from two improvements not available during this project. First, the ability to detect events within apps, which has been recently added to the current version of the meter used. Second, increased coverage of panelists sharing multiple devices. These improvements should reduce false negatives, expanding the sample and/or shortening fieldwork times.

Improved Ensemble Predictive Modeling Techniques for Linked Social Media and Survey Data Sets Subject to Mismatch Error

Brady T. West[1], Martin Slawski[2] &
Emanuel Ben-David[3]

[1] *Institute for Social Research, University of Michigan-Ann Arbor*
[2] *Department of Statistics, University of Virginia*
[3] *U.S. Census Bureau*

Abstract

Modern predictive modeling tools, such as random forests (and related ensemble methods), have become almost ubiquitous in research applications involving innovative combinations of survey methodology and data science. However, an important potential flaw in the widespread application of these methods has not received sufficient research attention to date. Researchers at the junction of computer and survey science frequently leverage linked data sets to study relationships between variables, where the techniques used to link two (or more) data sets may be probabilistic and non-deterministic in nature. If frequent mismatch errors occur when linking two (or more) data sets, the commonly desired outputs of predictive modeling tools describing relationships between variables in the linked data sets (e.g., variable importance, confusion matrices, RMSE, etc.) may be negatively affected, and the true predictive performance of these tools may not be realized. We demonstrate a new methodology based on mixture modeling that is designed to adjust modern predictive modeling tools for the presence of mismatch errors in a linked data set. We evaluate the performance of this new methodology in an application involving the use of observed Twitter/X activity measures and predicted socio-demographic features of Twitter/X users to accurately predict linked measures of political ideology that were collected in a designed survey, where respondents were asked for consent to link any Twitter/X activity data to their survey responses (exactly, based on Twitter/X handles). We find that the new methodology, which we have implemented in R, is able to largely recover results that would have been seen prior to the introduction of mismatch errors in the linked data set.

Keywords: modern predictive modeling, ensemble methods, record linkage, mismatch error, mixture modeling, linked survey and social media data

In recent years, social media platforms such as Instagram and Twitter/X have provided social scientists with a wealth of user-content data (Agarwal et al., 2011; Bello-Orgaz et al., 2016; Ghani et al., 2019; McCormick et al., 2017). These data are often collected from multiple sources and then combined by probabilistic record linkage; for example, a research team might link two social media data sets, or link one social media data set to survey data (Al Baghal et al., 2021; Conrad et al., 2021; Eady et al., 2019; Karlsen & Enjolras, 2016). Researchers analyzing these linked data sets often apply advanced machine learning techniques, such as random forests, boosting (and related ensemble methods), neural networks, etc., whether the objective of the research project is accurate prediction of categorical survey outcomes (e.g., indicators of survey cooperation) or regression-based prediction of continuous outcomes (e.g., Gautam & Yadav, 2014; Liu & Singh, 2021; Wan & Gao, 2015).

There is, however, a potential pitfall in the widespread application of these modern predictive modeling techniques to linked data sets that needs more research attention. Although linking these types of new data sources provides the required information for novel studies of the relationships between variables, errors in the record linkage process may distort the true relationships between variables that are brought together from different data sources due to *mismatch errors* and *missed-match errors*. Missed-match errors refer to the inability to link a record in one data source to a matching record in a second data source, ultimately preventing that record from being included in an analysis of the relationships between variables from the two data sources. This type of error can lead to a form of selection bias in estimates of relationships, in a setting where the records with missed matches are unique in terms of the relationship of interest (Little & Rubin, 2019). In the setting of linking social media data with survey data, this type of error can arise when survey respondents do not consent to researchers linking their survey data with the information extracted from a Twitter handle or other identifiers (e.g., full names) used for social media accounts (e.g., Al Baghal et al., 2020). In this paper, we do not consider the problem of *missed-match errors,* but we suggest future directions for research in this area in the Discussion.

Mismatch errors, which are the primary focus of the current study, arise when records from different data sources are incorrectly matched (see Figure 1). Several prior studies have demonstrated the attenuating effects of mismatch errors on estimates of relationships in classical parametric regression

Direct correspondence to
Brady T. West, Survey Research Center of the Institute for Social Research,
University of Michigan-Ann Arbor, USA.
E-Mail: bwest@umich.edu

modeling settings, and proposed approaches for correcting this attenuation (Dalzell & Reiter, 2019; Han & Lahiri, 2019; Lahiri & Larsen, 2005; Neter et al., 1965; Scheuren & Winkler, 1997, 1993; Slawski et al., 2021; Steorts et al., 2018; Tancredi & Liseo 2015). In the setting of linking social media data with survey data, obtaining consent from respondents to link their survey responses with the social media content that they generate is required (Stier et al., 2020). In this setting, mismatch errors may arise when the names provided by the consenting survey respondents do not match with the names used for social media accounts, the full names provided do not uniquely identify individuals, when social media platform handles corresponding to user accounts are provided with typos that prevent exact matching, or when consenting respondents change their platform handles over time (Beuthner et al., 2021; Stier et al., 2020).

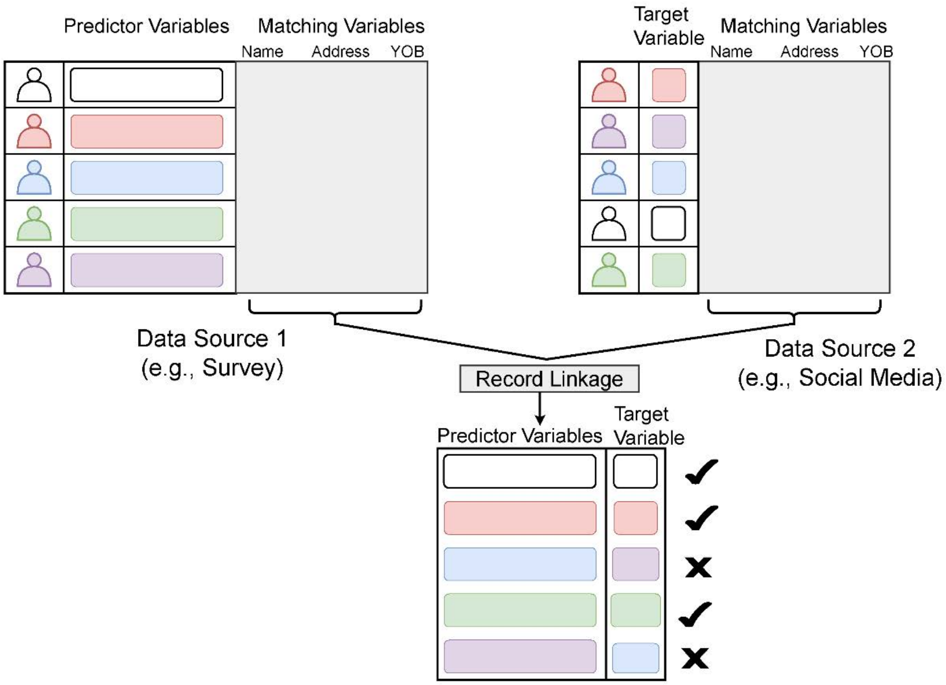


Figure 1 A visual overview of the mismatch error problem. Record linkage produces a linked file from two data sources containing predictor variables (Source 1) and the target (or dependent) variable (Source 2), respectively, based on a set of matching variables common to both data sources. The resulting linked file consists of correct matches (checkmarks) and mismatches (crosses).

This type of "fuzzy matching" can produce record linkages where the probability of a correct match is lower than 1 for certain records in the linked data set. This type of error in record linkage can produce outliers in terms of relationships of interest and may adversely alter the performance and outputs of applied predictive modeling techniques, such as variable importance, confusion matrices, RMSE, etc. Mismatch errors may ultimately prevent the realization of the actual predictive performance of these machine learning techniques, introducing a need for adjustments to the predictions that correct for this problem. Addressing the general absence of such adjustment approaches in the literature, Ben-David et al. (2023) derived and described novel adjustment techniques for the machine learning context based on a general mixture modeling framework (Hof & Zwinderman, 2015; Slawski et al., 2024). Via theoretical development and empirical simulation studies, these authors demonstrated that the proposed adjustment approaches can effectively improve predictions based on selected machine learning algorithms in the presence of various levels of mismatch error.

In this paper, our goal is to apply the methodology presented by Ben-David et al. (2023) to the specific context where 1) survey researchers are interested in linking survey and social media data, 2) fuzzy matching in the record linkage process is likely to introduce mismatch errors, and 3) the researchers wish to apply machine learning techniques to study relationships of interest in the linked data set. We evaluate the performance of this new adjustment methodology in an application involving the use of observed Twitter activity measures and predicted socio-demographic features of Twitter users to accurately predict linked measures of political ideology that were collected in a designed survey, where respondents were asked for consent to link any Twitter activity data to their survey responses (exactly, based on Twitter handles). We aim to demonstrate the use and importance of this new adjustment methodology to survey researchers interested in linking new sources of social media to survey data and ultimately applying machine learning techniques to the resulting linked data sets. We also summarize the limitations of the current adjustment approaches and make recommendations for future work in this area.

Methodology

An Overview of Adjustment Approaches Based on Mixture Modeling

We begin with an overview of our general approaches to adjusting modern predictive modeling algorithms for the presence of mismatch error. This paper focuses on possible adjustment techniques for *ensemble methods,* including bagging (or bootstrap aggregating) and random forests (distinguished from bagging by the selection of a random subset of predictors at each step of decision tree construction). For brevity, we focus on a heuristic explanation of the approaches and do not provide explicit mathematical or technical details here; interested readers can find these details in Ben-David et al. (2023).

In general, we are interested in using an ensemble method to estimate some general regression function $\mu_{y|\mathrm{x}} = E[y|\mathrm{x}]$, where y corresponds to a dependent variable of interest and x represents a vector of values on predictor variables of interest. The new adjustment methods introduced in this paper assume that the x variables are measured without error; we revisit this issue in the Discussion section. After a record linkage process, we have values on these variables of interest available for each subject in a study denoted by i, with $i = 1, \ldots, n$. In the *permuted* linked data file that arises due to a record linkage procedure subject to mismatch error (Figure 1), we (unfortunately) observe $\tilde{y}_i$ instead of y_i, where some fraction of the cases in the linked data file have a mismatched value on the dependent variable y. These mismatches are the source of the attenuation in the estimated relationships of interest defined by the regression function.

Following a mixture modeling approach, the overall distribution of the permuted version of y is a *mix* of two distributions: the conditional distribution of y defined by the regression function for those correctly matched cases (which gets a weight of $1 - \alpha$, where α is the probability of a mismatch error, meaning that the weight is the probability of a *correct* match), and the *marginal* distribution of y for the mismatched cases (without conditioning on the covariates), which gets a weight of α. The mixture model is flexible enough to allow a *unique* value of α for each case, denoted by α_i.

This mixture model implies that we can write the regression function as follows (where μ_y is the marginal mean of the variable y):

$$\mu_{y_i\,|\,\mathbf{x}_i} = \frac{1}{1-\alpha_i}\mu_{\tilde{y}_i\,|\,\mathbf{x}_i} - \frac{\alpha_i}{1-\alpha_i}\mu_y,\ i = 1, \ldots, n \tag{1}$$

When analyzing real data in practice, we would first apply the analyst's favorite predictive modeling algorithm to the linked data including mismatch errors. Given the resulting estimates of $\hat{\mu}_{\tilde{y}_1\,|\,\mathbf{x}_1}, \ldots, \hat{\mu}_{\tilde{y}_n\,|\,\mathbf{x}_n}$, along with the sample mean of the observed $\tilde{y}_i$, we can then substitute these quantities in (1). As a result, we can write the overall distribution of the permuted y as a function of α_i alone. Then, we can use maximum likelihood methods (or other optimization methods) to find an optimal $\hat{\alpha}_i^{opt}$ (see Algorithm 1 in Ben-David et al., 2023). This $\hat{\alpha}_i^{opt}$ can then be used in (1) to obtain an *improved* estimate of $\mu_{y_i\,|\,\mathbf{x}_i}$. We can also simply work with the mean of the $\hat{\alpha}_i^{opt}$, $\hat{\alpha}^{opt} = \sum_{i=1}^{n} \hat{\alpha}_i^{opt}/n$, in (1). We refer to this as a "mean optimal alpha" approach, which has the potential to save computational time. This is because we can efficiently estimate $\hat{\alpha}^{opt}$, the population mean of the $\hat{\alpha}_i^{opt}$, using the mean of a small random sample of the $\hat{\alpha}_i^{opt}$, with size much smaller than n.

The improvement in estimates of $\mu_{y_i \mid \mathbf{x}_i}$ based on this approach thus depends on (1) $\hat{\alpha}^{opt}$ being a good estimate of α, (2) $\hat{\mu}_{\tilde{y}_i \mid \mathbf{x}_i}$ being a good estimate of $\mu_{\tilde{y}_i \mid \mathbf{x}_i}$ (i.e., the regression function is specified correctly), and (3) the mixture model being a good fit for the overall distribution of the permuted y values. We note that this "optimal alpha" adjustment method would generally be applied *after* any other predictive modeling algorithm has been used to generate initial predictions $\hat{\mu}_{\tilde{y}_1 \mid \mathbf{x}_1}, \cdots, \hat{\mu}_{\tilde{y}_n \mid \mathbf{x}_n}$ for all cases in the linked data file.

Extending this idea to the more general context of the ensemble methods that are the focus of the current study, the α values described above can play the role of *weights* in the algorithms used to build the decision trees. We distinguish between two different approaches to using weights in the construction of decision trees: *adj-trees*, where differential case weights are used at each step of the tree construction process to determine optimal splits, and *adj-rf*, where differential case weights are used when the bootstrap samples are selected for the ensemble method (and cases with a higher weight would have a higher probability of selection).

Given no prior information about the mismatch probabilities, we would assign a weight of 1 to each case and set $\alpha_i = 0.5$ for all cases. We can then take, say, 100 bootstrap samples from the data (this number could be modified). For each sample, we first obtain $\hat{\mu}_{\tilde{y}_1 \mid \mathbf{x}_1}, \cdots, \hat{\mu}_{\tilde{y}_n \mid \mathbf{x}_n}$ from a decision tree, or random forests, with our initial weights. We can then use methods described in Ben-David et al. (2023) to compute the *posterior probability* of a mismatch given the predicted values according to the regression function, and then update the weight of each observation i as $1 - \alpha_i$. We then re-run the decision tree, or random forests, with these updated weights (which again either affect how the bootstrap samples are selected *or* how the tree is split at each node) to compute a new set of predictions $\hat{\mu}_{\tilde{y}_1 \mid \mathbf{x}_1}, \cdots, \hat{\mu}_{\tilde{y}_n \mid \mathbf{x}_n}$. We repeat this procedure, updating the weights and then updating $\hat{\mu}_{\tilde{y}_1 \mid \mathbf{x}_1}, \cdots, \hat{\mu}_{\tilde{y}_n \mid \mathbf{x}_n}$, until there is no numerical evidence of a significant improvement in the predictions obtained with the new weights. In the end, we average over the $\hat{\mu}_{\tilde{y}_1 \mid \mathbf{x}_1}, \cdots, \hat{\mu}_{\tilde{y}_n \mid \mathbf{x}_n}$ obtained from the final set of bootstrap samples and report this as the adjusted predictions $\hat{\mu}_{\tilde{y}_1 \mid \mathbf{x}_1}, \cdots, \hat{\mu}_{\tilde{y}_n \mid \mathbf{x}_n}$.

Ben-David et al. (2023) refer to this general approach as a *weighting-reweighting* adjustment method (Algorithm 2). Figure 2 visualizes this general approach. In theory, this adjustment procedure that assigns greater weight to cases with higher estimated probabilities of being a correct match will yield ensemble predictions with improved accuracy; simulations reported by Ben-David et al. (2023) provide empirical support for this concept.

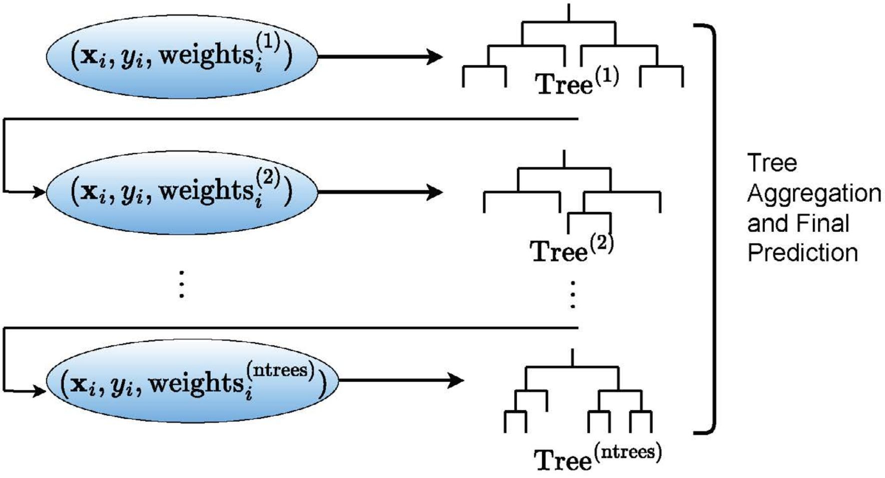

$(\mathbf{x}_i, y_i, \text{weights}_i^{(1)})$
$\text{Tree}^{(1)}$
$(\mathbf{x}_i, y_i, \text{weights}_i^{(2)})$
$\text{Tree}^{(2)}$
$(\mathbf{x}_i, y_i, \text{weights}_i^{(\text{ntrees})})$
$\text{Tree}^{(\text{ntrees})}$
Tree Aggregation and Final Prediction

Figure 2 A visual overview of the weighting-reweighting adjustment method in the context of ensemble methods such as bagging and random forests.

There are, therefore, several possible combinations of approaches that one could use when applying these ensemble methods to a linked data set. We distinguish between four methods that do not include the computation of optimal alpha values (referred to as basic bagging, basic random forests, adj-trees, and adj-rf) and four methods that do include the subsequent computation of optimal alpha values (optimal-alpha-bagging, optimal-alpha-rf, optimal-alpha-adj-trees, and optimal-alpha-adj-rf). In our analyses, we evaluate the performance of these eight alternative methods, summarized below in Table 1.

Data Source

We conduct secondary analyses of a linked data set (n = 448) that includes data from web survey respondents and aggregated measures of social media activity based on their linked Twitter profiles (we refer to Twitter, rather than X, as this data collection occurred prior to the change in the name of that platform). The web survey data, capturing measures of social media use, political attitudes and knowledge, and other related topics, were collected from a random sample of the Ipsos KnowledgePanel in January and February of 2020 (response rate = 76%); see Mneimneh (2022) for the original study design details. The record linkage was based on actual Twitter handles for those respondents who consented to this linkage, meaning that the record linkage was largely deterministic, exact, and error-free.

Given the objectives of our study, we randomly permuted the linked social media data to simulate mismatch errors (as the actual record linkage process used was unlikely to result in mismatch errors). As we noted in the Introduction, these types of mismatch errors may arise for several reasons when linking survey and social media data, but this mismatch error scenario may be even more common in other applications that involve linking survey data and administrative data (e.g., Patki & Shapiro, 2023).

Table 1 Alternative adjustment methods under consideration (none = no adjustment).

Adjustment method	Description
Bagging (none)	This is a standard application of bootstrap aggregating (bagging) using the original linked data and no random selection of predictors at each step of the tree construction.
Random forests (none)	This is a standard application of random forests similar to bagging but including the random selection of possible predictors at each step of the tree construction.
Adj-trees	The weighting-reweighting adjustment method, starting with default values of alpha (0.5) for all cases (and equal weights of 1), and then proceeding iteratively with applying weights to cases when splits are determined to construct individual trees. Improved estimates of the regression function are based on the mixture model.
Adj-rf	Like adj-trees, but applying the weights in the selection of the bootstrap samples (rather than in the formation of splits).
Optimal-alpha-bagging	A modification of bagging including a subsequent application of the optimal alpha algorithm to improve adjusted estimates based on the mixture model. Given our results and the additional computational burden introduced by using a unique optimal alpha for each case (without apparent benefits of this approach), we focus on the mean optimal alpha value for all “optimal alpha” approaches.
Optimal-alpha-rf	This is a modification of random forests, including the application of the optimal alpha algorithm to improve adjustment estimates based on the mixture model.
Optimal-alpha-adj-trees	This is a modification of adj-trees to include the optimal alpha algorithm.
Optimal-alpha-adj-rf	This is a modification of adj-rf to include the optimal alpha algorithm.

Measures

In our analysis, we focus on applying predictive modeling where we wish to predict a dependent variable representing an ordered measure of political ideology collected in the web survey. This question asked, "In general, do you think of yourself as…" and provided the following response options: 1 = extremely liberal; 2 = liberal; 3 = slightly liberal; 4 = moderate, middle of the road; 5 = slightly conservative; 6 = conservative; and 7 = extremely conservative. Given the roughly symmetric distribution of this variable among the survey respondents, we treated the variable as a continuous outcome in our analyses. Candidate predictors of this survey measure were all derived from the linked Twitter data. These included predictions of the person's gender (male vs. female) and age (>45 or <= 45) based on a neural network model (Liu & Singh, 2021), along with predictions of gun ownership (yes or no) and political party (Democrat or Republican) based on a random forest classifier using features of tweets and Twitter biographies. We also included as a predictor the overall number of tweets generated by the survey respondent (based on actual Twitter activity for the linked Twitter handle). We assume that all of these measures derived from the Twitter data are error-free; we return to this issue in the Discussion section.

Analytic Approach

In our evaluation of the eight alternative adjustment approaches described in Table 1, we first applied each of the eight approaches to the exactly matched Twitter and survey data (i.e., a 0% mismatch rate), evaluating the mean squared error (MSE) of the predictions for political ideology based on the correctly linked data. This initial analysis provided a benchmark for evaluating the success of the adjustment methodology after varying levels of mismatch error were introduced via random permutations (10%, 15%, …, 35%, 40%). We then evaluated the ability of the eight different approaches to recover this "ideal" MSE of the predictions based on the correctly-linked data. We constructed 100 trees based on bootstrap replicate samples for each ensemble method. We repeated these analyses 100 times and averaged the estimated MSE values across these 100 iterations.

Because ensemble methods may also be computationally expensive depending on the size of the data set and the number of predictors under consideration, we also compared the computational times associated with executing each adjustment procedure (based on a single run of each procedure). We provide separate computational times for each of the two algorithms described earlier, given that the use of optimal values of alpha for the weighting-reweighting adjustment approach also requires execution of the first algorithm to identify optimal values of alpha (possibly for each individual case). We weigh the comparisons of the procedures in terms of MSE based on the computational run times to identify an optimal adjustment procedure that is also computationally efficient.

Results

Table 2 compares the alternative adjustment methods in terms of the average estimated MSEs of the predictions of political ideology, averaged across the 100 iterations of each analysis and separately for different simulated mismatch rates.

Table 2 Relative performance of each adjustment procedure in terms of average estimated MSE (across 100 iterations) of the predictions for political ideology (best performance indicated in boldface).

	Mismatch rate							
Adjustment method	0%	10%	15%	20%	25%	30%	35%	40%
Bagging	**1.63**	1.68	1.74	1.76	1.80	1.82	1.87	1.93
Random forests	2.02	2.05	2.10	2.13	2.13	2.14	2.19	2.23
Adj-rf	1.99	2.00	2.01	2.02	2.03	2.04	2.06	2.09
Adj-trees	**1.63**	1.68	1.73	1.76	1.80	1.82	1.87	1.93
Optimal-alpha-bagging	1.64	**1.64**	**1.68**	**1.70**	**1.74**	**1.76**	**1.80**	**1.86**
Optimal-alpha-rf	2.09	2.07	2.10	2.12	2.12	2.12	2.16	2.20
Optimal-alpha-adj-rf	2.15	2.09	2.09	2.09	2.10	2.11	2.11	2.13
Optimal-alpha-adj-trees	1.64	**1.64**	**1.68**	**1.70**	**1.74**	**1.76**	**1.80**	**1.86**

The performance of each procedure when all matches are correct (i.e., when analyzing the original linked Twitter data) can be found in the Mismatch rate column of Table 2 labeled “0%.” In this setting, basic bagging and adj-trees have the best predictive performance (MSE = 1.63), and we use this as a benchmark to evaluate the performance of the alternative adjustment procedures when mismatches are introduced in the linked data. Examining the other columns of Table 2 corresponding to increasing mismatch rates (introduced by randomly permuting the values of the dependent variable for the indicated percentage of cases in the linked data set), we observe that the optimal-alpha-bagging and optimal-alpha-adj-trees approaches yield predictions that are consistently closest to the benchmark performance, with larger deviations from the benchmark as mismatch rates increase (as would be expected).

Given the results in Table 2, we next consider the computational run times associated with each procedure. Table 3 presents run times in seconds for the various components of the adjustment procedures.

Table 3 Run times in seconds for the various components of the adjustment procedures.

Optimal alpha	Mean optimal alpha	Bagging	Random forests	Adj-trees	Adj-rf
6.349	0.864	0.502	0.016	36.208	0.001

We note that a particular adjustment procedure may introduce the run times associated with each of the two algorithms. For example, the optimal-alpha-adj-trees approach requires subsequent execution of the optimal-alpha algorithm (6.349 seconds) following the adj-trees algorithm (36.208 seconds). Table 2 shows that the adj-trees approach tends to be computationally expensive. Combining these results with those in Table 2, it therefore seems that the optimal-alpha-bagging approach has the best overall performance in the setting considered here.

We have included the R code needed to carry out these analyses in the GitHub repository https://github.com/ehb2126/Data-Analysis-after-Record-Linkage.

Discussion

Summary of Contributions

Mismatch errors are common in probabilistic record linkage procedures. In the specific setting of linking survey data with social media data, these errors can arise for several reasons, including names provided by the consenting survey respondents that do not match with the names used for social media accounts, full names provided by consenting survey respondents that do not uniquely identify individuals, social media platform handles corresponding to user accounts containing typos that prevent exact matching, or consenting respondents changing their platform handles over time (Stier et al., 2020; Beuthner et al., 2021). At the same time, machine learning methods are becoming increasingly popular for studying complex relationships in the analyses of linked data sets from different sources (e.g., social media and survey data, or survey data and administrative data).

Much of the record linkage literature has focused on adjustment procedures for mismatch errors in classical parametric regression modeling. Recently, Ben-David et al. (2023) addressed an important gap in this area, focusing on optimal methods for adjusting for mismatch errors when applying modern prediction tools (specifically bagging and random forests) and describing alternative adjustment procedures for ensemble prediction methods within a mixture modeling framework. This paper applies these new adjustment methods to a case study linking survey data with social media (specifically Twitter/X) data, and demonstrates that these methods improve the performance of modern predictive modeling methods that were applied to this linked data set under various simulated rates of mismatch error.

We find that in the presence of these various rates of mismatch error, an adjustment methodology that combines bagging with optimal estimation of the probability of correct linkage for each case tends to have the best predictive performance, from the perspectives of both MSE of predictions and computational runtime. This procedure is straightforward to implement using available software, and we have implemented it using the R software (see the GitHub repository https://github.com/ehb2126/Data-Analysis-after-Record-Linkage).

Limitations and Directions for Future Research

We note that studies linking social media data with survey data generally use exact platform handles or other types of unique identifying information in the record linkage, and do not attempt the linkage at all if respondents do not consent to provide these handles or other user account information, such as full names (e.g., Al Baghal et al., 2021). This introduces the possibility of missed-match errors, a type of selection bias that could affect the performance of predictive modeling methods. Selection bias due to missed-match errors could affect machine learning algorithms that are focused on prediction in three ways (Quiñonero-Candela et al., 2022):

1) *covariate shift,* where the distribution of the predictors x would differ across successfully linked cases and missed matches;

2) *label shift,* where the distribution of the dependent variable *y* would differ across successfully linked cases and missed matches; or

3) *concept drift,* where the distribution of *y* conditional on x would differ across successfully linked cases and missed matches, and the classification rule would depend on the successfully linked cases.

If we assume that an indicator of successful linkage is independent of *y* when conditioning on x, then *concept drift* does not hold, but this is a strong assumption that needs to be evaluated in future simulation studies. Adjustment approaches accounting for these types of missed match errors and allowing for violations of this assumption are still needed in the machine learning context; we only focused on mismatch errors in the current application.

The new methodologies illustrated in this paper also assume that the mismatch errors occur *completely at random*, using the terminology of Little and Rubin (2019) in the missing data context. This strong assumption may not hold in real applications, since the probability of a mismatch error may at least depend on the values of observed covariates. We designed an additional simulation study to evaluate the performance of the methodology in a setting where the probability of a mismatch depends on the value of the covariate that had the strongest relationship with political ideology in a linear regression model fitted to the political ideology outcome in the original data: the binary prediction of preferring the Republican party (1 = yes, 0 = no). The supplemental materials describe the design of this additional simulation study and the corresponding results.

Summarizing those results here, we find that the methods identified as having the best performance in the “mismatch completely at random” scenario have equally strong performance in this informative mismatch error scenario. Despite these positive results, additional theoretical development is still needed to understand why the current methodologies also seem to work well in this informative mismatch error setting; they are presently designed for mismatches occurring completely at random. Future research on this methodology should also aim to accommodate more complicated types of informative mismatch error scenarios.

We also did not quantify variable importance in our application of the adjustment methods. We have not yet developed a procedure for identifying the most important predictors that emerge from one of these adjustment approaches, and work on the development of adjusted variable importance measures is ongoing. This is another worthwhile direction for future research.

We also note that we assumed that all of the social media measures were of sufficiently high quality. These variables computed from the Twitter/X data were either predictions of user characteristics or counts of tweets that may themselves be subject to prediction error and sampling error. Future applications involving predictive modeling of linked survey and social media data need to carefully consider potential sources of error in derived variables from social media activity and ensure that these errors are either corrected, adjusted for, or transparently described in written summaries of the modeling applications.

Finally, while the adjustment approaches in this paper were evaluated in the context of mismatch error in linked social media and survey data, we anticipate that they will also have widespread application in other substantive settings where probabilistic record linkage is used (e.g., Patki & Shapiro, 2023) and researchers are interested in predictions based on machine learning procedures.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)* (pp. 30–38). Association for Computational Linguistics. https://dl.acm.org/doi/abs/10.5555/2021109.2021114

Al Baghal, T., Sloan, L., Jessop, C., Williams, M., & Burnap, P. (2020). Linking Twitter and survey data: The impact of survey mode and demographics on consent rates across three UK studies. *Social Science Computer Review, 38* (5), 517–532. https://doi.org/10.1177/0894439319828011

Al Baghal, T., Wenz, A., Sloan, L., & Jessop, C. (2021). Linking Twitter and survey data: Asymmetry in quantity and its impact. *EPJ Data Science, 10* (1), 32. https://doi.org/10.1140/epjds/s13688-021-00286-7

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45–59. https://doi.org/10.1016/j.inffus.2015.08.005

Ben-David, E., West, B. T., & Slawski, M. (2023). A novel methodology for improving applications of modern predictive modeling techniques to linked data sets subject to mismatch error. In *Big Data Meets Survey Science (BigSurv)* (pp. 1–8). IEEE. https://doi.org/10.1109/BigSurv59479.2023.10486610

Beuthner, C., Breuer, J., & Jünger, S. (2021). *Data linking – Linking survey data with geospatial, social media, and sensor data* (GESIS Technical Report, Version 1.0). https://doi.org/10.15465/gesis-sg_en_039

Conrad, F. G., Keusch, F., & Schober, M. F. (2021). New data in social and behavioral research. *Public Opinion Quarterly, 85* (S1), 253–263. https://doi.org/10.1093/poq/nfab027

Dalzell, N., & Reiter, J. (2018). Regression modeling and file matching using possibly erroneous matching variables. *Journal of Computational and Graphical Statistics, 27*(4), 728–738. https://doi.org/10.1080/10618600.2018.1458624

Eady, G., Nagler, J., Guess, A., Zilinsky, J., & Tucker, J. A. (2019). How many people live in political bubbles on social media? Evidence from linked survey and Twitter data. *Sage Open, 9* (1), 1–21. https://doi.org/10.1177/2158244019832

Gautam, G., & Yadav, D. (2014). Sentiment analysis of Twitter data using machine learning approaches and semantic analysis. In *Seventh International Conference on Contemporary Computing (IC3)* (pp. 437–442). IEEE. https://doi.org/10.1109/IC3.2014.6897213

Ghani, N. A., Hamid, S., Targio Hashem, I. A., & Ahmed, E. (2019). Social media big data analytics: A survey. *Computers in Human Behavior,* 101, 417–428. https://doi.org/10.1016/j.chb.2018.08.039

Han, Y., & Lahiri, P. (2019). Statistical analysis with linked data. *International Statistical Review, 87*(S1), 139–157. https://doi.org/10.1111/insr.12295

Hof, M., & Zwinderman, A. (2015). A mixture model for the analysis of data derived from record linkage. *Statistics in Medicine, 34*(1), 74–92. https://doi.org/10.1002/sim.6315

Karlsen, R., & Enjolras, B. (2016). Styles of social media campaigning and influence in a hybrid political communication system: Linking candidate survey data with Twitter data. *The International Journal of Press/Politics, 21*(3), 338–357. https://doi.org/10.1177/1940161216645335

Lahiri, P., & Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association, 100*(469), 222–230. https://doi.org/10.1198/016214504000001277

Little, R. J., & Rubin, D. (2019). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119482260

Liu, Y., & Singh, L. (2021). Age inference using a hierarchical attention neural network. In *Proceedings of the ACM International Conference on Information & Knowledge Management* (pp. 3273–3277). Association for Computing Machinery. https://doi.org/10.1145/3459637.3482055

McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., & Spiro, E. S. (2017). Using Twitter for demographic and social science research: Tools for data collection and processing. *Sociological Methods & Research, 46* (3), 390–421. https://doi.org/10.1177/0049124115605339

Mneimneh, Z. (2022). Evaluation of consent to link twitter data to survey data. *Journal of the Royal Statistical Society Series A: Statistics in Society, 185*(Supplement 2), S364–S386. https://doi.org/10.1111/rssa.12949

Neter, J., Maynes, S., & Ramanathan, R. (1965). The effect of mismatching on the measurement of response error. *Journal of the American Statistical Association, 60*(312), 1005–1027. https://doi.org/10.1080/01621459.1965.10480846

Patki, D., & Shapiro, M. D. (2023). Implicates as instrumental variables: An approach for estimation and inference with probabilistically matched data. *Journal of Survey Statistics and Methodology, 11*(3), 597–618. https://doi.org/10.1093/jssam/smad005

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds). (2022). *Dataset shift in machine learning*. The MIT Press. https://mitpress.mit.edu/9780262545877/dataset-shift-in-machine-learning/

Scheuren, F., & Winkler, W. (1993). Regression analysis of data files that are computer matched. *Survey Methodology, 19*(1), 39–58. https://www150.statcan.gc.ca/n1/pub/12-001-x/1993001/article/14476-eng.pdf

Scheuren, F., & Winkler, W. (1997). Regression analysis of data files that are computer matched – Part II. *Survey Methodology, 23*(2), 157–165. https://www150.statcan.gc.ca/n1/pub/12-001-x/1997002/article/3613-eng.pdf

Slawski, M., Diao, G., & Ben-David, E. (2021). A pseudo-likelihood approach to linear regression with partially shuffled data. *Journal of Computational and Graphical Statistics, 30*(4), 991–1003. https://doi.org/10.1080/10618600.2020.1870482

Slawski, M., West, B. T., Bukke, P., Wang, Z., Diao, G., & Ben-David, E. (2024). A general framework for regression with mismatched data based on mixture modelling. *Journal of the Royal Statistical Society Series A: Statistics in Society,* qnae083. https://doi.org/10.1093/jrsssa/qnae083

Steorts, R. C., Tancredi, A., & Liseo, B. (2018). Generalized Bayesian record linkage and regression with exact error propagation. In J. Domingo-Ferrer & F. Montes (Eds.), *Privacy in statistical databases. PSD 2018. Lecture Notes in Computer Science* (Vol. 11126, pp. 295–306). Springer. https://doi.org/10.1007/978-3-319-99771-1_20

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review, 38*(5), 503–516. https://doi.org/10.1177/0894439319843

Tancredi, A., & Liseo, B. (2015). Regression analysis with linked data: Problems and possible solutions. *Statistica, 75*(1), 19–35. https://www.proquest.com/docview/1765123569?pq-origsite=gscholar&fromopenview=true&sourcetype=Scholarly%20Journals

Wan, Y., & Gao, Q. (2015). An ensemble sentiment classification system of Twitter data for airline services analysis. In *IEEE International Conference on Data Mining Workshop (ICDMW)* (pp. 1318–1325). IEEE. https://doi.org/10.1109/ICDMW.2015.7

Appendix

Improved Ensemble Predictive Modeling Techniques for Linked Social Media and Survey Data Sets Subject to Mismatch Error

Simulation Study of Informative Mismatch Error

To compare the MSE of methods where mismatches correlate with a given predictor x (in this case, the binary prediction of preferring the Republican party) versus ones where the mismatches occur completely at random, we need to consider the average percentage of cases where x = 1 is swapped with x = 0 when mismatches are introduced completely at random. Table A1 below shows this average for various mismatch rates using the same simulation approach described in the paper, with each average computed numerically based on 100,000 replications.

Table A1 x = 1 permutation rates introduced by mismatches occurring completely at random.

Percentage of mismatches (completely at random)	Average percentage of x = 1 swapped with x = 0
10%	10%
15%	17%
20%	21%
25%	23%
35%	31%
40%	35%

For this supplemental simulation study, we selected probabilities of changing values from x = 1 to x = 0 that were consistent with the table above, to ensure that overall mismatch rates were similar to those already evaluated in the paper.

Table A2 below shows the MSEs of bagging, random forests, and the new adjustment methods for these informative mismatch error scenarios. We used the same simulation approach described in the paper, but allowed the probability of a mismatch error to change for cases with x = 1. Each column of Table A2 below shows the MSEs of these methods in a different mismatch scenario, where P(x’ = 0 | x = 1) varies according to percentages comparable with the “completely at random” mismatch rates given in Table A1 above. In addition, the probability of a mismatch error was set to be larger for cases with x = 1: P(x’ = 1 | x = 0) = (71/377) × P(x’ = 0 | x = 1), which shows how the covariate was related to the probability of mismatch error. The MSEs in the table are averaged over 250 iterations.

Table A2 Informative mismatch error simulation results (MSEs).

P(x' = 0 | x = 1)	**0**	**0.1**	**0.17**	**0.21**	**0.23**	**0.25**	**0.31**	**0.35**
P(x' = 1 | x = 0)	**0**	**0.019**	**0.032**	**0.040**	**0.043**	**0.047**	**0.058**	**0.066**
Bagging	1.63	1.64	1.66	1.67	1.69	1.69	1.74	1.76
Random forests	2.02	2.05	2.08	2.10	2.11	2.12	2.17	2.20
Adj-rf	1.99	1.99	2.00	2.00	2.01	2.01	2.03	2.05
Adj-trees	1.63	1.64	1.66	1.67	1.69	1.69	1.74	1.76
Optimal-bagging	1.64	1.60	1.59	1.59	1.60	1.60	1.63	1.65
Optimal-rf	2.09	2.08	2.09	2.10	2.10	2.11	2.15	2.17
Optimal-adj-rf	2.15	2.12	2.10	2.09	2.09	2.08	2.08	2.09
Optimal-adj-trees	1.64	1.60	1.59	1.59	1.60	1.60	1.63	1.65

Overall, we see performance quite similar to that in the mismatch completely at random scenario that was analyzed in the paper. The adj-trees, optimal-bagging, and optimal-adj-trees approaches all tend to have the best performance, and while the MSEs increase somewhat as the "conditional" mismatch probabilities increase, these methods consistently have the best performance in this scenario.

Reflective Appendix

1. If you had been required to pre-register your methodological approach, in advance of conducting your research, how would you have described it?

The main objective of the paper was to evaluate the ability of a new methodology, presented by the authors in previously published work, to improve the accuracy of predictions of respondents' self-reported political ideology using their Twitter account information. For purposes of the analysis, we obtained a secondary dataset of US Ipsos KnowledgePanel participants (n = 448) who responded to a web survey measuring social media use, political attitudes and knowledge, and other related topics. All 448 web survey respondents had consented to let the study designers link their survey responses to their Twitter/X accounts, and all 448 had provided their correct Twitter handles for the linkage. Given the 100% matching of respondents to their Twitter accounts, we had to simulate mismatch error to test our methodology. The methodological approach used bagging and random forest techniques to predict an ordinal dependent variable measuring political ideology from the web survey. The predictors of interest included predicted socio-demographic information and aggregated measures of activity from the linked accounts on the Twitter/X platform.

2. To what extent did you make any modifications to your plans as described above in the course of producing the final version of the paper?

Since we were using secondary data that was selected explicitly for the purposes of this analysis, we were able to specify clearly and precisely the methodology adopted in advance and we did not deviate from what was originally proposed. Having a dataset with no mismatches provided us with an important and valuable benchmark against which to assess the performance of the varying methods for predicting survey answers on ideology from individuals' Twitter data. However, having a pre-cleaned dataset generates some "costs" to the analysis in terms of limiting the diagnostics we were able to perform and our capacity to make adjustments or modifications to the analysis. Having access to the larger original dataset from which this subset of 448 correctly matched individuals were drawn would have provided the opportunity to draw more generalizable conclusions. Specifically, prior work using this larger dataset (Mneimneh, 2022) revealed that of the 58.6% of survey respondents that consented to the Twitter linkage, less than half (48%) provided a useable handle. Had the dataset included the correctly matched and larger sample of mismatched respondents it would have been possible to conduct diagnostics on the former sample to see how closely they resembled the latter, and whether mismatches were more likely for certain types of individuals. For example, were our sample respondents more active on Twitter than the average user? The counts of tweets and retweets ranged from 21 to 75,077 for the fully matched sample of 448 we investigated and no individuals had counts of zero or missing data. If bias did exist, this type of "informative" mismatch error would have been useful for adjusting our hypothetical simulated mismatch error to correspond to ”organic” mismatch error, thereby enhancing the robustness of its application to larger samples that are more at risk of the "organic" mismatch error.

Additionally, having access to the larger sample that included organically mismatched survey respondents would have allowed us to replicate the analysis performed here, and assess the results of the predictive algorithms employed for the correctly matched subsample in the larger (more representative) sample of Twitter users. It is possible that our methodology is more effective than is demonstrated here in the context of stronger predictive models (where the mismatch error is likely to have larger attenuating effects on predictive performance). For this we would need to have access to aggregate information regarding the characteristics of the Twitter-using population at the time when this survey was collected, and microdata from non-consenting respondents to determine whether the power of the predictive models was lower in our subsample compared to a larger set of Twitter users.

3. Can you list up to 3 practical steps that you would recommend, based on what you learned doing this research project, future researchers take into

account when working with similar data sources? These would ideally be relevant to the methods, data and analysis chosen.

Four practical steps that we would recommend when working with linked survey and social media data and using predictive modeling techniques to study relationships among variables from the two linked data sources:

1. Evaluate the quality of the record linkage process. Were all linkages correct, or was there evidence of problems in the process that would lead to potential mismatches? For example, are the useable Twitter handles provided actually those of the survey respondents, or those of *other* individuals? Providing the Twitter handle of another individual could lead to another type of mismatch error (above and beyond the sources mentioned in the paper), and our methodology would still be able to accommodate this alternative type of linkage error. Asking consenting survey respondents for their two most recent Twitter handles to facilitate record linkage is one possible approach for dealing with non-useable handles or handles of other users that survey respondents have provided. All else being equal, we would avoid the collection of names, addresses, and other personal identifying information in an attempt to resolve these problems, as this raises new ethical concerns. Our methodology is designed to address the mismatch errors that may result from this “less than ideal” type of respondent behavior.

2. If alternative variables aside from the Twitter handles are ultimately used to perform the linkage for selected cases, try to obtain information on the correctness of each link, when possible, taking into account ethical considerations regarding the protection of respondent confidentiality. For example, this could involve calculating the predicted probability of correct linkage arising from a probabilistic record linkage procedure. Such information will be helpful in informing adjustment approaches like the one evaluated in this paper. Alternatively, if such information is not available, anticipated mismatch rates or block-wise mismatch rates can still be helpful. Blocks define different groups of cases based on discrete observable characteristics (sex, race, age, etc.) within which linkage is considered.

3. If there is a risk of mismatch error and the analyst ultimately wants to use a modern predictive modeling technique to predict values of a variable in one data source with predictors from the other data source, consider using the R software provided with this paper to perform the bagging and random forests (rather than standard procedures implementing these techniques, which would be adversely affected by the mismatch error). We find in this paper that an adjustment methodology that combines bagging with optimal estimation of the probability of correct linkage for each case tends to have the best predictive performance.

4. Compare the performance of the adjusted predictive modeling methods with that of the standard predictive modeling methods to quantify the potential effects of the mismatch error on the performance of the techniques. If notable differences in performance are observed, report predictions based on the adjusted methods, as the standard techniques were likely affected by the mismatch errors engendered by the record linkage process. If a fraction of the linked data set has records that were linked deterministically (i.e., with no probability of linkage error), predictions based on a machine learning algorithm applied to those "exactly matched" cases could be used as a benchmark, and predictions based on an application of our methodology to the *full* data set (including mismatch errors) could be compared in a validation analysis to assess the effectiveness of our methodology.

More generally, we believe that the methodology illustrated in this case study can be generalized and transported to other applications involving the linkage of records in novel data sources, and we also believe that several extensions of our approach are possible. First, mismatch errors may not necessarily cause estimation problems (for example, in a binary classification problem, if the mismatched record has the same value on the binary variable, there is no impact). Second, we may want to use information on "local" mismatch rates (e.g., changing with covariates or with information about the quality of the record linkage) rather than using a global mismatch rate model (as was the case in this study). The availability of metadata in the specific setting of linking social media information (e.g., geographic location, age of account, employment history, etc.) may be helpful for improving estimation of these "local" mismatch rates, in turn improving the adjustments engendered by our methodology. In the simulation study that we considered to look at "informative" mismatch errors, we found that the methods identified as having the best performance in the "mismatch completely at random" scenario have equally strong performance in the informative mismatch error scenario. However, we also found that the conditional mismatch rate did have an impact on predictive performance overall, meaning that additional enhancements of our methodology to implement larger adjustments for certain subgroups of cases with larger associated mismatch rates may be important.

Third, missed matches may be even more problematic than mismatch errors, given that we might be training a machine learning model on a non-representative sample. Fourth, our approach can be connected to robustness (Slawski et al., 2021). In this setting, other sources of measurement error or outliers can be handled simultaneously. At the same time, mismatch errors and other measurement errors cannot be distinguished, which makes sense, since their impact is often identical. For example, an incorrect link or a data entry error in terms of computed Twitter activity are equivalent in terms of impact. Fifth, the size of the data set may be important when choosing the specific adjustment method. The optimal-alpha method may be prohibitive from the point of view of computational runtime. Finally, the general methodology illustrated here can be applied to other types of variables of interest; regression functions based on specified link functions in generalized linear models could be used as part of the algorithm to accommodate other types of dependent variables of interest (binary, count, etc.) in other linked data sets.

We believe that it would be easy to apply our methodology (designed for the secondary analysis setting) in other settings where novel data sources have been linked, a data analyst who did not perform the linkage is working with the linked data file, and mismatch errors are suspected. As noted in the paper, having the probability of a correct match for each case can be helpful for the algorithms, but is *not required*. The methodology can therefore proceed in the absence of quality indicators in the linked data file, making applications of the methodology easy for data users (no matter what types of data sources have been linked). We are confident that the optimal adjustment approach identified in this case study would also emerge in other larger data sets or other applications involving the linkage of other types of novel data sources. However, as we note above, the optimal alpha adjustment method may be computationally prohibitive in larger data sets, in which case the “next best” methods may need to suffice.

Additional research involving applications of our methodology in other settings would shed more light on these practical and computational issues, and likely lead to additional refinements of the new methodology presented here. As we note in the Discussion, additional research examining the performance of the adjustment methodology in more complex settings of *informative* mismatch error is necessary. We performed a relatively simple simulation study, and depending on the variables of interest in the analysis and the nature of the informative mismatch error as a function of these variables, other more optimal approaches may emerge. There is a possibility that linking alternative types of data sources could result in more complex patterns of informative mismatch error (e.g., certain socio-demographic subgroups are less likely to provide correct or truthful information related to handles for a particular type of social media platform), and this should be a focus in future research that seeks to refine our adjustment methodology.

methods, data, analyses | Vol. 19(2), 2025, pp. 146-176 DOI: 10.25521/mda.706

Improving Assessments of Group-Based Appeals in Political Campaigns by Systematically Incorporating Visual Components of Ads

Niamh Cashell

University of Manchester

Abstract

Existing research on group-based appeals primarily uses text-based methods, and while many studies show the importance of visuals in implicitly cueing groups, this data is rarely captured in a systematic way. This paper seeks to make the first important step towards filling this gap by outlining a coding scheme to evaluate how group-based appeals are used multimodally in modern political campaigns. This paper builds categories from a qualitative sample of 182 images taken from 28 television and 63 Facebook ads from candidates running in the US 2020 House of Representatives elections. Direct appeals are captured as explicit group mentions and I present new categories for indirect and baseline appeals, which incorporate primarily visual indicators of groups. Intercoder reliability tests were conducted, and the schema was applied to a larger sample of 2480 images from 125 television ads from candidates running in the three most populous states (California, Texas, Florida). This paper finds that candidates use direct and indirect appeals at similar rates, often using them in combination. Capturing visual data therefore enables greater coverage of the range of group-based appeals that political campaigns conduct. Secondly, candidates are more likely to cue occupational groups indirectly, and capturing only direct cues may lead to skewed findings in terms of which groups candidates appeal to. I find that this new coding scheme may reduce bias in measures of both the prevalence of group-based appeals and the types of groups that campaigns appeal to in modern political discourse.

Keywords: group appeals, visual methods, election campaigns, group targeting, political communication

Political campaigns' appeals towards voter groups is receiving increased interest in political science (Huber, 2022; Huber & Dolinsky, 2023). Currently, group-based appeals are measured from text-based sources such as speeches and manifestos (Dolinsky, 2022; Horn et al., 2021; Huber, 2022; Thau, 2019, 2021). Where multimodal media such as print campaign materials are analyzed, only the textual content is typically evaluated (Dolinsky, 2022). Textual content, however, only reveals part of the picture, and studies suggest that groups can be indicated indirectly, through visuals (McIlwain & Caliendo, 2011; Swigger, 2012). Visual communication is even more prominent today as newer forms of social media such as Instagram and TikTok are primarily visual platforms. Despite the evident importance of images, we lack a methodology to systematically incorporate indirect measures into assessments of group-based appeals. The focus on text-based methods is part of a broader pattern in political science that prioritizes textual over visual content, partly due to methodological reasons such as the volume of images available on social media raising questions around size and scope, as well as images being viewed as more subjective and a hinderance to reasoning (Coleman, 2010; Dean, 2019; Graber, 2012). However, by incorporating multimodal features of campaign content into assessments of group-based appeals, researchers can unlock new insights into the wider range of ways in which campaigns signal groups. In this paper I propose an approach for systematically coding group-based appeals using visual data such as campaign advertisements or social media posts.

This paper first reviews the literature and demonstrates that group-based appeals are currently measured by text-based methodologies, before moving to argue why visual and indirect appeals are likely to be important for political campaigns and how this data can be incorporated. The paper then outlines how the coding scheme was developed from a qualitative analysis, before presenting the schema. The coding scheme is tested for intercoder reliability with a second coder and applied to television ads in the 2020 US House of Representatives elections for candidates running in the three most populous states. Application of the coding scheme reveals that indirect cues both provide additional context to direct appeals and constitute group-based appeals in themselves and are therefore important to capture to make accurate assessments of how campaigns target groups.

Direct correspondence to
Niamh Cashell, University of Manchester, Manchester, UK
E-mail: niamh.cashell@postgrad.manchester.ac.uk

Background

Existing Approaches to Measuring Group-Based Appeals: Textual Analysis Methods

Since the 1960s political scientists have recognized the importance of groups in politics and campaigns, and it is therefore important to capture the full range of ways in which they are appealed to (Ford & Jennings, 2020; Lipset & Rokkan, 1967). Group-based voting is a two-step process in which voters naturally link themselves to a social group, and this group is associated with a political party (Butler & Stokes, 1969; Campbell et al., 1980; Conover, 1988). Political parties play a role in this process, by representing some groups over others, they help foster and sustain group identities and articulate and mobilize group demands (Lipset & Rokkan, 1967). Group-based appeals are used by political parties to signal which groups they will represent if elected.

Huber and Dolinsky (2023) define a group-based appeal as *"an intentional act that associates a political actor with or dissociates them from a social group"* (p. 11) and distinguish between direct and indirect appeals. Direct appeals constitute overt and unambiguous communication toward a group, for example, explicit statements of endorsement (Huber & Dolinsky, 2023). For example, statements referencing demographic groups such as 'women' and 'Latino voters', economic groups such as 'farmers', 'workers', and religious groups such as 'Catholics' and 'Muslims', would count as direct appeals. In this way, the *intentional* linking to a group is clear, a campaign explicitly mentions a group or they do not. In contrast, indirect appeals are when "no overt mention of a group is observed but the party instead uses symbols or language associated" with a group or proposes policies that impact a group without naming them directly (Huber & Dolinsky, 2023, p. 17). Huber and Dolinsky (2023) observe that indirect appeals can be conducted through symbols, language or policies associated with a group. These have been less well explored in the literature due to the challenge of reading the intention behind such group linkage (Huber & Dolinsky, 2023). While direct appeals constitute the foundation of what has already been evaluated by existing methods, indirect appeals constitute the methodological gap, or 'new' data that this paper seeks to address.

Existing methods evaluating textual content demonstrate that group-based appeals play an important role in political campaigns' electoral communication. Most studies focus on Europe and show that group-based appeals are moving away from class-based groups and towards demographic and identity-based groups such as lifecycle groups (for example, the young, elderly, pensioners) (Dolinsky 2022). In the UK, political parties appeal to a greater number of groups today than they did in the 60s, even when the length of manifestos is controlled for (Thau, 2019). Many of these studies have relied on hand-coding textual data (Dolinsky, 2022; Horn et al., 2021; Huber, 2022; Thau, 2019, 2021), although work on appeals increasingly uses and develops computational methods such as detection of keywords and supervised classification language models (Licht & Sczepanski, 2024). This leaves a notable gap for two reasons. Firstly, such approaches are missing a significant component of group-based appeals, particularly the more cultural, symbolic, and implicit which is now arguably a primary component. Secondly, this component is likely to be even more important in the US context, where implicit cues have been shown to signal cultural and particularly racial messages (McIlwain & Caliendo, 2011; Mendelberg, 2001).

The Need for New Methods to Capture Indirect and Visual Group-Based Appeals

There are three key reasons why campaigns may be motivated to use visual communication for group-based appeals, underscoring the importance of capturing this data. Firstly, audiovisuals provide rich opportunities for audiences to learn about politics, particularly for those with low political interest and knowledge, and therefore are likely to be useful for political campaigns (Graber, 2001). The inclusion of visuals enhances memory and accuracy in recalling news, as well as invoking emotion in audiences, and political campaign professionals will try to maximize the benefits of this (Graber, 2001). Secondly, including images of groups in ads may be a lower-risk strategy for political campaigns than conducting direct group-based appeals. Showing a group visually involves less commitment than explicitly stating which group you will represent and, according to Swigger (2012), is an effective way of positively associating candidates with a group while avoiding committing to potentially unpopular policies. Finally, developments in media technologies are likely to further incentivize candidates to conduct group-based appeals with visual cues. Ads are cheaper to run on social media than television, and campaigns can microtarget messages, which may motivate campaigns to target groups through this medium (Fowler et al., 2023). Television ads are already inherently visual, and visual platforms such as TikTok and Instagram are therefore likely to continue these trends.

Having argued that visuals are likely to be used for group-based appeals, I now turn to the question of how such appeals can be measured. As highlighted by Huber and Dolinsky (2023), studies do explore how parties associate or disassociate from social groups, although they are rarely framed as group appeals. Scholars of race in particular have studied implicit messages about racial groups, and Mendelberg's (2001) analysis of the infamous Willie Horton ad is a good example of this. Textually, the 1988 Bush ad references 'murderers', while showing a threatening image of an African American man who committed violent crimes while on weekend release from prison. In this way, the ad disassociates from African Americans as a group using visual racial stereotypes. Using this example, traditional text-based approaches would record the ad as an appeal against 'murderers' as a collective group. The visual information of showing an African American man however provides additional context as to the racial aspects of who is meant by this term, and capturing the visual changes our interpretation as an appeal against criminals to disassociating from and demonizing African Americans as a group.

Candidates can indirectly associate with groups through the visible demographic characteristics of people included in ads, as well as through additional signifiers such as clothing and symbols. McIlwain and Caliendo (2011) conduct a systematic content analysis of House and Senate ads between 1970 and 2006, coding ads into 56 variables relating to racist potential. One variable evaluates whether non-candidates included in images are white, and although they argue that this kind of imagery is not enough to constitute a racist appeal in itself, it does indicate who is being included and excluded (McIlwain & Caliendo, 2011). Taken together with Mendelberg's (2011) analysis, these studies suggest that the visible demographics of people included in ads could be one variable in which groups are signaled and provide important information about who a candidate represents. Furthermore, Benoit's (2019) study of visual and verbal symbols in presidential campaign posters dating back to 1828 found that images depicted groups such as blacksmiths, farmers and miners through the setting, such as factories, and clothing. Therefore, while not studying group appeals specifically, this suggests that campaigns have been depicting and signaling groups through various symbols since the early days of political campaigning in the US.

Visual features may enhance emotions and perceptions of groups included in ads and are therefore important to consider in developing a methodology. These can be studied using a visual social semiotic approach, which focuses on relationships between the viewer and the image, such as camera angle and gaze (Feng & O'Halloran, 2012; Kress & Van Leeuwen, 1996). Setting, facial expression, gaze, distance and gestures all contribute towards perceptions of a candidate's credibility (Kaid & Johnston, 2001; Page & Duffy, 2009). Close-ups suggest intimacy, smiling and eye contact increase perceptions of likeability and authenticity, while casual clothing and setting convey authenticity (Kaid & Johnston, 2001; Kress & van Leeuwen 1996; Page & Duffy 2009). Additionally, color tone of an ad is likely to be important, as negative ads cue fear through shadowed lighting and contrasts (Brader, 2005; Jamieson, 1992). More recent studies have used automated image analysis to detect facial expressions and emotions of people included in ads (Bossetta & Schmøkel, 2023). Commercially available image labelling tools such as Google Cloud Vision, Amazon Rekognition and Clarifai have been used to tag images and conduct automated visual content analyses (Araujo et al., 2020; Bossetta & Schmøkel, 2023; d'Andrea & Mintz, 2019; Geboers & Van De Wiele, 2020). Although these automated methodologies do frequently exhibit gender and racial biases (Barlas et al., 2021; Neumayer & Rossi, 2022), it is promising for future visual political communication research that such tools are being developed. For automated image analysis tools to be used successfully in the future, it is important to develop an understanding of the use of different kinds of visual features and create gold-standard human data to compare these automated approaches against.

Data and Methods

To build a methodology to capture visual aspects of political campaigns and to measure group-based appeals, a qualitative approach was firstly undertaken to build categories from the bottom up. This ensured that the coding scheme was data-driven and responsive to appeals present in ads. The scheme was then checked for intercoder reliability through a second coder and applied to a larger sample to evaluate whether the categories work more broadly.

Data Collection and Sampling

To conduct the qualitative exploration, intercoder reliability and proof of concept, three samples were created. Table 1 outlines how these samples were developed and for what purpose, the column 'Referenced as' indicates how each is referenced throughout the paper.

Firstly, a *qualitative sample* of television and social media ads was used to explore how group-based appeals may be conducted multimodally to develop the coding scheme (Table 1). Television adverts were accessed through the Wesleyan Media Project (WMP; Fowler et al., 2023), which provides the video files of television adverts put out by candidates. Social media adverts were accessed through the Meta ad library. The first sweep for the qualitative analysis included 134 images. 78 images were screenshotted from 23 television ads, and 56 images were screenshotted from 58 sampled social media ads. I subsequently collected more images from different candidates using the same random sampling and technique, resulting in a further 48 images (182 total), to ensure saturation of the categories developed.

A primary dataset was created of television ads for each of the top two candidates running in all 435 House of Representatives elections from the WMP. Only television ads were included for analysis despite the qualitative sample containing Facebook ads, because television ads are an important part of political campaigns and contain a strong visual element.

To test the reliability and internal validity of the coding scheme, an *intercoder reliability sample* was created by randomly sampling 25 ads from the primary dataset, which were coded by the author and a second coder, as described in more detail below.

Table 1 Data samples used in the paper

Element of research process	Data source	Sample	Referenced as
Coding scheme development	US House of Representatives 2020, all 435 races. Stratified into three equal-sized groups: the third closest races, the third least close races, and the third middle races, depending on the percentage margin of victory.	Four races randomly sampled from each group (24 candidates), and two television and social media adverts sampled per candidate. 182 total images screenshotted Sweep 1: 134 images (78 from 23 television ads, 56 from 56 Facebook ads) Sweep 2: 48 images (30 from 5 television ads, 18 from 7 Facebook ads)	Qualitative sample
Primary dataset for which samples can be taken	Top two candidates running in all 435 districts, US House of Representatives 2020 election. Television ads running from 1st September to 3rd November 2020. Primaries excluded.	For each candidate, two weeks randomly sampled, and one ad for each week sampled proportionately to the estimated amount of money spent on the ad slot (variable taken from WMP).	Primary dataset
Intercoder reliability	Main dataset of ads.	Random sample of 25 ads. Videos screenshotted every two seconds resulting in 375 images.	Intercoder reliability sample
Proof of concept	Main dataset of ads.	125 unique ads from candidates running in 3 most populous states, California, Texas and Florida who had ads in the period. Videos screenshotted every 2 seconds resulting in 2480 images.	Proof of concept sample

Finally, to evaluate the application of the schema, the two ads for candidates running in the three most populous states (California, Texas, and Florida) were taken for the *proof of concept sample* from the primary dataset and coded. This final sample was created to apply the schema to a larger set of ads to evaluate what new information the inclusion of visual and indirect appeals provides compared with textual-only appeals.

Coding Scheme Development

The coding scheme categories were developed through an inductive exploration of the qualitative sample (Table 1), analyzing different ads to identify patterns and letting categories emerge (Thomas, 2006). I watched each ad in the first sweep of the qualitative sample for instances where a candidate appeared to make a group-based appeal, and screenshotted and uploaded these images to NVivo. I evaluated why I believed an image contained an appeal, and annotated the image to note which features led me to believe this. This general inductive method shares similarities with visual discourse analysis, which emphasizes viewing the visual as a whole (Albers, 2013), however, instead of uncovering the discourse emerging within the sample, features indicating potential indirect group-based appeals were assessed. Following this approach, uses of groups were noted, and categories were created. I then returned to watch the full ads again to ensure that all categories were applied, to ensure saturation and that no features were missed.

This exploration of the qualitative sample revealed that many groups could be signaled in one shot, some more strongly than others, and the coding scheme was therefore developed to take account of this. For example, some ads contained shots of the candidate talking to a group of people, and from this image the demographic characteristics of age, gender and race of each person may be assumed by the viewer. In other cases, people in an image were shown in clothes which signaled occupation, such as a nurse's uniform or symbols of the military, and in this way, many groups could be signaled in one shot. As a result, the coding scheme was developed to code group attributes, such as gender, occupation and household position, through a variety of different cues, which can then be aggregated post-coding to gauge the overall group-based appeal. The coding scheme presented below uses examples from this qualitative approach to demonstrate why categories were incorporated and provides examples from this sample.

Moffitt's (2022) study was used as a guide for incorporating demographic characteristics of age, gender and race into the coding scheme. Moffitt (2022) codes for the majority characteristics within an image. In the case of gender, this uses the categories 'majority feminine in appearance', 'majority masculine in appearance', 'mix/balance' and 'unsure/difficult to discern' (Moffitt, 2022). This structure was used as a guide to capture where demographics were signaled, and additional types of group features included as categories, as outlined in the next subsection.

The Coding Scheme

As a result of this process, the following coding scheme was created as a method to capture multimodal group-based appeals. The coding scheme has three parts to capture 1) *what* group attributes are coded for, 2) *how* these attributes are cued, and 3) *how* group members are visually presented. The coding scheme applies to a still image as a unit of analysis.

As outlined above, the coding scheme evaluates how individual group attributes are cued, which can then be reaggregated post-coding. Table 2 depicts these attributes, and the subcategories that are coded for. This scheme proposes that the following group attributes can be cued in either their visual, verbal or textual content: *age, gender, race/ethnicity, occupation, industry, wealth/income, sexuality, religion, disability, health, partisan, recreational activity, household position,* and *ideological group*. These attributes are coded for people who are not the candidate or another politician. Where these attributes are not present, NA is recorded. The benefit of including group attributes separately is that groups covering multiple identities can be coded systematically. For example, the phrase 'working families' would be coded under both occupation and household position. If the family shown in the image is Latino, race can then be coded as 'demographics of person/people'. The image was taken as the unit of analysis, the scheme could be applied to each person/group member in an image for more granular data, and the categories changed from 'majority feminine in appearance' to 'feminine in appearance' for example.

Capturing Group-Based Appeals: Baseline, Indirect, and Direct Appeals

Table 3 shows *how* the group attributes presented in Table 2 can be measured across multimodal indicators, starting with the data that is captured by existing studies.

Direct Appeals

Direct appeals indicate the data captured by existing methods of measuring explicit mentions of groups in text or verbal aspects: *explicit mention of attribute in voiceover, text*, or *caption*. For direct group-based appeals, these terms are recorded as an appeal when they refer to a collective of people who are not politicians or public figures. For example, the phrase 'protecting pre-existing conditions' would not be recorded as a direct appeal, but 'protecting people with

Table 2 Group attributes that can be cued through multi-modal group cues and their subcategories

Attribute	Attribute categories	Rationale
Age	Baby/small child (0–3) Child (4–12) Teenager (13–19) Adult (20–40) Adult (41–64) Retirement (65+) Unsure/difficult to discern	
Gender	Majority masculine in appearance Majority feminine in appearance Mix/balance Unsure/difficult to discern	
Race	White Black/African American Asian American American Indian/Alaska Native Native Hawaiian or other Pacific Islander Hispanic/Latinx Unsure/difficult to discern Mix/balance	Moffitt (2022) keeps racial categories broad in his coding scheme using the categories ‘majority white’ and ‘majority non-white’ to avoid making problematic assumptions. However, conflating distinct groups arguably erases important distinctions between people that would be more meaningful to them, and therefore this schema uses distinct categories for race and ethnicity.[a] Importantly, the researcher is not making claims about what a person’s identity is, but what the intended audience may assume it is from watching the ad.
Occupation	What occupations are depicted? (open text box answer)	
Industry	What industry/industries are the people linked to? (open text box answer)	
Sexuality	Heterosexual Homosexual Bisexual Other (open text box answer)	For example, same sex parents in an image would be coded as homosexual.
Religion	Christian Muslim Hindu Sikh Jewish Buddhist Atheist Other	

Table 2 (continued)

Attribute	Attribute categories	Rationale
Disability	Does the image show a person who appears visibly disabled? (yes/no)	
Health	Is there a person with assumed health issues? (yes/no)	For example, if the ad talks about people with health issues, or a person is shown is hospital.
Recreational activity	Are they taking part in a recreational activity? Yes/no (open text box answer)	For example, is a person in the ad engaging in hobbies or recreational activities not related to their occupation
Household position	Parents Children Grandparents Family/mix Unsure/difficult to discern	
Partisanship	Democrat Republican Party switcher (i.e., ex-Republican/ex-Democrat) Non-partisan/independent Other party (e.g., Green, Libertarian, Socialist) Unsure/difficult to discern	This category was added because both Republican and Democrat candidates used examples of party switchers explaining why they no longer support the opposing party. Visually, these cues could include party logos and symbols.
Ideological group	What is the ideological group reference? (open text box answer)	Added as a category due to the consistent use of terms such as 'radical leftists' and 'radicals' by Republican candidates
Candidate association with group	Strong association Broad/assumed association Broad/assumed disassociation Strong disassociation Neither Unsure	
Opponent association with group	Strong association Broad/assumed association Broad/assumed disassociation Strong disassociation Neither Unsure	

[a] More importantly, as a white, British researcher there is a reasonable question of whether I will make such coding decisions in the same way as the target audience. As discussed above, I was conscious to be aware of the different context and research cues such as locations and activities.

Table 3 How direct and indirect group-based appeals are cued

Group cue	Indicator	Captured by existing methods?
Direct	Explicit mention of group/attribute in voiceover Explicit mention of group/attribute in text Explicit mention of attribute in caption	Yes
Indirect	Historical context of person Characteristic accentuated by activity Setting indicates group Symbol indicates group Clothing indicates group Issue of ad indicates group Inferred from voiceover Inferred from text Inferred/would be assumed by viewer Language	Some indicators explored
Baseline	Demographics of person/people	Captured by Moffitt (2022) under the framing of showing who populists represent rather than group-based appeals

pre-existing conditions' would. Similarly, references were recorded as ideological groups when they refer to a collective of people. For example, 'radicals' was included as a group cue when referring to people such as protestors or citizens with socialist beliefs, but not when used to describe an opponent or other politicians. In the qualitative sample, candidates did occasionally use group terms to refer to themselves, such as referencing that they used to be a 'doctor' or talking about their 'family'. These were not counted as direct appeals because they emphasize candidate characteristics rather than a generic grouping of people.

Indirect Appeals

Indirect appeals are instances where a group attribute is implied through visual or verbal cues without directly mentioning the group. The coding scheme evaluates whether the group is signaled visually by *characteristics accentuated by activity, historical context of the individual, setting, symbols, clothing, language* (e.g., if the ad is in Spanish)*, issue of ad,* or verbally *inferred from voiceover* or *text*. Similar to how direct appeals are not counted when the candidate is speaking about themselves, indirect appeals are assessed based on people who are not the candidate. For example, if a candidate is wearing a military uniform, this would not be coded as a group, unlike if an ordinary person is wearing the same.

Attributes can be cued by *characteristic accentuated by activity*. For example, Figure 1 shows Republican Minnesota candidate Lacy Johnson talking to the camera while men walk into a barbershop behind him. In the voiceover, Johnson discusses how conversations start in neighborhoods, not government and 'in here'. Black barbershops are important cultural spaces for Black men to talk about their experiences (Mills, 2013). Therefore, by having men walk into a barbershop behind him while he talks, Johnson is accentuating the characteristics of being both African American and a man. A mix of both white and Black men walk into the barbershop, which signals masculinity, seemingly regardless of race.

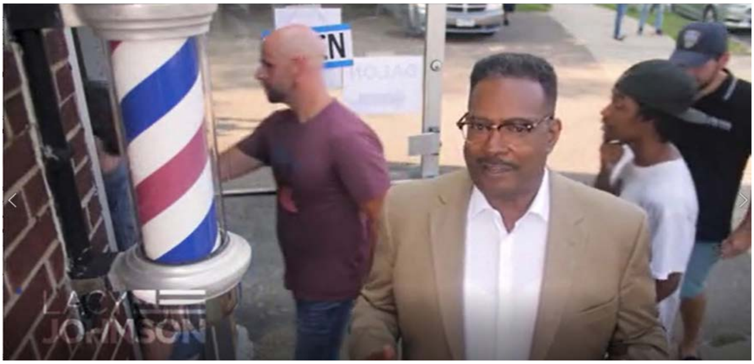


Figure 1 Screenshot from Lacy Johnson's (Minnesota 5) ad 'Breaks my heart'

The *historical context of the person* can cue a demographic group. For example, Ilhan Omar's television ad 'Broken' (Figure 2) shows a photograph of George Floyd taped to a bus stop. George Floyd was murdered by a police officer in Omar's district of Minneapolis, prompting protests against racism and police brutality (Taylor, 2021). The image of George Floyd therefore connects to a broader movement against racism, which very much connects to the demographics of being Black, and particularly male in this instance.

The *setting* of a group was noted where it indicated a group, and when there were people (non-politicians) in the image. Some ads did include settings which could indicate a group, such as showing a field of wheat. On one hand, this could indicate a group-based appeal because of the implied link to farmers. However, it could also indicate farming as an issue, or values of rural life more generally. Because of this ambiguity, only images with people included are coded as 'setting' under the group-based appeals schema.

The *clothing* of a group was again noted if it indicated or implied that the person was a member of a group. A person wearing military clothing, or army camouflage would be recorded as clothing implying a group. *Symbols* were again

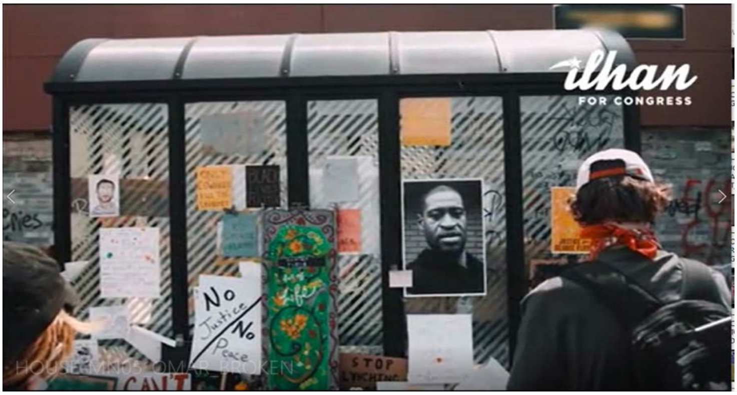


Figure 2 Screenshot from Ilhan Omar's (Minnesota 5) television ad 'Broken'

noted as implying a group where there was an identifiable and recognizable symbol, for example the LGBTQ, Irish flag and military symbols.

Textually, a group can be *inferred through the voiceover, text, or the name of the ad*. Finally, the category *inferred/would be assumed by viewer* captures instances where an identity may be inferred or assumed. For example, a person talking to the candidate may be assumed to be American.

Baseline Appeals

A baseline appeal is defined when the characteristics of age, gender and race are cued through the visible/assumed demographics of the people shown in the ad, not including indirect indicators listed above. A viewer might make inferences about a person's age, gender, and race, and therefore these characteristics cued through *demographics* are coded as a 'baseline' cue towards that group rather than an indirect appeal. For example, in Figure 3 Texas candidate Gina Oritz Jones is talking to a white woman. Here, the demographic characteristics of age (65+), gender (majority feminine in appearance) and race (white) can be coded, and there are no further indicators of other group attributes. It is difficult to conceptualize this image as a group-based appeal in itself, yet the demographics of people included is revealing when ads are evaluated systematically. Therefore, instead of constituting an indirect appeal, these baseline appeals are coded where demographics are coded for and convey information, yet do not signal a group through any cues other than the visible characteristics.

Social Semiotics

Finally, the coding scheme captures social semiotic features around the presentation of group members as outlined in the literature review. These features build upon the previous categories to gain a deeper understanding about how a group is portrayed.

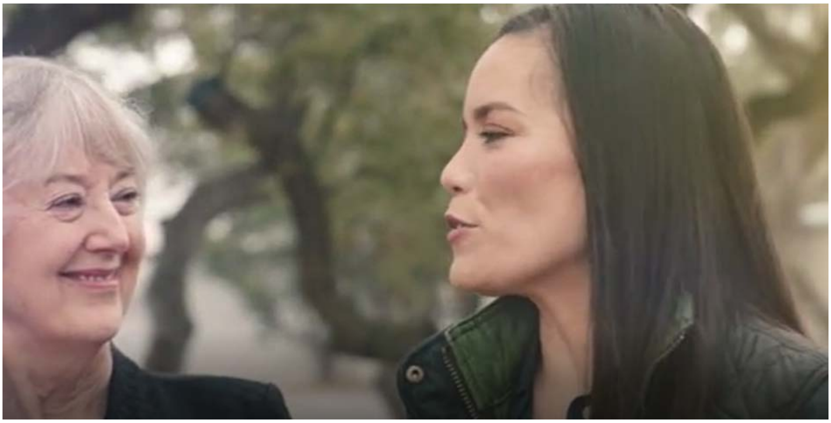


Figure 3 Screenshot from Gina Oritz Jones's (Democrat, Texas 23rd District) ad 'Gina vs Tony

Gaze captures whether the group member is looking at the *camera, candidate, 'stock imagery' (*looking neither at the camera nor candidate), *face obscured, mix/ balance,* or *unsure/difficult to discern*. I have termed the phrase 'stock imagery' where a person is engaging with neither the camera nor candidate as in Figure 4.

When the gaze is at the camera, it demands more attention from the viewer and is therefore more powerful (Kress & Van Leeuwen, 1996). *Shot type* addresses whether the image of the group member of person is *close-up, medium/torso,* or *long shot*. Close-ups can be polarizing in that a close personal distance is acceptable when we are comfortable with the person, but aggressive if not (Kress & van Leeuwen, 1996). *Color tone* evaluates whether the image is *black and white, sepia/ low saturation, normal* or *other.* In the US, black and white color tone often conveys fear, and therefore may indicate negative campaigning or disassociating from an outgroup (Gorn et al., 1997).

Facial expression captures the emotions: *happy, sad, anger, disgust, calm, surprised, confused, fear, neutral,* and an open-ended text-box option for '*other*', following Bossetta and Schmøkel's (2023) categorization.

In the qualitative sample, the social semiotic indicators suggest that the emotional intensity of group-based appeals can be increased with visual features. Black and white color toning was used in ads disassociating from groups to create fear around an outgroup. The connection between black and white and fear is likely to be both universal (in the human physiology of reacting to color) and culturally specific in how it has been used in the US (Gorn et al., 1997). Genevieve Collins uses black and white in her negative ad (Figure 5) of a protestor alongside the message that 'Dallas radicals' are trying to defund the police. The protestor's face is obscured in the edited image, overlaid with the image of Collins' opponent. The black and white color tone therefore conveys fear, and the

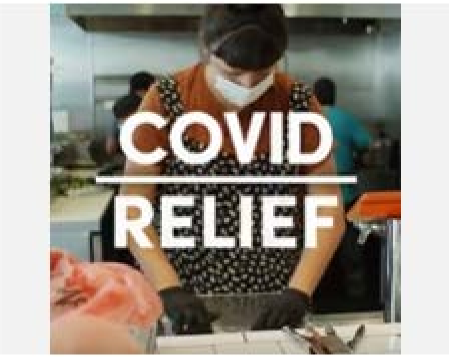


Figure 4 Screenshot from a Facebook ad by Gina Oritz Jones (Democrat, Texas 23rd District)

red text stands out in the image. The fact that both the protestor and the opponent's faces are obscured by the mask creates fear and uncertainty.

Collins' ad demonstrates how multiple types of groups can be used within one image. The ad tells us a clear narrative of a villain (Dallas radicals), and a victim (the police). The phrase 'Dallas radicals' is an example of an ideological group, and this phrase is emblematic of the ideological phrases Republican candidates used across their ads in the qualitative sample. However, the fact that the protestor is not white adds a demographic aspect to the group being dissociated from and implicit racial messaging. Collins herself is white, and the fact that she uses a non-white woman overlaid with her African American opponent implies a racial message, while the editing and use of black and white creates fear.

Intercoder Reliability and Proof of Concept

To test the reliability and internal validity of the coding scheme, a second coder was included for a random sample of 25 ads (see Table 1). Videos were screenshotted every 2 seconds and the resulting images taken as the unit of analysis, with a typical 30-second ad producing 15 images, creating a final sample of 375 images. Coding the ads in this way ensures that the verbal content can be matched to the visual and captures how frequently groups appear across an ad. The second coder was trained and provided with guidance and examples on the scheme. Videos were watched for context when coding the individual images where required. Auto-captioning software was used to automatically add subtitles to the videos to ensure that verbal information was captured. Intercoder reliability tests were conducted to calculate percentage agreement and Cohen's kappa scores, which is perceived to be the best choice when the distribution of categories is not expected to be equal (Di Eugenio & Glass, 2004).

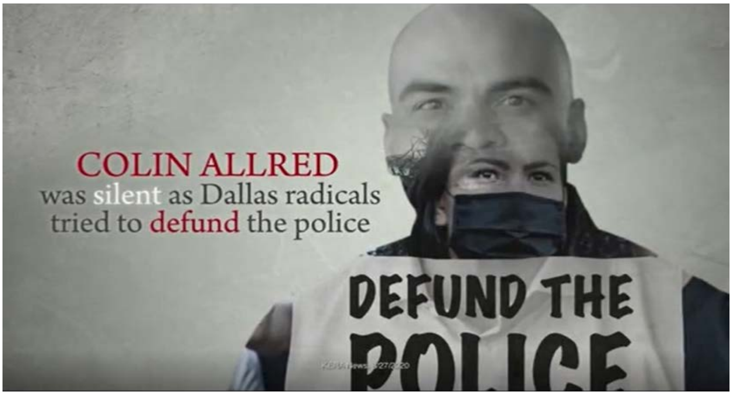


Figure 5 Screenshot from Genevieve Collins' (Texas 32) ad 'The real Colin'

Finally, the scheme was applied to a larger proof of concept sample (see Table 1) of candidates running in the three most populous states using the same methodology outlined above. Some television images were too blurry or dark to evaluate and 199 were removed, leaving 2281 coded screenshots.

Results

Intercoder Reliability

Percentage agreement between the two coders on whether an image contained group cues as outline in the scheme above was generally very good, as shown in Figure 6, as all but three agreement scores are over 80%.

Figure 7 depicts the Cohen's kappa scores for the presence of group cues by type. Baseline appeals cued through the visible demographics of people included in images have good scores above .75. Low kappa scores (< .20) were recorded for direct appeals mentioning household position, ideology, occupation and industry for three reasons. Firstly, there were differences in how coders coded explicit mentions. The schema presented in this paper posits that group-based appeals are not coded where a candidate talks about themselves, however this was not equally applied. For instance, in one ad, a candidate discussed their previous career as a 'midwife', which was coded as a missing value by Coder 1 and occupational by Coder 2, emphasizing the need for clarity amongst coders on the definition of a group-based appeal. As a result, indirect cues are not less reliable than explicit text-based mentions, suggesting against general perceptions that visual cues are more subjective.

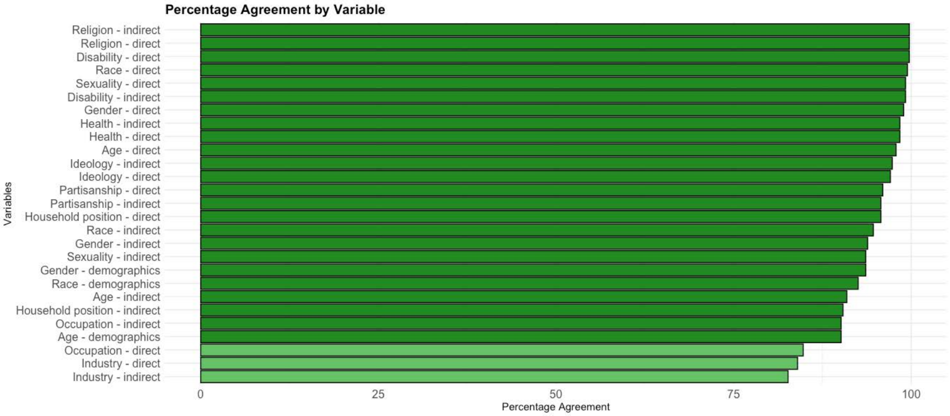


Figure 6 Percentage agreement between two coders on whether an ad contains a baseline, indirect and/or direct group-based appeal. Dark green shows agreement of over 80% and light green over 75%.

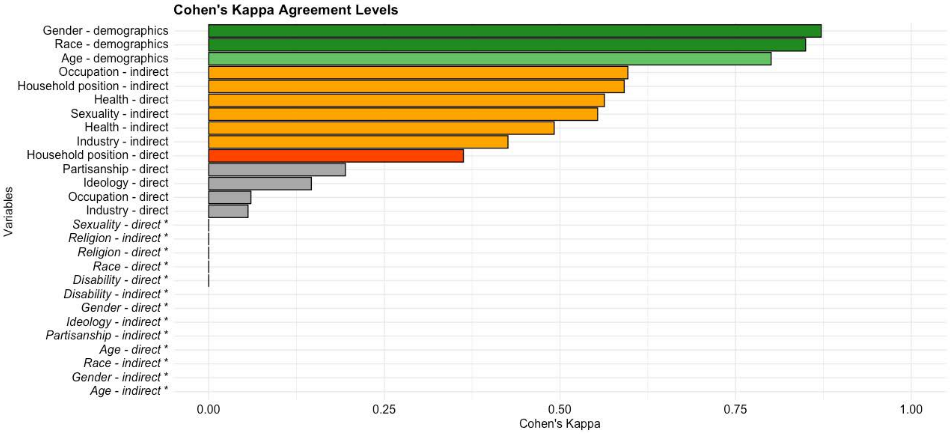


Figure 7 Cohen’s kappa scores for the two coders on presence of a baseline, indirect, and/or direct group-based appeal. Dark green shows very good scores of .81–1.00; light green for substantial scores of .61–.80; orange for moderate scores .41–.60; red for fair scores of .21–.40; and grey for slight 0–.20. Variables marked with an asterisk indicate low prevalence of <16 instances (6.5%).

Secondly, there were differences in the coding of indirect cues relating to a candidate. For instance, in one ad, the second coder consistently coded demographics of the candidate, alongside additional indirect cues relating to the candidate, where the first coder logged this as a missing value. This could be explained by the selection of the House of Representatives, where candidates are less recognizable.[1] To improve reliability in future, researchers and coders should visually identify the candidate before beginning the coding. Researchers could provide images of the candidate along with the ad data, particularly in instances where the candidate is less well-known.

Thirdly, Cohen's kappa is affected by skewed data, such as low prevalence, which may lower scores (Di Eugenio & Glass, 2004; Viera & Garrett, 2005). This is indicated in Figure 7 where variables in grey (marked *) were used in fewer than 16 instances, with 6 categories being used fewer than 6 times. These categories could still be included in future applications of the coding scheme, as the percentage agreement was generally very good.

Table 4 Cohen's kappa scores for social semiotic features of visual group-based appeals

Variable	Kappa
Gaze	.50
Image type	.49
Color tone	.46
Facial expression	.33

The categories capturing social semiotic characteristics image type, gaze, color tone and facial expression had generally moderate Cohen's kappa scores (.41–.60) (Table 4). This is perhaps to be expected from categories which are more subjective and future use of the coding scheme could therefore omit them from a quantitative application of the scheme. These features may be better explored qualitatively to understand the depiction of groups in specific contexts as opposed to scaling up.

Proof of Concept: Application of the Coding Scheme

Table 5 shows the frequency of baseline, indirect and direct appeals by political party in the television ads of candidates running for election in the three most populous US states (Table 1, proof of concept sample). The rows show the type of appeal, and what proportion of images screenshotted from television ads by Democrat and Republicans contain each appeal. Direct appeals are explicit mentions of groups that are captured by traditional methods and constitute 13.5% of images screenshotted every 2 seconds in television ads. A further

[1] The second coder was a PhD student selected due to knowledge of US elections; however, it is unlikely that they would recognize all candidates in the sample. The coding instruction sheet does advise to look up an image of the candidate before coding the particular ad. See appendix for future recommendations.

10.6% of all images contain a combination of both direct and indirect appeals. In these instances, existing methods capture only the direct aspect. Indirect-only appeals constitute 15.7% of all images captured from television ads, representing the new data captured by incorporating visual and more implicit references. Finally, baseline appeals, where age, gender and/or race are assumed through *only* the demographics of people included in images with no other kind of appeal present, constitute 4% of all images screenshotted from ads of candidates from both parties.

Table 5 The proportion of images (taken from ads every 2 seconds, $N = 2{,}281$) containing baseline (demographic cues only), indirect, and direct appeals on group attributes by party

Appeals	Democrat		Republican		Both	
Baseline	56	5.10%	34	2.88%	90	3.95%
Indirect	196	17.83%	163	13.79%	359	15.74%
Indirect and direct	131	11.92%	111	9.39%	242	10.61%
Direct	169	15.38%	139	11.76%	308	13.50%
No appeal	547	49.77%	735	62.18%	1,282	56.20%
Total	1,099	100.00%	1,182	100.00%	2,281	100.00%

Table 6 further decomposes the appeals to show the frequency of each sub-type of cue by party. The visible demographic characteristics (age, gender, race) can be inferred in 35.6% of the total number of images taken from ads. This seems surprising given that only 4% of images contained a baseline appeal as shown in Table 5. The low level of baseline appeals is therefore not explained by demographics rarely being visible, but by other formats of group cue being alongside these.

Of the indirect cues towards groups, clothing (11.1%), setting (12.7%), characteristic accentuated by activity (9.8%) and inferred from voiceover (9.8%) are the most frequently used. Clothing was often used for occupation and industry, cueing groups of workers. Therefore, indirect appeals are primarily derived from the person included in the image and what they are wearing, doing and their location. Significantly, these are primarily visual cues towards groups.

Table 7 shows the frequency of types of appeal by group attribute. The rows list the group attributes captured by the coding scheme and the columns how the group is cued. Industry and occupation are the most likely attributes to be cued indirectly, with 14.7% and 11.4% of the total images containing indirect cues towards these attributes. Industry and occupation are highly related, as showing workers such as nurses often cues both aspects. In the sample this was largely driven by campaigns showing workers, particularly manual, industrial and construction workers, as well as small business owners working in hospitality or retail. Industry and occupation are therefore the most likely group attributes to be missed from existing text-based methods.

Table 6 Frequency of use of group cues by party

	How different cues are used	Democrat	Republican	Total references across groups	Number of images containing appeal
Baseline	Demographics of people	1,305	995	2,300	812 (35.59%)
Indirect	Clothing indicates/ accentuates attribute	212	273	485	253 (11.09%)
	Setting indicates/ accentuates attribute	259	208	467	291 (12.76%)
	Characteristic indicated/ accentuated by activity	224	168	392	224 (9.82%)
	Symbol indicates/ accentuates attribute	26	127	153	95 (4.16%)
	Issue of ad	170	135	305	192 (8.41%)
	Historical context of person	171	110	281	114 (3.97%)
	Language	93	82	175	175 (7.67%)
	Inferred from voiceover	159	117	276	224 (9.82%)
	Inferred from text	114	107	221	172 (7.54%)
	Inferred from ad name	15	8	23	23 (1.01%)
Direct	Explicit mention of attribute in text	157	181	338	269 (11.79%)
	Explicit mention of attribute in voiceover	168	152	320	317 (13.90%)
	Explicit mention of attribute in ad name	3	8	11	11 (0.48%)
	Total				2,281

Household position is cued indirectly in 6.4% of images, with candidates showing images of families, parents and children. Other group attributes including wealth, health, sexuality, disability, partisanship and religion are rarely cued in any format. Of the 45 ideological group-based appeals, 42 of these were made by Republican candidates appealing directly against groups such as ‘radicals’ and ‘leftists’ and indirectly showing BLM protestors and symbols.

Table 7 Frequency of indirect and direct cues used for different group attributes.

	Indirect	Indirect and direct	Direct	None	
Industry	351 14.72%	35 1.53%	71 3.11%	1,824 79.96%	2,281
Occupation	261 11.44%	78 3.42%	117 5.13%	1,825 80.01%	2,281
Household position	146 6.40%	13 0.57%	36 1.58%	2,086 91.45%	2,281
Health	47 2.06%	14 0.61%	23 1.01%	2,197 96.32%	2,281
Sexuality	47 2.06%	0 0.00%	0 0.00%	2,234 97.94%	2,281
Ideological	26 1.14%	3 0.13%	16 0.70%	2,236 98.03%	2,281
Disability	37 1.62%	0 0.00%	0 0.00%	2,244 98.38%	2,281
Partisan	17 0.75%	0 0.00%	6 0.26%	2,258 98.99%	2,281
Religion	15 0.66%	0 0.00%	2 0.09%	2,264 99.25%	2,281

Household Position As a Case Study

To demonstrate how the inclusion of indirect appeals impacts our conclusions, household position was selected as a case study. Figure 8 shows the proportion of images cueing household position by party, appeal type (direct or indirect) and the type of family member cued. If we look only at direct appeals in the darkest shades of red and blue, Republicans explicitly appeal to parents and children in 26% and 22% of images cueing household position, compared with no direct appeals by Democrats to these groups. Democrats are more likely to mention ‘families’ explicitly in 38% of images cueing household position, compared with 6% of Republican images. Neither party directly appeals to grandparents, suggesting a lack of interest.

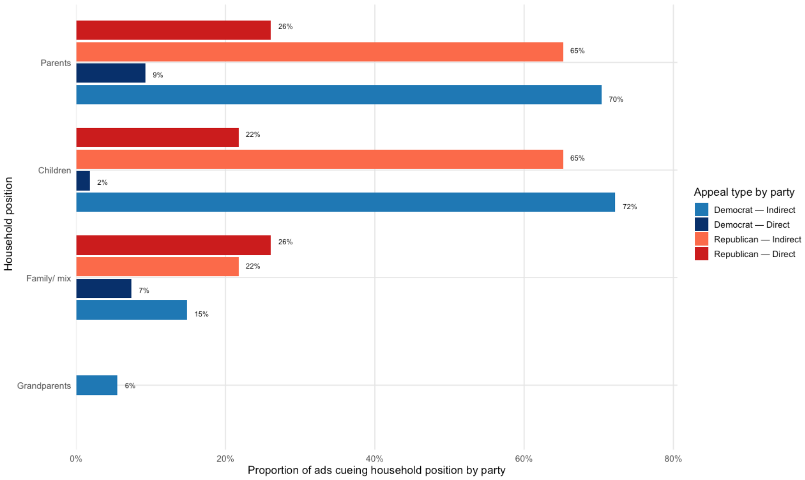


Figure 8 Appeals cueing household position by party and appeal type

The picture looks different when we include indirect appeals as a measure, and the two parties start to look more similar. Despite Republicans explicitly mentioning parents and children more than Democrats, Democrats are slightly more likely to signal these groups indirectly, with 72% and 70% of images cueing household position containing indirect appeals to parents and children respectively. Republicans mention families explicitly more often, and are more likely to signal them indirectly (22%) than Democrats (15%). Democrats appeal to grandparents in 6% of shots cueing household position, indicating that this group would be missed if direct appeals were studied alone. Including indirect appeals therefore alters our conclusions from Republicans appealing to all household positions more than Democrats, to Republicans appealing to these groups more directly, while Democrats do so more indirectly. Furthermore, using traditional methods we would conclude that neither party appeals to grandparents, whereas inclusion of indirect appeals shows Democrats do appeal to this group, albeit rarely. Not including indirect measures therefore biases our understanding of how the two parties appeal. Interestingly, the finding that Republicans appeal more explicitly to household position counters the perception that Democrats are more interested in social groups than Republicans (Grossmann & Hopkins, 2016) and therefore suggests that party differences in this type of campaigning are worth further exploration.

Indirect appeals can occur either 1) with direct appeals to provide additional meaning and context to an appeal, or 2) alone to constitute a group appeal without an explicit group reference, and Figures 9 and 10 are taken from the qualitative analysis to demonstrate this. Figure 9 depicts both a direct appeal towards ‘working families’ and an indirect appeal towards blue-collar workers, as repre-

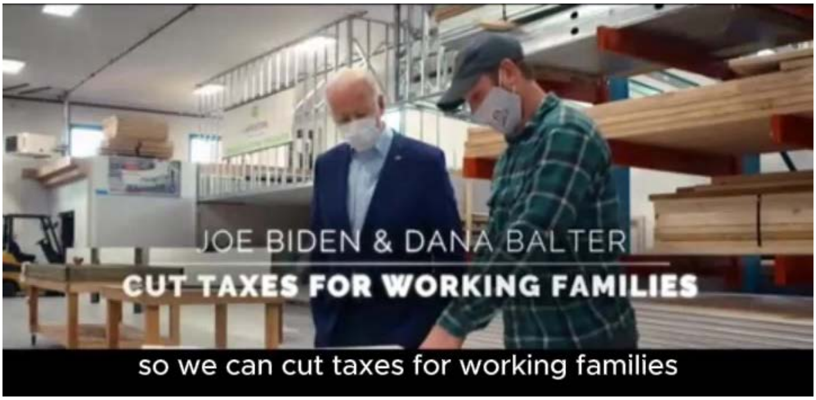

JOE BIDEN & DANA BALTER
CUT TAXES FOR WORKING FAMILIES
so we can cut taxes for working families

Figure 9 Screenshot from Dana Balter's (New York 24) ad 'The last four years'

sented through a white man. Using traditional methodologies, only the phrase 'working families' would be captured, exemplifying how indirect appeals can provide additional context as to who is inferred to be part of this group. In contrast, Figure 10 depicts an instance where the appeal is only indirect because using text-based methods no group would be detected in this image. Without incorporating the visual, this scene is a policy statement. Including the visual indicates a young Black family, with a small child with health conditions (as symbolized through the characteristic accentuated by activity of using an asthma inhaler), an appeal signaling the kinds of families the candidate seeks to represent. It is therefore important to include indirect appeals both to capture the additional meaning and context of direct appeals and to ensure that indirect appeals that occur without textual references are included in the analysis.

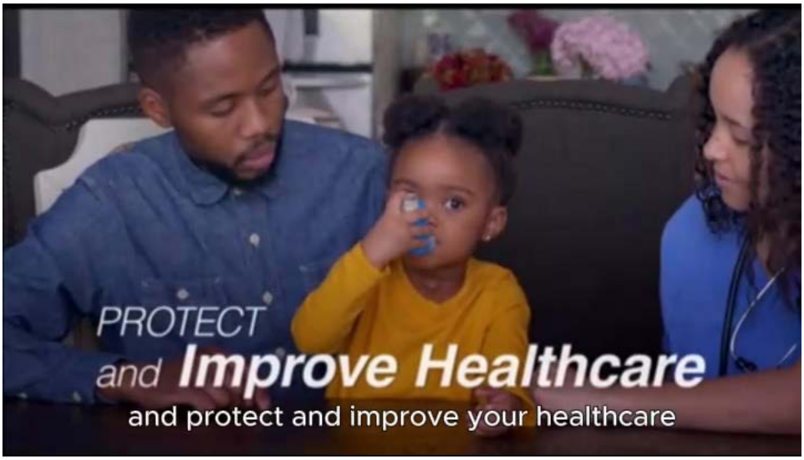


Figure 10 Screenshot from Charlie Crist's (Florida 13) ad 'Called a Lot'

Discussion and Conclusion

Existing studies of group-based appeals focus primarily on textual mentions of groups, and are not equipped to incorporate more indirect, and particularly visual cues towards groups. This paper has proposed and briefly evaluated a novel schema to conduct content analyses incorporating visual and indirect cues.

This paper has demonstrated the importance of incorporating visual cues into methodologies of group-based appeals for two reasons. Firstly, candidates running in the three most populous states for the 2020 US House of Representatives elections engage in indirect appeals at similar levels to direct appeals. Furthermore, indirect appeals are conducted alongside direct appeals in 10.6% of images screenshotted from television ads. The exploration of household position as a case study demonstrates that indirect appeals alone can convey information about who a candidate seeks to represent, and combined with a direct appeal provide more context as to who is meant by text-based groups. Capturing visual data therefore provides both a greater *coverage* of appeals conducted and deepens understanding and meaning with further insights as to who is included in groups. Secondly, candidates appear to use different cue types for different groups, and capturing only direct appeals may lead to skewed findings in terms of *which* groups parties appeal to. Industry and occupation in particular are more likely to be cued indirectly, and therefore capturing only textual appeals to these groups may bias results. In the case of household position, the inclusion of indirect appeals alters our conclusions to reveal that Republicans are more interested in families and Democrats in grandparents than direct only appeals would suggest.

This study has limitations which could be addressed through future research. Firstly, some categories in the scheme scored low on intercoder reliability. Particularly, social semiotic categories such as gaze and facial expression had low reliability, and such features may be better explored qualitatively to understand how members of groups are visually depicted and positioned. A second reason for low intercoder reliability for some variables was due to low prevalence of many categories, and the scheme should therefore be tested on a larger scale. Secondly, the schema was tested only on television ads despite being developed from a qualitative sample of ads from television and social media. Television ads were selected because they are widely used in US elections and they are inherently visual. Future research however could apply the schema to social media to evaluate media differences in the use of group-based appeals and to explore reliability across mediums. Finally, this study uses US ads while most group-based appeals studies focus on European political communication. This brings benefits in broadening the literature to include North America, however could be applied to other countries to evaluate whether indirect appeals are as widely used beyond the US context.

Group appeals are important aspects of political campaigns, yet if we only measure direct mentions, we do not get an accurate picture of which groups exactly are being appealed to. As a result, we would get biased results as to which group identities are being activated and made salient, which is particularly important given than visual signals are more easily processed by audiences and may be more emotive. Inclusion of indirect appeals data may reduce bias in measures of both the *prevalence* of group-based appeals and *which* groups political parties appeal to in modern political discourse. I hope that this paper adds to the accurate measurement of group-based appeals in political advertisements moving forward.

References

Albers, P. (2013). Visual discourse analysis. In P. Albers, T. Holbrook & A. S. Flint (Eds.), *New methods of literacy research* (pp. 85–97). Routledge.

Araujo, T., Lock, I., & van de Velde, B. (2020). Automated visual content analysis (AVCA) in communication research: A protocol for large scale image classification with pre-trained computer vision models. *Communication Methods and Measures, 14*(4), 239–265. https://doi.org/10.1080/19312458.2020.1810648

Barlas, P., Kyriakou, K., Guest, O., Kleanthous, S., & Otterbacher, J. (2021). To "see" is to stereotype: Image tagging algorithms, gender recognition, and the accuracy-fairness trade-off. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW3), 1–31. https://doi.org/10.1145/3432931

Benoit, W. L. (2019). A functional analysis of visual and verbal symbols in presidential campaign posters, 1828–2012: Functional analysis of visual and verbal symbols. *Presidential Studies Quarterly, 49*(1), 4–22. https://doi.org/10.1111/psq.12503

Bossetta, M., & Schmøkel, R. (2023). Crossplatform emotions and audience engagement in social media political campaigning: Comparing candidates' Facebook and Instagram images in the 2020 US Election. *Political Communication, 40*(1), 48–68. https://doi.org/10.1080/10584609.2022.2128949

Brader, T. (2005). *Campaigning for hearts and minds: How emotional appeals in political ads work*. University of Chicago Press. https://doi.org/10.7208/9780226788302

Butler, D., & Stokes, D. E. (1969). *Political change in Britain: Forces shaping electoral choice*. Palgrave Macmillan. https://doi.org/10.1007/978-1-349-00140-8

Campbell, A., Converse, P. E., Miller, W. E., & Stokes, D. E. (1980). *The American voter*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/A/bo24047989.html

Coleman, R. (2010). Framing the pictures in our heads: Exploring the framing and agenda-setting effects of visual images. In P. D'Angelo & J. A. Kuypers (Eds.), *Doing news framing analysis: Empirical and theoretical perspectives* (pp. 249–278). Routledge. https://doi.org/10.4324/9780203864463

Conover, P. J. (1988). The role of social groups in political thinking. *British Journal of Political Science, 18*(1), 51–76. https://doi.org/10.1017/S0007123400004956

d'Andrea, C., & Mintz, A. (2019). Studying the live cross-platform circulation of images with computer vision API: An experiment based on a sports media event. *International Journal of Communication, 13*, 1825–1845.

Dean, J. (2019). Sorted for memes and gifs: Visual media and everyday digital politics. *Political Studies Review, 17*(3), 255–266. https://doi.org/10.1177/1478929918807483

Dolinsky, A. O. (2022). Parties' group appeals across time, countries, and communication channels—Examining appeals to social groups via the Parties' Group Appeals Dataset. *Party Politics, 29*(6), 1130–1146. https://doi.org/10.1177/13540688221131982

Eugenio, B. D., & Glass, M. (2004). The kappa statistic: A second look. *Computational Linguistics, 30*(1), 95–101. https://doi.org/10.1162/089120104773633402

Feng, D., & O'Halloran, K. L. (2012). Representing emotive meaning in visual images: A social semiotic approach. *Journal of Pragmatics, 44*(14), 2067–2084. https://doi.org/10.1016/j.pragma.2012.10.003

Ford, R., & Jennings, W. (2020). The changing cleavage politics of Western Europe. *Annual Review of Political Science, 23*(1), 295–314. https://doi.org/10.1146/annurev-polisci-052217-104957

Fowler, E. F., Franz, M. M., Ridout, T. N., Baum, L. M., & Bogucki, C. (2023). *Political advertising in 2020* (Version 1.0) [Datset]. The Wesleyan Media Project, Department of Government, Wesleyan University.

Geboers, M. A., & Van De Wiele, C. T. (2020). Machine vision and social media images: Why hashtags matter. *Social Media + Society, 6*(2). https://doi.org/10.1177/2056305120928485

Gorn, G. J., Chattopadhyay, A., Yi, T., & Dahl, D. W. (1997). Effects of color as an executional cue in advertising: They're in the shade. *Management Science, 43*(10), 1387–1400. https://doi.org/10.1287/mnsc.43.10.1387

Graber, D. A. (2001). *Processing politics: Learning from television in the Internet age*. University of Chicago Press.

Graber, D. A. (2012). *On media: Making sense of politics*. Paradigm Publishers.

Horn, A., Kevins, A., Jensen, C., & van Kersbergen, K. (2021). Political parties and social groups: New perspectives and data on group and policy appeals. *Party Politics, 27*(5), 983–995. https://doi.org/10.1177/1354068820907998

Huber, L. M. (2022). Beyond policy: The use of social group appeals in party communication. *Political Communication, 39*(3), 293–310. https://doi.org/10.1080/10584609.2021.1998264

Huber, L. M., & Dolinsky, A. O. (2023). *How parties shape their relationship with social groups: A roadmap to the study of group-based appeals*. OSF. https://doi.org/10.31219/osf.io/szaqw

Jamieson, K. H. (1992). *Dirty politics: Deception, distraction, and democracy*. Oxford University Press. https://doi.org/10.1093/oso/9780195078541.001.0001

Kaid, L. L., & Johnston, A. (2001). *Videostyle in presidential campaigns: Style and content of televised political advertising*. Praeger.

Kress, G. R., & Van Leeuwen, T. (1996). *Reading images: The grammar of visual design*. Routledge.

Licht, H., & Sczepanski, R. (2024). *Detecting group mentions in political rhetoric. A supervised learning approach*. OSF. https://doi.org/10.31219/osf.io/ufb96

Lipset, S. M., & Rokkan, S. (1967). Cleavage structures, party systems, and voter alignments: An introduction. In S. M. Lipset & S. Rokkan (Eds.), *Party systems and voter alignments: Cross-national perspectives* (pp. 1–64). Free Press.

McIlwain, C., & Caliendo, S. M. (2011). *Race appeal: How candidates invoke race in U. S. political campaigns*. Temple University Press. http://ebookcentral.proquest.com/lib/manchester/detail.action?docID=650405

Mendelberg, T. (2001). *The race card: Campaign strategy, implicit messages, and the norm of equality*. Princeton University Press. https://doi.org/10.1515/9781400889181

Mills, Q. T. (2013). *Cutting along the color line: Black barbers and barber shops in America*. University of Pennsylvania Press. https://doi.org/10.9783/9780812208658

Moffitt, B. (2022). How do populists visually represent 'the people'? A systematic comparative visual content analysis of Donald Trump and Bernie Sanders' Instagram accounts. *The International Journal of Press/Politics, 29*(1), 74–99. https://doi.org/10.1177/19401612221100418

Neumayer, C., & Rossi, L. (2022). Seeing images from conflict through computer vision: Technology, epistemology and humans. In M. Mortensen & A. McCrow-Young (Eds.), *Social media images and conflicts* (pp. 122–133). Routledge. https://doi.org/10.4324/9781003176923

Page, J. T., & Duffy, M. E. (2009). A battle of visions: Dueling images of morality in U.S. political campaign TV ads. *Communication, Culture & Critique, 2*(1), 110–135. https://doi.org/10.1111/j.1753-9137.2008.01031.x

Swigger, N. (2012). What you see is what you get: Drawing inferences from campaign imagery. *Political Communication, 29*(4), 367–386. https://doi.org/10.1080/10584609.2012.722174

Taylor, D. B. (2021, November 5). George Floyd protests: A timeline. *The New York Times*. https://www.nytimes.com/article/george-floyd-protests-timeline.html

Thau, M. (2019). How political parties use group-based appeals: Evidence from Britain 1964–2015. *Political Studies, 67*(1), 63–82. https://doi.org/10.1177/0032321717744495

Thau, M. (2021). The social divisions of politics: How parties' group-based appeals influence social group differences in vote choice. *The Journal of Politics, 83*(2), 675–688. https://doi.org/10.1086/710018

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine, 37*(5), 360–363.

Reflective Appendix

Data Samples and Collection

I originally planned to test the scheme on a large dataset of two television and social media ads for each candidate running in the 2020 US House of Representatives elections.

I used both Facebook and television ads for the qualitative sample to create the schema, but decided to code only television ads for intercoder reliability and proof-of-concept due to time constraints. Firstly, the coding scheme is complex, taking between 1 and 20 minutes to code an image depending on the content of the ad. Secondly, collecting Facebook ads specifically was time-consuming because the API allows researchers to download a maximum of 3 CSV files a day. Instead of downloading a file per candidate, I therefore searched for phrases like 'for congress' or 'for Illinois' and manually extracted the ads in bulk through the image IDs, cross-referencing against the number of ads in the library. Some ads were removed from Facebook for breaching advertising guidelines, therefore potentially biasing the sample. Therefore the WMP television ads were a more comprehensive dataset for testing. Furthermore, I reduced the n to the three most populous states to reduce coding time while still providing a large enough sample (2,480 images from 125 ads) to test the schema.

Method: Developing the Coding Scheme

I planned to evaluate group-based appeals by making a judgment on which type of group was being appealed to (such as economic, lifecycle, religious), and then evaluating in what format the appeal was conducted (text, visual).

When I started looking at the ads however this was often challenging to do visually. Compared to a textual statement like 'farmers', images allow for many individuals with group characteristics to be shown at once, and it was often difficult to pinpoint one or two singular groups being indicated. I then tried to organize appeals by 'types' or 'frames', such as whether a group member appeals directly to the camera or represents an issue (e.g., nurses = healthcare). Again, it was often challenging to make an overall judgment, with multiple types used in the same shot to differing degrees. I finally decided to code individual attributes (e.g. gender, occupation) along with the social semiotic features to capture this aspect of engagement with a group.

Development of the schema was iterative and I removed national identity and migration status as categories because they were largely 'assumed American'/ 'assumed native to US' without additional indicators, and therefore repetitive to code without producing useful insight. For 'disability', I began coding every person as disabled or not, however due to the repetition of coding 'not disabled' in most instances, I decided to only code when disability was visibly present. I removed 'wealth' as it was often difficult to judge and could be inferred through other variables such as occupation. Group cues were added iteratively, with common cues setting, clothing and symbol added early on, and less frequently used categories such as 'historical figure' later.

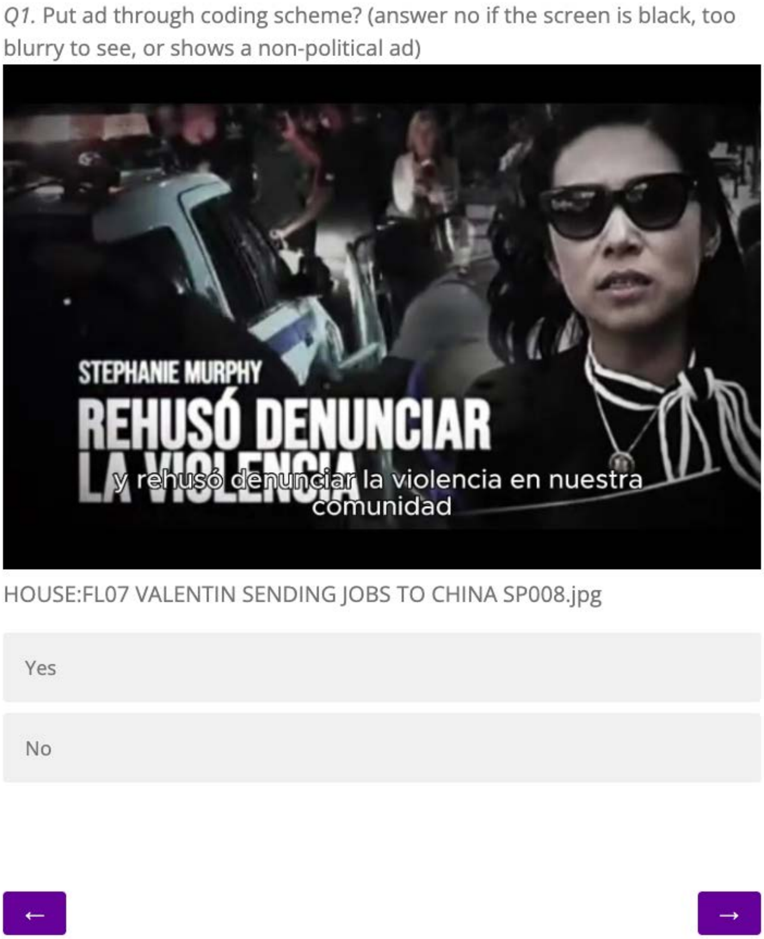
Recommendations

1. The number of variables made coding in a spreadsheet unwieldy and so I uploaded the images to Qualtrics. Qualtrics limits 100 image uploads per survey (it crashes if it goes above this) so I uploaded images in batches and piped the ad title in with the looped image so that coders can view the candidate and ad name for further context (see Figure A1).

2. Do not underestimate how long it will take to code images. At the beginning, I completed around 50–100 images in a day, increasing up to 200 in an 8-hour working day towards the end. Using Qualtrics did make this quicker through an initial filtering question which asks if any of the attributes are present or can be assumed, answering 'none' moves on to the next image

(Figure A2). Part of the initial slowness at the beginning however was part of learning the complexity of the coding scheme. Therefore, I recommend reviewing the coding scheme for clarity and understanding periodically before and during coding.

3. Spend time looking up the candidate and local politicians so that group members are not confused with political figures. This caused a particular challenge for the second coder, who, on multiple occasions, coded the demographics of the candidate, despite coding instructions to only code these features where a person who is not a politician is included. This was likely because House candidates are less likely to be recognized.


STEPHANIE MURPHY
REHUSÓ DENUNCIAR LA VIOLENCIA
y rehusó denunciar la violencia en nuestra comunidad
←
→

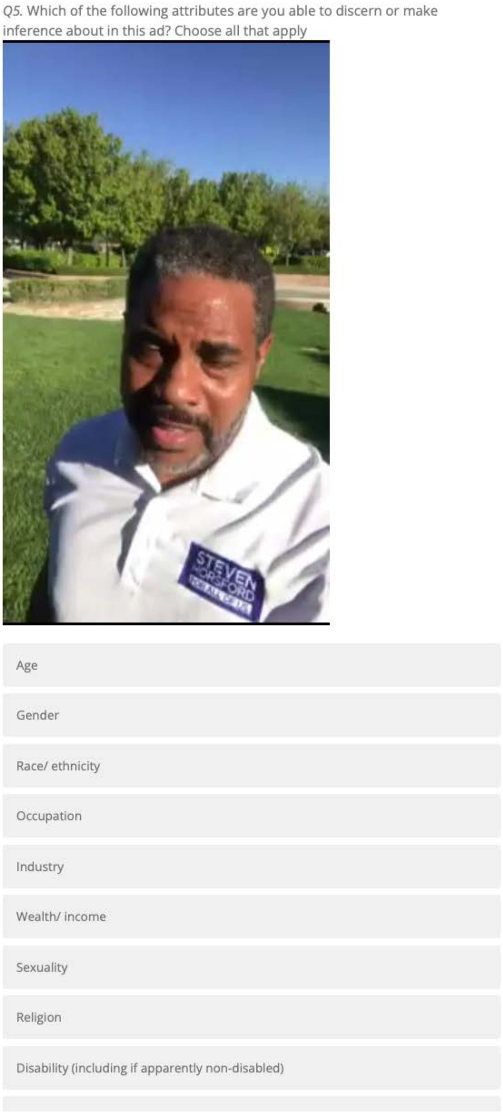
Q1. Put ad through coding scheme? (answer no if the screen is black, too blurry to see, or shows a non-political ad)

HOUSE:FL07 VALENTIN SENDING JOBS TO CHINA SP008.jpg

Yes

No

Figure A1 Screenshot 1 from Qualtrics coding scheme



Q5. Which of the following attributes are you able to discern or make inference about in this ad? Choose all that apply

Age

Gender

Race/ ethnicity

Occupation

Industry

Wealth/ income

Sexuality

Religion

Disability (including if apparently non-disabled)

Figure A2 Screenshot 2 from Qualtrics coding scheme