Original Manuscript

Harnessing Big Data, Hindered by Bias: Evaluating TikTok Research API for Fair and Optimal Social Sciences

Social Science Computer Review
2026, Vol. 0(0) 1–31

DOI: 10.1177/08944393251413277
journals.sagepub.com/home/ssc

Dan Bai[1] **and Yan Gu**[2,3]

Abstract

Digital platforms now serve as crucial archives for analysing societal trends, yet their research APIs pose methodological challenges. This study critically evaluates TikTok Research API through comparative analysis of 6,373 videos from 14 creators in the United States and United Kingdom (2020–2022), contrasting API-derived outputs with manual collection and third-party analytics. The API demonstrated scalability, retrieving more videos than alternative methods and providing 22 variables, including eight unavailable elsewhere. However, limitations were substantial: transcriptions covered about 10% of the content, with more transcripts returned from American male creators. Engagement metrics exhibited inconsistent accuracy across data sources, with the API showing systematically lower view counts but higher comment and share counts compared to manual collection. The number of videos varied depending on sample composition, indicating that small changes in inclusion criteria could shift outcomes disproportionately. These results highlight systematic inconsistencies, showing why multi-method approaches remain necessary despite automation. While TikTok Research API offers valuable scale and ethical compliance, its demographic biases and metadata inconsistencies compromise validity. The study advocates integrated auditing protocols and targeted API refinements to improve representativeness and accuracy in platform-based research.

Keywords

digital platform, TikTok, API, big data, data audit

[1]Institute of Social and Economic Research, University of Essex, Colchester, UK
[2]Department of Psychology, University of Essex, Colchester, UK
[3]Department of Experimental Psychology, University College London, UK

Corresponding Author:
Dan Bai, Institute of Social and Economic Research, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK.
Email: db22181@essex.ac.uk

Introduction

Digital platforms have become integral to contemporary social life, producing vast datasets that can be systematically examined to address pressing research questions. Social media (SM) platforms generate digital traces that offer unprecedented opportunities to study societal trends (Reisach, 2021) and human behaviour across diverse contexts. These data enable researchers to explore phenomena ranging from conflict dynamics (Zeitzoff, 2011) to polarisation in voting behaviour (Sharma et al., 2022), underscoring their relevance to computational social science (Lazer et al., 2021). The dynamic interplay of platform algorithms, user interactions, and societal events necessitates robust open science practices to ensure transparency and reproducibility (Bezjak et al., 2018; Mosnar et al., 2025). Application programming interfaces (APIs) provide structured protocols for accessing these data, facilitating large-scale analyses of networked behaviours (Acker & Kreisberg, 2020; Davidson et al., 2023; Perriam et al., 2020).

Among social media platforms, TikTok has emerged as a focal point for researchers due to its rapid growth and appeal to younger demographics. Access to TikTok's Research API necessitates an endorsement letter, a requirement that has grown common, though the frequency of approvals remains unclear due to limited public data on application outcomes. By late 2023, TikTok expanded its Research API to European countries, offering a more permissive approach compared to platforms like Meta Transparency Centre (2024) and X Developer Platform (2024), which have curtailed academic access (Chang, 2018; TikTok, 2024a, 2024b). TikTok's substantial revenue growth, reaching £875 million by October 2022 (Sweney, 2022), underscores its expanding influence across cultural and political domains (Literat, 2021; Rejeb et al., 2024). The platform's rich, multimodal data, encompassing videos, audio, text, and engagement metrics, support a diverse array of research methodologies, ranging from thematic analysis (Moir, 2023; Lu & Shen, 2023) and digital ethnography (Entrena-Serrano, 2025; Haime & Biddle, 2025; Yang, 2022) to interviews (Klug et al., 2021; Liang, 2021), surveys (Chen & Zhang, 2021; Kirkpatrick & Lawrie, 2024), and machine-learning applications (Agrawal, 2024; Parisi et al., 2023). However, platform-imposed restrictions arising from Terms of Service (ToS), such as data retention, pre-production, post-publication, and risks of indemnity, often constrain research validity and replicability Bak-Coleman, 2023; Venkatagiri, 2023; related items in Appendix 1).

Additionally, significant methodological and ethical challenges continue to limit the validity of societal insights derived from TikTok research. Small sample sizes, often the result of manual scraping (e.g. Cervi & Divon, 2023), restrict representativeness. Ethical concerns also arise when data collection contravenes TikTok's Terms of Service, whether through manual or automated scraping (Thole, 2022). Unstable API performance can return data inconsistent with webpage information (Ruz et al., 2023), while reliance on third-party tools frequently produces incomplete datasets requiring extensive cleaning and validation (Mimizuka et al., 2025). Meanwhile, despite the widespread use of API-driven data, little is known about potential algorithmic biases favouring certain demographics or about inconsistencies in metadata accuracy. This lack of understanding undermines the ability to assess the generalisability and validity of current acquisition methods, and, by extension, the robustness of findings based upon them.

Existing research highlights persistent technical and ethical challenges in using TikTok's Research API. There are three key limitations: inconsistent data delivery, rate restrictions on video retrieval, and restricted API access for independent researchers, which undermines the quality of societal insights that can be derived from platform data during critical events such as the COVID-19 pandemic. Notably, algorithmic curation and demographic disparities, including gender and cultural skews, introduce biases that call into question the fairness of trend analyses on the platform. Crucially, integrating multi-method approaches, such as combining API data with manual collection or external analytics, offers a means to address these inconsistencies. This is

consistent with calls to enhance the reliability of social media research by triangulating sources, particularly when examining time-bound phenomena such as lockdown periods.

This study critically evaluates TikTok Research API by comparing API-derived data with manual collection and third-party analytics, using a dataset of 6,373 videos (after manual calibration) from 14 U.S. and U.K. creators (2020–2022). Videos were selected for their relevance to social commentary, humour, and public discourse during the COVID-19 period. This study addresses key research questions:

RQ1. To what extent does TikTok's Research API enable effective multimodal data collection for computational social science, and how do its technical limitations and ethical constraints influence the validity of societal insights derived from such data?
RQ2. In what ways do algorithmic and demographic biases within TikTok's Research API affect the fairness of societal trend studies based on its multimodal data?
RQ3. How can integrating multi-method data collection with TikTok Research API mitigate inconsistencies to yield valid and equitable insights into societal trends?

This paper proceeds as follows: it first examines the functionalities of the TikTok Research API, then evaluates its limitations, and finally proposes methodological refinements. The literature review synthesises prior scholarship on social media APIs, covering data collection practices, the affordances and constraints of the TikTok Research API, and associated issues of data quality, governance, and ethics. The methodology outlines data acquisition from three sources, such as API retrieval, paid datasets, and manual calibration, followed by statistical evaluation of data quality across these streams. The findings present the results of the comparative analysis. The discussion then interprets these outcomes considering the research questions, drawing out implications for platform-based data acquisition strategies. The paper concludes with recommendations to strengthen methodological rigour in future social media research.

Literature Review

Data Collection in Computational Social Science

Computational social scientists typically employ one or more of four approaches to data collection: web scraping, commercial analytics services, data donation, and platform-

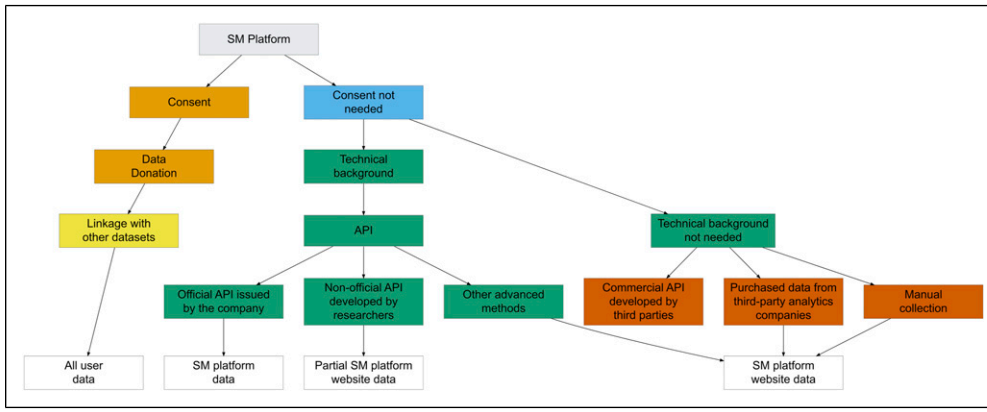


Figure 1. Accessing SM Data. *Note.* The diagram is built on Davidson et al. (2023)

Table 1. Description of each route from Figure 1 with the types of data that are typically obtained from that route

	Type of data collection	Type of data collected	Example	Pros	Cons
Data donation	Participants from a different dataset agree to offer their SM platform information.	Typically, all user data available on the platform (content, posts)	Researchers ask for consent from participants in a nationally representative survey of the U.K. if they can use their SM platform information (such as Twitter and LinkedIn[1]	Legal and academically ethical	A smaller sample size
Official API	Issued by the specific platform or company as the only official way to access data	SM platform data	TikTok Research API	Legal and academically ethical	Immensely restricted to SM platform terms of service: Specific application procedure (sometimes requires reference letters)
Commercial API developed by third-party companies	Also known as scraper, spider, or crawler, developed by third-party companies for commercial use	SM platform website data	TikTok Scraper by APIFY[2] or TikAPI[3]	Legal for the researcher and academically ethical; no technical background required	Payment (relatively affordable)
Non-official API developed by other researchers/ scientists	Developed and maintained by researchers	SM platform website data (partially)	GitHub wrapper[4]	Free; legal for the researcher	It may not work correctly when the platform changes its algorithms to prevent scraping

(continued)

Table 1. (continued)

	Type of data collection	Type of data collected	Example	Pros	Cons
Other advanced software or methods	Developed and maintained by experts who may not use common programming language or software (such as R and Python)	SM platform website data	Interface (previously known as Stiftung Neue Verantwortung)[5]	Free; legal; sometimes more data than official API	Ambiguous in research ethics[6]
Manual scraping	Researchers go to the webpage and manually copy information	SM platform data	Researchers may watch a specific group of videos under a hashtag[7]	Free; more reliable and increased familiarity with the datasets; legal; no technical background required	Labour-intensive; less effective or a smaller sample; ambiguous in research ethics
Purchased data	Researchers can choose from diverse services in the market	SM platform data	Analisa.io	Data has already been cleaned for use with customised variables, which are legal and academically ethical; no technical background required	Higher costs and still require minor manual calibration[8]

Note. The table is built on Davidson et al. (2023).

provided APIs, with the choice shaped by participant consent and the researcher's technical expertise. Among these, APIs offer a sanctioned and legally compliant route to structured, large-scale public social media data, with clear advantages over manual collection and unofficial scraping (see Figure 1; Table 1). Official APIs, such as those provided by TikTok, can grant access to archived content not visible on public-facing interfaces, thereby ensuring a degree of compliance with platform terms (Davidson et al., 2023; Sato, 2023). However, restrictive usage agreements, including rate limits and potential costs, can narrow the research scope and hinder open dissemination (Ohme et al., 2024). Manual collection remains accessible to those without advanced technical skills, but is labour-intensive, error-prone, and poorly suited to large datasets (e.g., Cervi & Divon, 2023). Unofficial scraping and some third-party crawlers present an alternative yet may breach terms of service and raise significant ethical and legal concerns (Bruns, 2019). Taken together, these constraints underscore the need to critically evaluate API-based methods for fairness, validity, and their capacity to support robust social media research.

The TikTok Research API: Promise and Pitfalls

Launched in 2023, the TikTok Research API represents a milestone in platform transparency and academic collaboration. Initial applications of the API have enabled researchers to analyse political communication, content virality, and algorithmic demographic bias through machine-learning models (Corso et al., 2024; Pinto et al., 2024a; Pinto et al., 2024b). The structured provision of multimodal data, including text, video, and audio metadata, marks a significant step forward compared to earlier TikTok studies reliant on ad hoc methods (Rogers & Zhang, 2024). This has enabled analyses ranging from content engagement statistics to semiotic and linguistic interpretations (Leedham et al., 2020; Parikh, 2025).

Nevertheless, the practical implementation of the API is far from seamless. Scholars have reported incomplete and inconsistent data delivery, often receiving only a fraction of the requested materials. Burnat and Davidson (2025) note that TikTok Research API imposes highly selective access criteria and suffers from documented data quality issues, corroborating reports of inconsistent data delivery (Corso et al., 2024; Ruz et al., 2023). Pfeffer et al. (2018) demonstrate that streaming sample APIs are neither random nor immune to distortion: deliberate flooding can manipulate topic coverage, while rate limits and batching introduce temporal artefacts and omit low-activity content. Furthermore, high-engagement posts are over-represented, challenging the notion that API data provide a faithful snapshot of platform dynamics. These constraints do not only limit the API's utility for comprehensive societal trend analysis but also can severely impact dataset completeness, thereby compromising analytical accuracy and interpretive robustness (Pfeffer et al., 2018). Additionally, server instability and unreliable pagination further undermine the reliability of automated data extraction processes (Mimizuka et al., 2025). If API workflows encourage fragmented or low-fidelity datasets, then the integrity of empirical findings is inevitably weakened.

Data Quality, Governance, and Ethics

Concerns over the epistemological trustworthiness of API-derived data extend beyond purely technical limitations. TikTok's internal moderation algorithms and content-curation systems introduce layers of bias and obfuscation rarely addressed in dataset evaluations. These 'black-box' mechanisms can distort both the authenticity and representativeness of retrieved data, creating systemic blind spots in research outputs (Corso et al., 2024).

Governance frameworks further shape who can access platform APIs, often reinforcing epistemic privilege for university-affiliated researchers. Under the European Union's *Digital Services Act*, applicants must demonstrate formal affiliation with a recognised academic or non-profit institution to obtain 'vetted' status for internal datasets (European Centre for Algorithmic Transparency, 2025). In practice, this requirement compels independent researchers, freelance journalists, and civil society organisations to partner with universities, think tanks, or research consortia to meet eligibility criteria. As a result, API access remains largely confined to academic investigators, excluding many non-academic actors from direct engagement with platform data (Mimizuka et al., 2025).

Burnat and Davidson (2025) identify critical 'audit blind-spots' in platform APIs, including TikTok's, where moderation and algorithmic processes remain opaque, an 'accountability paradox' in which platforms control both the data and the terms of its release. This lack of transparency, coupled with epistemic asymmetries favouring platform operators, raises significant ethical challenges for ensuring fairness and validity in computational social science.

A further point of contention concerns user consent and data privacy. The granularity of engagement metrics available through the API, when combined with other datasets, can enable

inadvertent re-identification of individual users. Such risks create potential legal liabilities and challenge the ethical legitimacy of research reliant on these data (Mimizuka et al., 2025), especially for those who are not qualified as public figures.

In sum, while the TikTok Research API represents a notable advance in platform-based research, the literature depicts its utility as mixed. Technical constraints, governance opacity, and ethical exclusions collectively limit its broader adoption and impact. These unresolved issues form the rationale for this study's central aim: to conduct a robust, comparative audit of TikTok's Research API against alternative data collection methods, including web scraping, and commercial analytics services, thereby contributing to methodological best practices in computational social science and informing future platform policy.

Methodology

This study adopts a mixed-method design to evaluate the efficacy and limitations of TikTok Research API in analysing societal trends, with a specific focus on comedic content during the COVID-19 pandemic. Data were obtained from three sources, the official TikTok API, a third-party analytics service (Analisa.io), and manual calibration, enabling a comparative assessment of accuracy, completeness, and representativeness. In doing so, the research addresses persistent gaps in platform-based studies (Davidson et al., 2023). The following subsections outline the sampling strategy, data acquisition procedures, and analytical approach, with particular attention to methodological robustness and ethical compliance.

Sample Selection

The COVID-19 pandemic offers a critical context for examining societal responses under conditions of collective stress, revealing patterns of isolation, mental-health strain, and the circulation of misinformation (Davidson et al., 2023). Comedy on TikTok, often rooted in shared social truths, provides a distinctive lens through which to interrogate cultural narratives and patterns of user engagement. As neuroimaging research indicates, the brain's response to vividly presented content can mirror real-world experiences, harnessing neuroplasticity to capture attention and reflect the prevailing zeitgeist (Nishimoto et al., 2011; Pearson, 2019; Pearson et al., 2015). Against this backdrop, prominent comedic creators were selected as a purposive sample for examining real-time societal discourse.

The raw dataset comprised 6,373 videos posted between 1 January 2020 and 31 December 2022 from 14 high-profile TikTok comedians: seven from the United Kingdom and seven from the United States, representing both genders. Creators were identified through a combination of Google search visibility and viral-content indicators, ensuring coverage of influential socio-cultural narratives (see Table 2 and Appendix 2 for video counts and detailed sampling procedures respectively). Two accounts, '*@steven_he*' and '*@nicholas_flannery*', were designated as optional. The former could not be unambiguously classified as British or American, identifying as Chinese Irish and residing in the United States during the study period, while the latter began uploading only in 2021, resulting in incomplete temporal coverage. Thus, the initial dataset for the following analysis includes 12 creators (6 from each country, balanced by gender in each country, totalling 5,246 videos).

The core analytical materials comprised video metadata (e.g. upload date, video description) and engagement metrics (e.g. views, likes, comments, shares) (see Figure 2). These provided both quantitative indicators and discursive content, enabling a multi-layered analysis of comedic expression and audience interaction during a period of significant social disruption.

Table 2. Total number of videos of selected content creators

Username handle	Gender	Nationality	No. of videos (research API)	No. of transcriptions (research API)	Transcription%	Period of transcriptions
lizzza	Female	American	159	1	1%	2022
daniel.labelle	Male	American	302	19	6%	2022
hannahstocking	Female	American	606	28	5%	2022
jeremylynch	Male	British	792	92	12%	2021.9~2022
adamw	Male	American	632	111	18%	2021.6~2022
spencewuah	Male	American	2,346	188	8%	2021.6~2022
adrianbliss	Male	British	106	0	0%	N/A
steven_he	Male	American (Chinese Irish actor)	199	35	18%	2021.12~2022
hayleygeorgiamorris	Female	British	187	0	0%	N/A
jackjos3ph	Male	British	832	0	0%	N/A
nicholas_flannery	Male	British	842	13	2%	2021.2~2022
ameliadimz	Female	British	141	13	9%	2021.8~2022
itscaitlinhello	Female	American	160	28	18%	2021.4~2022
maddiegracejepson	Female	British	580	136	23%	2022

Note. The content creators analysed in this study have not been anonymised, as they are public figures (Stevens et al., 2015). This approach parallels established conventions in SM research, where participant identities are likewise reported without pseudonymisation (e.g. Alexandre et al., 2022; Arora et al., 2019).

Data Acquisition and Annotations

To enable a triangulated assessment of data fidelity, three distinct sources (Table 3) were employed: a commercial analytics platform (Analisa.io), the official TikTok Research API, and manual calibration. This multi-source design allowed systematic cross-validation of completeness, accuracy, and consistency across retrieval methods.

The first dataset was obtained in March 2023 from Analisa.io, a commercial social media analytics service (Lachief, 2023). However, access to several creator profiles, including *@lizzza*, *@ameliadimz*, *@maddiegracejepson*, and *@hannahstocking*, was subsequently lost following the service's closure, resulting in partial discontinuity.

The second dataset was collected in April 2024 via the official TikTok Research API (the raw data are shown in Figure 3), accessed in R using packages such as `httr` and `jsonlite` for HTTP requests and JSON parsing (Pascual, 2020). Notably, fields in Figure 2 are publicly available information, which does not qualify as an 'internal' dataset under the *Digital Services Act* (European Commission, 2025), nor do they contain sensitive or private information beyond what is already publicly accessible. All codes are available on GitHub.[9] At the time of collection, three primary endpoints were available (see the list of endpoints in Appendix 3), including video metadata, comments, and creator profiles, with video metadata forming the principal analytical focus (Figure 3). Data were filtered by publication date (2020–2022), geographic region (United Kingdom and United States), and key content metrics (Appendix 4).

As is typical for behavioural trace data, raw API output required extensive preprocessing. Technical constraints, including rate limits, pagination handling, emoji decoding (Figure 2),

Table 3. Data Description from Three Data Sources

Username	Analisa.io	TikTok Research API	Manual calibration
lizzza	N/A	159	168
daniel.labelle	319	302	319
hannahstocking	N/A	606	691
jeremylynch	738	792	734
adamw	567	632	584
spencewuah	457	2,346	449
adrianbliss	69	106	113
steven_he	237	199	237
hayleygeorgiamorris	204	187	201
jackjos3ph	806	832	1,028
nicholas_flannery	902	842	890
ameliadimz	N/A	141	156
itscaitlinhello	170	160	170
maddiegracejepson	N/A	580	592
Number of participants	10	14	14
Number of videos	4,469	7,884	6,373
Engagement metrics	View, likes, comments, shares	View, likes, comments, shares and transcriptions	View, likes, comments, shares
Time of collection	Mar-23	April 2024	Between April and July 2023

and incomplete exposure of video-level features, restricted both the volume and granularity of extractable data. Every cleaning step was systematically documented, with derived variables cross-checked against supplementary metadata sources (Lazer et al., 2021; Salganik, 2018). This approach ensures transparency, reproducibility, and interpretability in subsequent modelling.

Finally, manual calibration was undertaken between April and July 2023, with videos verified and collected directly from TikTok.com. This step was designed to check whether there were missing videos between actually available metadata on the website and the purchased source or API source. Therefore, it is reasonable to compare the number of videos from three sources. This served as a benchmark against which to evaluate the completeness and accuracy of both API- and analytics-based retrieval. Research-ready version is presented in Figure 4.

Data Analyses

To evaluate the TikTok Research API's efficacy for computational social science, the analysis proceeded in two phases, addressing both macro-level data coverage and micro-level metric precision (Figure 5). This dual approach aligns with RQ1 by assessing the API's technical capacity for multimodal data collection and with RQ2 by examining whether systematic discrepancies across data sources could introduce demographic or algorithmic biases that compromise the fairness of societal insights. In other words, if certain groups or content types are systematically under- or over-represented due to data source effects, subsequent interpretations of societal trends risk being skewed.

Example Output from API

create_time	like_count	comment_count	share_cou	view_count	video_description	hashtag_names	voice_to_text	id	music_id	effect_ids	region_cod	username
28/12/2022 21:07	1340936	14533	33668	4750903	ðŸ˜³I LOVE YOUR SON ð christmas, foryou, pr		hi. I know everyone's	7.18231E+18	7.18231E+18	0	US	itscaitlinhello
24/11/2022 18:45	68249	1114	1304	326738	Happy thanksgiving, res foryou, fyp, thankagi		hey, I know it's morn	7.16966E+18	7.16966E+18	0	US	itscaitlinhello
15/11/2022 23:16	162607	2260	4432	627859	She has a secret. And iti movies, foryou, pov,		well, well, well. if it	7.16639E+18	7.16639E+18	0	US	itscaitlinhello
30/10/2022 18:59	79734	491	333	369475	Naked, WET. #fyp #foryı halloween, cousins, f		hey, sure, sure, you n	7.16038E+18	7.16038E+18	0	US	itscaitlinhello
13/10/2022 22:55	165465	1225	6129	791826	Good friend dinner with kimkardashian, foryo		Jeff Bezos is like, whi	7.15413E+18	7.15413E+18	0	US	itscaitlinhello
11/10/2022 22:21	741690	11690	24469	2254431	Avery:ðŸ§â€™ð, #fyp f foryou, parents, fyp,		free can you please l	7.15E+18	7.15338E+18	0	US	itscaitlinhello
08/10/2022 15:07	62819	636	1342	383807	Nicola Peltz-Beckham w vogue, foryou, fyp, ni		hi British Vogue it is f	7.15216E+18	7.15216E+18	0	US	itscaitlinhello

Figure 2. Example output from API. *Note.* The screenshot is intended solely to illustrate the workflow and data structure, not to expose raw data. Only a partial view is shown to minimise any risk of unnecessary disclosure

Phase 1: Number of Videos. The first phase compared the monthly total number of videos (dependent variable) retrieved from the three data sources (API, manual calibration, and Analisa.io) to assess the API's ability to provide a comprehensive dataset. Raw dataset included 14 creators over 36 months including two optional candidates (*@steven_he* and *@nicholas_flannery*), and then initial specification included 12 creators, yielding a balanced panel of 1,251 observations.

After data harmonisation, linear mixed-effects models were fitted using the `lme4` package in R, with creator gender (male/female), nationality (UK/US), and year of publication (2020, 2021, 2022) as control variables. A random intercept was added to account for creator-level clustering.

Robustness analyses were conducted in two subsets (1) excluding four creators lacking purchased-source data, leaving 829 observations across eight creators, and (2) excluding both the four creators with missing purchased data and one high output outlier, yielding 733 observations across seven creators.

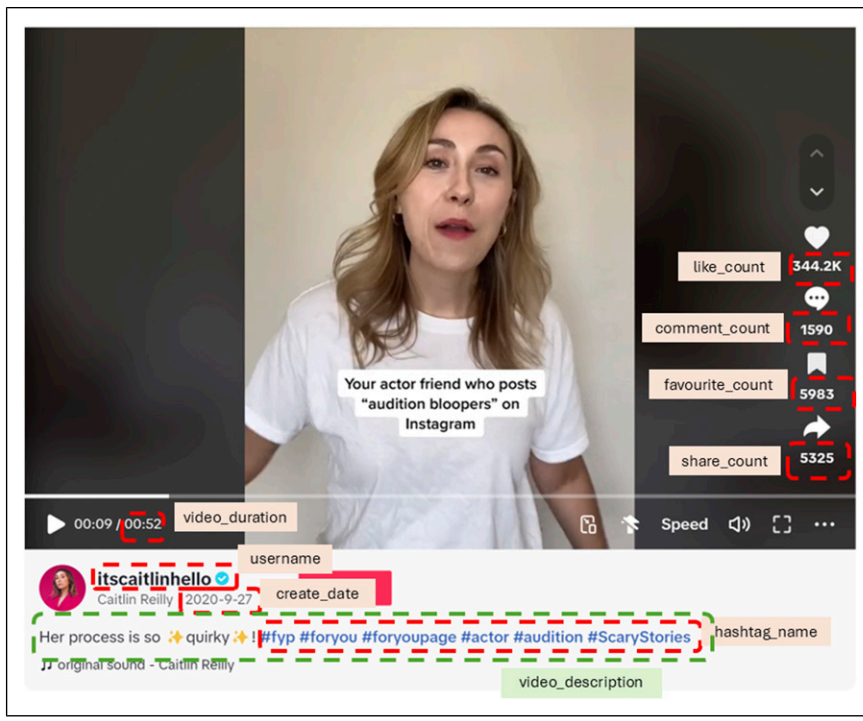
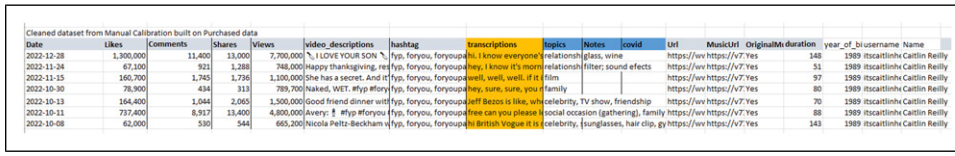


Figure 3. Sample video by @itscaitlinhello. *Note.* 'favourite_count' is not included in this paper due to its unavailability by the time of data collection (April 2024)



Cleaned dataset from Manual Calibration built on Purchased data

Date	Likes	Comments	Shares	Views	video_descriptions	hashtag	transcriptions	topics	Notes	covid	Url	MusicUrl	OriginalMu	duration	year_of_bi	username	Name
2022-12-28	1,300,000	11,400	13,000	7,700,000	↖, I LOVE YOUR SON ↖	fyp, foryou, foryoupa	hi. I know everyone'	relationsh	glass, wine		https://ww	https://v7.	Yes	148	1989	itscaitlinhe	Caitlin Reilly
2022-11-24	67,100	921	1,288	748,000	Happy thanksgiving, res	fyp, foryou, foryoupa	hey, I know it's morn	relationsh	filter; sound efects		https://ww	https://v7.	Yes	51	1989	itscaitlinhe	Caitlin Reilly
2022-11-15	160,700	1,745	1,736	1,100,000	She has a secret. And it'	fyp, foryou, foryoupa	well, well, well. if it i	film			https://ww	https://v7.	Yes	97	1989	itscaitlinhe	Caitlin Reilly
2022-10-30	78,900	434	313	789,700	Naked, WET. #fyp #fory	fyp, foryou, foryoupa	hey, sure, sure, you a	family			https://ww	https://v7.	Yes	80	1989	itscaitlinhe	Caitlin Reilly
2022-10-13	164,400	1,044	2,065	1,500,000	Good friend dinner with	fyp, foryou, foryoupa	Jeff Bezos is like, wh	celebrity, TV show, friendship			https://ww	https://v7.	Yes	70	1989	itscaitlinhe	Caitlin Reilly
2022-10-11	737,400	8,917	13,400	4,800,000	Avery: 🧍 #fyp #foryou #fory	fyp, foryou, foryoupa	free can you please k	social occasion (gathering), family			https://ww	https://v7.	Yes	88	1989	itscaitlinhe	Caitlin Reilly
2022-10-08	62,000	530	544	665,200	Nicola Peltz-Beckham v	fyp, foryou, foryoupa	hi British Vogue it is	celebrity, sunglasses, hair clip, gy			https://ww	https://v7.	Yes	143	1989	itscaitlinhe	Caitlin Reilly

Figure 4. Cleaned dataset from manual calibration built on purchased data

Phase 2: Quantity and Quality of Variables. The second phase examined data completeness and accuracy. First, the number of available metadata variables per video was compared across sources to provide a holistic measure of coverage. Next, the accuracy of key engagement metrics, like count, view count, comment count, and share count, was assessed, as these are widely used indicators of content impact in social media research. Their reliability is critical: systematic under- or over-reporting could distort comparative analyses and, by extension, the fairness of societal interpretations.

For comparability, only videos present in both sources were retained; those unique to a single source were excluded, thus leaving paired videos from 10 participants in total. Engagement analyses were restricted to API and manually calibrated data, as Analisa.io's rounded outputs closely mirrored manual counts, limiting their utility for precision testing. Notably, engagement metrics of *@lizzza, @ameliadimz, @maddiegracejepson,* and *@hannahstocking* from manual calibration are unavailable due to their lack of purchased source.

Linear mixed-effects models were also used, but with log-transformed outcome variables (via `log1p`) to normalise distributions. Fixed effects included data source, creator gender, nationality,

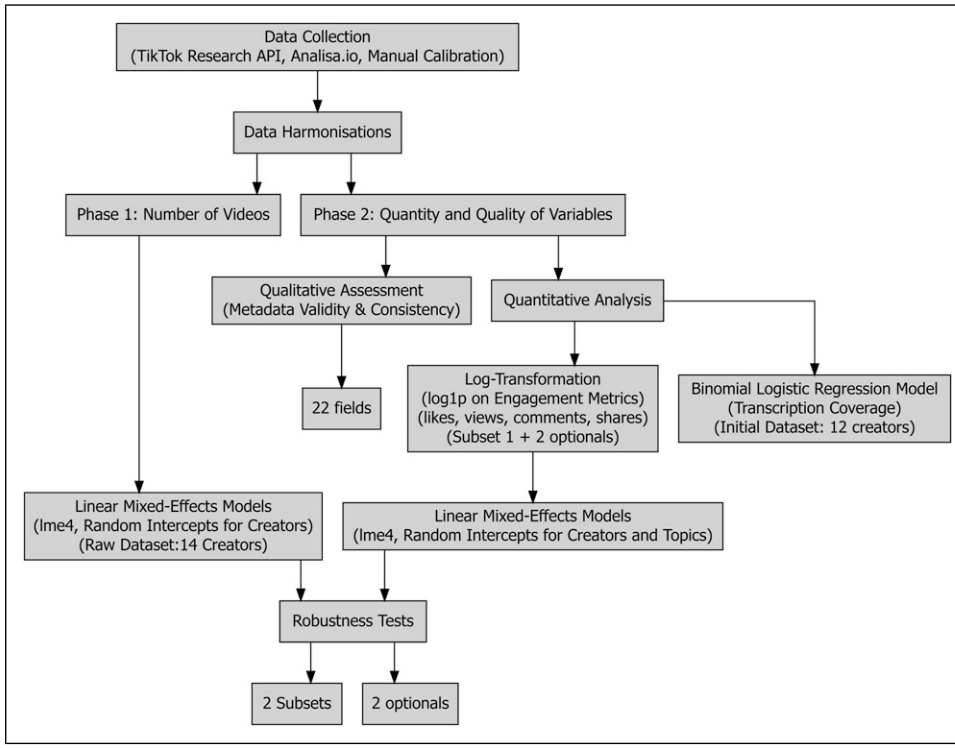


Figure 5. Workflow diagram illustrating the two-phase data analysis process

and year of publication, while random intercepts for the creators and topics accounted for repeated measures. Topics were coded manually (e.g. childhood, family, friendship, see details in Appendix 5). Alongside quantitative checks, a qualitative review assessed the internal consistency of metadata fields.

Additionally, the newly introduced voice_to_text (transcription) variable was evaluated for validity, given its potential to introduce further demographic or linguistic bias. A binomial logistic regression was used to compare the differences in proportion of videos of having a transcription (dependent variable: whether a video had a transcription) as a function of gender and nationality (independent variables).

Preliminary Findings

Number of Videos

Descriptively, on average each month, the API ($M = 19.70, SD = 27.30$) and Manual sources ($M = 12.41, SD = 16.31$) generally captured a higher number of videos in raw counts than the Purchased source ($M = 11.56, SD = 16.70$). Figure 6 presents the total number of videos between 2020 and 2022 across three data sources, API, Manual, and Purchased, for 14 users, with each user identified by username, gender, and nationality. Heterogeneous correlation analysis indicated negligible associations among nationality, data source, and gender (all $|r| < 0.05$).

The linear mixed-effects analyses revealed consistent differences in estimated upload counts between data sources. In the full model including all 14 creators, both the purchased dataset

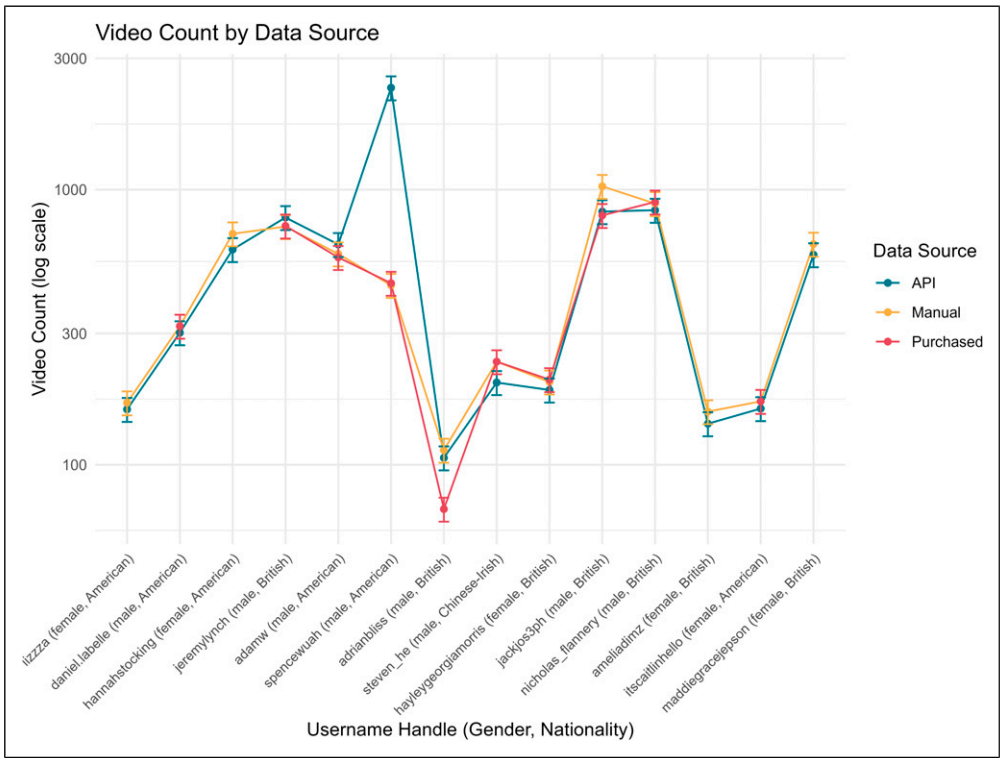


Figure 6. Video count by data source

($\beta = -7.98$, $p < .001$, 95% CI [−10.464, −5.523]) and the manual dataset ($\beta = -4.80$, $p < .001$, 95% CI [−7.000, −2.620]) had significantly lower counts than the API baseline. Gender was not significant, but male creators tended to upload more frequently ($\beta = 12.80$, $p = .068$, 95% CI [−1.12, 26.73]), while nationality effects were not significant ($\beta = -15.764$, $p = .235$, 95% CI [−37.899, 6.371]).

When the two optional candidates (*@steven_he* and *@nicholas_flannery*) were excluded, the pattern persisted, with larger estimated differences for purchased ($\beta = -10.17$, $p < .001$, 95% CI [−12.950, −7.408]) and manual ($\beta = -5.88$, $p < .001$, 95% CI [−8.279, −3.496]) sources than API. Further exclusions of the four creators without purchased data ($n = 8$, $N = 829$) showed that API-sourced monthly uploads were still higher than Purchased ($\beta = 11.764$, $p < .001$, 95% CI [8.560, 14.979]) and Manual ($\beta = 9.034$, $p < .001$, 95% CI [5.802, 12.300]).

As shown in Figure 6, there is an outlier of API (@spencewhuah). Thus, in a more restricted but balanced subset (e.g. excluding four creators absent from purchased data, two optional candidates, and the outlier, finally seven creators), the negative coefficients for purchased and manual sources were smaller and statistically insignificant (e.g. purchased: $\beta = -1.20$, $p = .214$, 95% CI [−3.093, 0.683]; manual: $\beta = -1.97$, $p = .937$, 95% CI [−1.970, 1.806]), indicating the results' sensitivity to sample composition.

Quantity and Quality of Variables

The API generally provides a greater number of variables and with higher quality than alternative sources. Specifically, the API offers approximately 22 variables, including eight unique fields unavailable in other sources: effect ID, hashtag description, hashtag ID, playlist ID, region code, transcription, video label, and verified STEM information. Additionally, four variables (creation time, like count, music ID, and view count) exhibited greater precision (Table 4).

Nonetheless, the API has limitations; for example, it fails to decode emojis in video descriptions but showed 'garbled text' or 'mojibake' (Kita et al., 2022) when data are exported into CSV or Excel formats, requiring manual correction and extra online tools (Faust, 2017/2024) during cleaning. Furthermore, some variables (e.g. favourite count, hashtag description, and video duration) were introduced incrementally after the API's initial release.

Variable Quality: Like Count. Across both manual and API sources ($N = 10$ creators, 8,424 observations), the average like count per observation was 5.47×10^5 ($SD = 1.23 \times 10^6$), with values ranging from 17 to 24.6 million. The mean like count was 5.50×10^5 ($SD = 1.23 \times 10^6$) for API entries and 5.45×10^5 ($SD = 1.22 \times 10^6$) for manual entries. Variance inflation factors indicated no evidence of problematic multicollinearity among predictors (all GVIF^(1/(2·Df)) ≤ 1.03). Figure 7 presents the monthly like counts with log-transformed values.

In the model including all 10 creators (excluding those without purchased data), the year of posting emerged as a significant predictor of like counts: relative to 2020, like counts were higher in 2021 ($\beta = 1.583$, $p < .001$, 95% CI [1.476, 1.690]) and 2022 ($\beta = 1.828$, $p < .001$, 95% CI [1.725, 1.931]), and 2022 was in turn even higher than 2021 ($\beta = 0.245$, $p < .001$, 95% CI [0.164, 0.326]). By contrast, there was no significant difference between manual and API sources ($\beta = 0.011$, $p = .737$, 95% CI [−0.051, 0.072]). Gender and nationality effects were not significant (p's > .171).

When the two optional candidates were excluded, the pattern persisted. Data source effects remained negligible ($\beta = 0.013$, $p = .698$, 95% CI [−0.053, 0.079]), while year effects were again substantial: 2021 ($\beta = 1.736$, $p < .001$, 95% CI [1.629, 1.844]) and 2022

Table 4. Variable Quality from Three Sources

Content made available for research								
Videos								
		Availability			Quality			
	Variable names	Research API	Analisa.io	Manual	Research API	Analisa.io	Manual	Notes
1	comment count (comment_count)	Available	Available	Available	Accurate to last digit	Accurate to last digit	Accurate to last digit	This is the total number of comments posted on a video.
2	create time (create_time)	Available	Available	Available	Accurate to minute	Accurate to minute	Accurate to date	This is the time when the video was created.
3	effect_ids	Available	Not available	Not available	Unique	NA	NA	The list of effects applied on the video.
4	favourite count (favourite_count)	Available after July 2024	Not available	Available	Accurate to last digit	NA	Accurate to last digit	The number of favourites a video receives.
5	hashtags (hashtag_name)	Available	Available	Available	Accurate	Accurate	Accurate	The list of hashtags used in the video.
6	hashtag_description	Available after July 2024 or later	Not available	Not available	Unique	NA	NA	Returns a description for a hashtag_name if one exists.
7	hashtag_id	Available after July 2024 or later	Not available	Not available	Unique	NA	NA	Returns the unique hashtag_ids for each hashtag.
8	like count (like_count)	Available	Available	Available	Accurate to last digit	Rounded number to hundred	Rounded number to hundred	The total number of likes on a TikTok video, created by users by clicking the 'Heart' icon.
9	music id (music_id)	Available	Available	Not available	Accurate to last digit	Binary (original or else)	NA	This is the music_id used in the video.
10	playlist_id	Available	Not available	Not available	Unique	NA	NA	The ID of the playlist that the video belongs to.

(continued)

Table 4. (continued)

Content made available for research							
Videos							
		Availability			Quality			
	Variable names	Research API	Analisa.io	Manual	Research API	Analisa.io	Manual	Notes
11	region code	Available	Not available	Not available	Unique	NA	NA	A two-digit code for the country where the video creator registered their account.
12	share count (share_count)	Available	Available	Available	Accurate to last digit	Accurate to last digit	Accurate to last digit	The total number of times a TikTok video has been shared by clicking the 'Share' button with the video.
13	transcription (voice_to_text)	Available	Not available	Not available	Unique (low coverage)	NA	NA	Voice to text and subtitles (for videos that have voice to text features on, show the texts already generated)
14	username	Available	Available	Available	Accurate	Accurate	Accurate	This is the username of the video creator.
15	video description (video_description)	Available	Available	Available	Hard to decode emojis sometimes	Accurate	Accurate	This is the description of the video.
16	video duration in seconds (video_duration)	Available after July 2024	Available	Available	Accurate	Accurate	Accurate	By the end of my data collection (June 2024), Research API only offered the field name of 'video_length', which will only return 'SHORT', 'MID', 'LONG', 'EXTRA_LONG', so I did not include that index in the first place in the R codes. But I checked the official codebook in July, 'video_duration' was added, and the precise length of the video can be returned through the query, along with another 2 field names of 'is_stem_verified' and 'favourites_count'

(continued)

Table 4. (continued)

Content made available for research							
Videos							
		Availability			Quality			
	Variable names	Research API	Analisa.io	Manual	Research API	Analisa.io	Manual	Notes
17	video id	Available	Not available	Available	Accurate	NA	Accurate	The unique identifier of the TikTok video. This is also called 'item_id' or 'video_id'. This is a number that can be used to reconstruct the URL link to access the video.
18	video_label	Available after July 2024 or later	Not available	Available	Unique	NA	NA	Returns any labels applied to a video such as 'election labels' (ex: Get info on the U.S. elections)
19	video_mention_list	Available after July 2024 or later	Available	Available	Available	Available	Available	Returns the other tagged users in a video.
20	view count (view_count)	Available	Available	Available	Accurate to last digit	Rounded number to hundred	Rounded number to hundred	This is the total number of views for a video on TikTok.
21	whether verified stem information (is_stem_verified)	Available	Not available	Available	Unique	NA	NA	Whether the video has been verified as being high quality STEM content.
22	video url	Partially available	Available	Available	NA	Accurate	Accurate	Video URL can be constructed with video ID.

Note. Details of variable names can be found in Appendix 4.

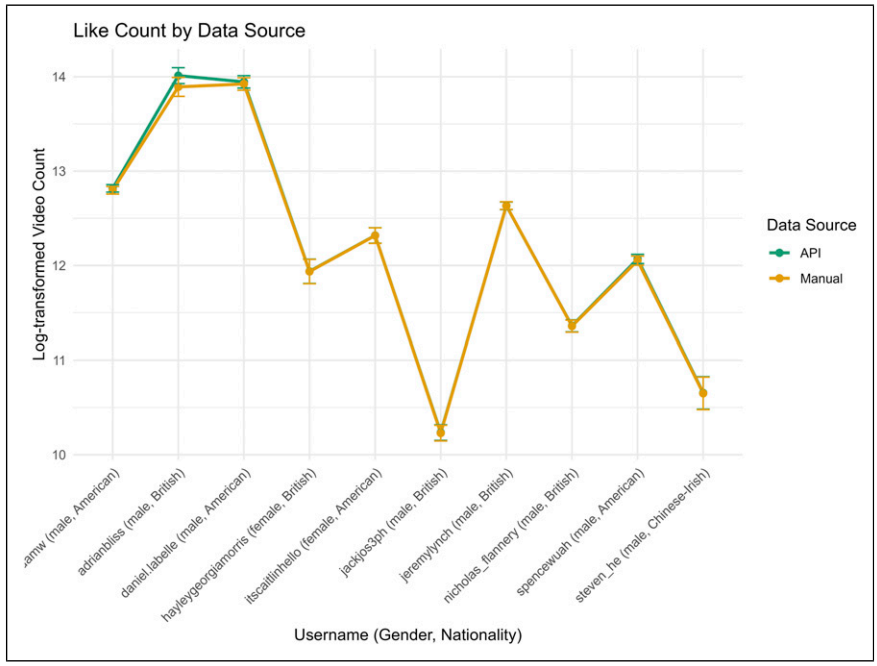


Figure 7. Monthly like count by data source (manual and API)

($\beta = 1.530$, $p < .001$, 95% CI [1.426, 1.633]) both exceeded 2020. There were no gender or nationality effects.

Variable Quality: View Count. As for the view count ($N = 8,424$ observatios, 10 creators), API-sourced data ($M = 3.69 \times 10^6, SD = 9.08 \times 10^6$) appeared to be lower than Manual ($M = 3.90 \times 10^6, SD = 9.17 \times 10^6$). As with the like-count model, predictors showed negligible intercorrelations (all GVIF^(1/(2·Df)) ≤ 1.03), confirming their suitability for inclusion. Figure 8 presents the monthly like counts with log-transformed values.

An analysis including all 10 creators showed that API-sourced counts were significantly fewer than manual counts ($\beta = -0.162$, $p < .001$, 95% CI [−0.218, −0.106]), while gender and nationality were not significant (p's > .327). Additionally, view counts in 2022 were higher than in 2021 ($\beta = 0.181$, $p < .001$, 95% CI [0.108, 0.254]), the latter of which were in turn higher than in 2020 ($\beta = 1.573$, $p < .001$, 95% CI [1.477, 1.669]). When the two optional candidates were excluded, reducing the sample to eight creators, the pattern remained (see details in Appendix 6-A).

Variable Quality: Comment Count. In terms of comment count ($N = 8,424$ observatios, 10 creators), API-sourced data ($M = 4239.60$, $SD = 1.06 \times 10^4$) appeared to be higher than Manual ($M = 3375.56$, $SD = 9923.08$). Predictor variables were effectively independent (all GVIF^(1/(2·Df)) ≤ 1.03; categorical associations were weak, Cramér's V < 0.12). Figure 9 presents the monthly like counts with log-transformed values.

The data source predicted comment counts: API-sourced counts were significantly higher than manual counts ($\beta = 0.211$, $p < .001$, 95% CI [0.148, 0.275]). While gender and nationality effects were not significant (p's > .172), years were significant. Comment count in 2021 exhibited substantially higher values than 2020 ($\beta = 1.501$, $p < .001$, 95% CI [1.391, 1.610]), but

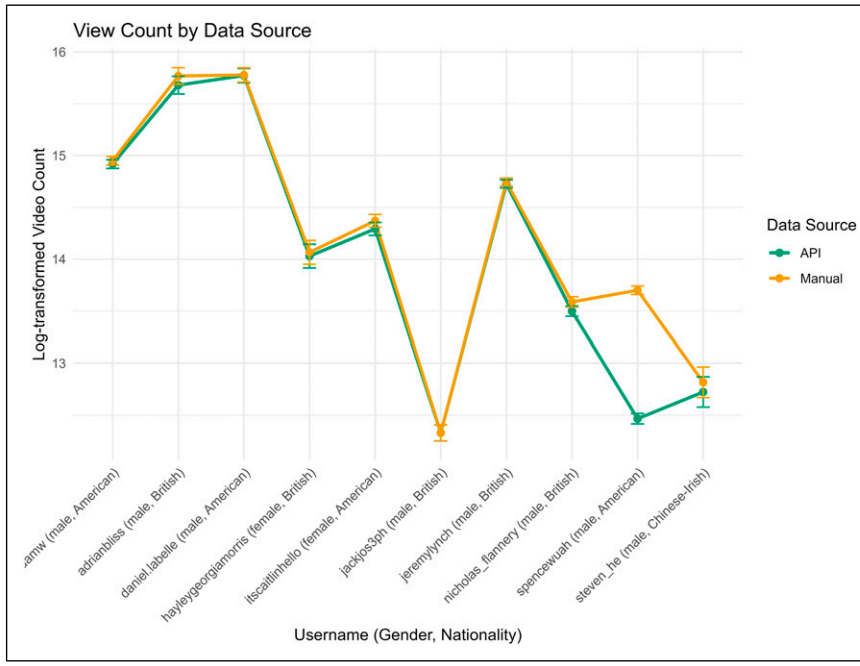


Figure 8. Monthly view count by data source (manual and API)

lower than 2022 ($\beta = -0.131,\ p = .002,\ 95\%\ CI\ [-0.215,\ -0.048]$). When the two optional candidates were excluded, reducing the sample to eight creators, the pattern remained (see details in Appendix 6-B).

Variable Quality: Share Count. As for share count ($N = 8,424,\ M = 1.75 \times 10^4,\ SD = 9.09 \times 10^4$), videos collected via the API ($M = 2.85 \times 10^4,\ SD = 8.92 \times 10^4$) had a higher mean values than manual collection ($M = 1.35 \times 10^4,\ SD = 1.16 \times 10^5$). Consistent with previous models, no problematic multicollinearity was detected (all GVIF$^{(1/(2\cdot Df))} \leq 1.03$). Figure 10 details the monthly share counts with log-transformed values ($n = 10$ creators).

Similar to comment counts, the analysis confirmed that API-sourced share counts were substantially higher than manual counts ($\beta = 0.922,\ p < .001,\ 95\%\ CI\ [0.840,\ 1.003]$), while gender and nationality were not significant (p's $> .229$). Additionally, relative to 2020, share counts were higher in 2021 ($\beta = 1.561,\ p < .001,\ 95\%\ CI\ [1.421,\ 1.701]$) and 2022 ($\beta = 0.529,\ p < .001,\ 95\%\ CI\ [0.394,\ 0.665]$), but 2022 were lower than 2021 ($\beta = -0.551,\ p < .001,\ 95\%\ CI\ [-0.678,\ -0.424]$).

Variable Quality: Transcription. The API returns voice-to-text transcriptions, one of its unique fields, though only for approximately 10% of videos on average (Figure 11) out of API-driven dataset. A binomial logistic regression model ($n = 12$) was fitted to examine whether transcription percentages (number of transcriptions divided by the number of videos) varied by creator nationality and gender, excluding two optional candidates (@steven_he and @nicholas_flannery). Main effects suggested that British creators had higher odds of transcription than American creators ($OR = 2.989,\ p < .001,\ 95\%\ CI\ [0.780,\ 1.422]$). Also, male creators had higher odds than female creators ($OR = 1.635,\ p < .001,\ 95\%\ CI\ [0.208,\ 0.793]$). However, there was a significant interaction between nationality and

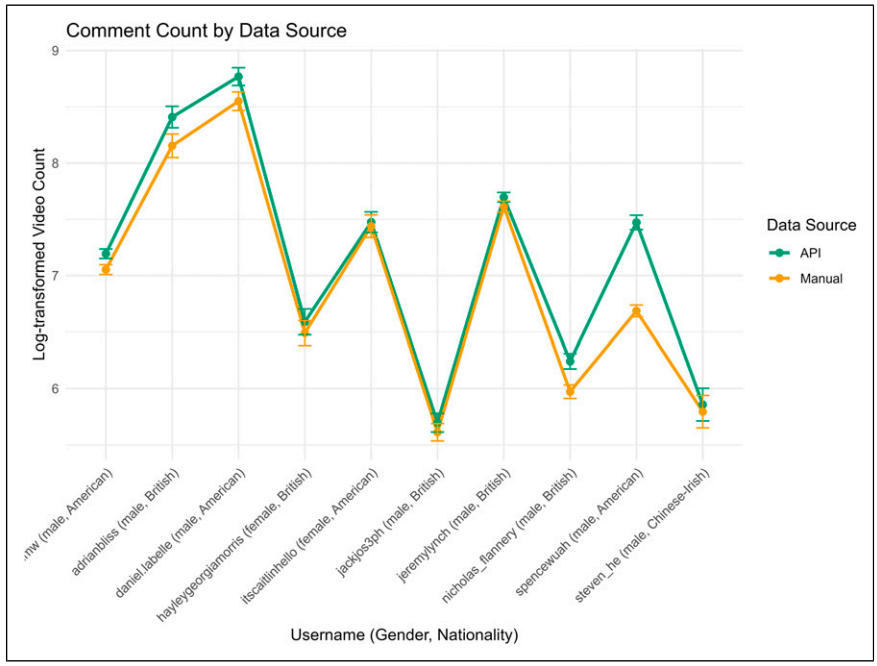


Figure 9. Monthly comment count by data source (manual and API)

gender ($OR = 0.175,\ z = -8.54,\ p < .0001,\ 95\%CI = [-2.15,\ -1.35]$), indicating that the nationality effect differed markedly by gender.

Post-hoc Tukey-adjusted pairwise comparisons revealed that British females had significantly higher odds of transcription coverage than American females (odds ratio = 2.99, $p < .001$, 95% CI [0.220, 0.509]). British females also had higher odds than American males (odds ratio = 1.83, $p < .001$, 95% CI [1.388, 2.409]) and British males (odds ratio = 3.50, $p < .001$, 95% CI [2.441, 5.004]). American males were more likely than American females to provide transcriptions (odds ratio = 1.63, $p = .0053$, 95% CI [0.417, 0.897]), whereas British males did not differ significantly from American females ($p = .805$, 95% CI [0.748, 1.827]). British males, however, had significantly lower odds than both British females and American males ($p < .001$ in each case).

Discussion

This study has critically evaluated the TikTok Research API's capacity to support computational social science, addressing its effectiveness for multimodal data collection (RQ1), the manifestation of biases in retrieved data (RQ2), and the potential of multi-method integration to mitigate these challenges (RQ3). Drawing on a dataset of 6,373 videos collected via multiple methodologies, the findings depict a nuanced landscape: the API offers clear strengths, notably scalable, ethically compliant retrieval, yet these are offset by inconsistent metadata and systematic demographic biases.

The analysis reveals marked variability in video counts, engagement metrics, and transcription coverage across creators, with patterns strongly influenced by data source, nationality and gender for transcriptions. First, the Research API retrieves marginally higher video volumes than alternative methods whilst providing access to unique variables unavailable through manual

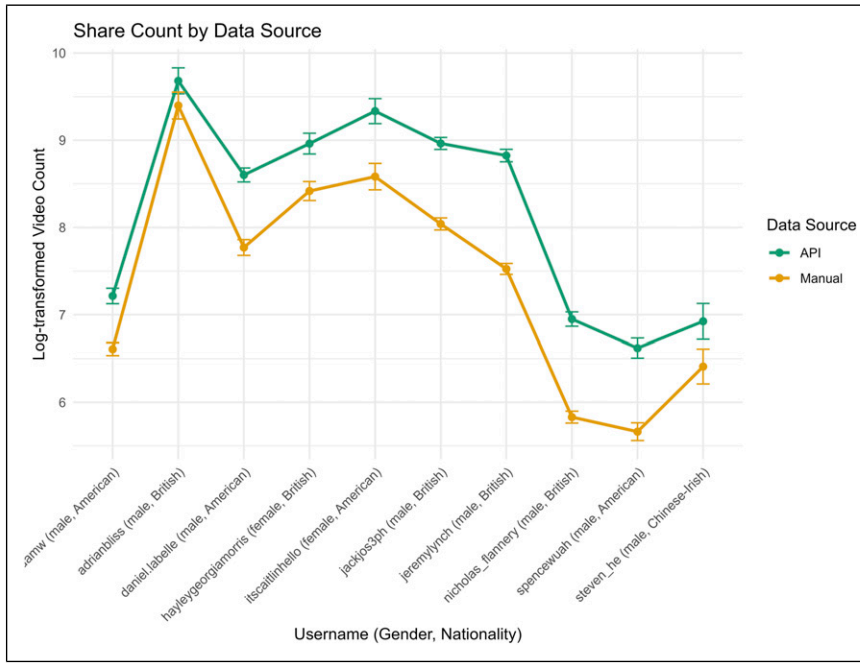


Figure 10. Monthly share count by data source (manual and API)

collection or third-party services. However, this advantage proves uneven, as differences between data sources appear concentrated among specific creators rather than reflecting universal measurement biases. Second, whilst data source and year exert a substantial influence on engagement metrics, gender and nationality do not. The consistent API advantage in engagement metrics raises questions about whether this reflects genuine behavioural differences or artefacts of the retrieval process.

Third, observed differences in transcription rates require careful interpretation. It is important to note that TikTok's Automatic Captions feature was only introduced in April 2021 (TikTok Newsroom, 2021), during the middle of our study period (2020–2022). Because creators had the option to enable captions, this may explain why transcripts generated via the API appear predominantly after that launch date and are absent beforehand. The decision to activate voice-to-text functionality rests with individual creators rather than the API itself. However, the uneven distribution of transcripts across demographic groups suggests that uptake patterns may reflect broader systemic factors. It is plausible that American creators adopted this functionality at higher rates than their British counterparts, which could account for regional variation. Similarly, differential adoption rates between male and female creators may point to underlying inequalities in platform literacy, technical confidence, or community norms around accessibility features. Whilst the API faithfully returns whatever transcriptions creators have enabled, these creator-level choices may themselves be shaped by demographic and cultural contexts. This raises important questions about equity in platform feature adoption and the extent to which automated systems can perpetuate existing social disparities even when those systems function as designed.

These results carry important implications for future research employing multiple data collection strategies in computational social science. While the API's quantitative advantages are evident, they are compromised by technical inconsistencies and demographic skews that raise fundamental questions about validity and representativeness in societal trend analyses. Addressing

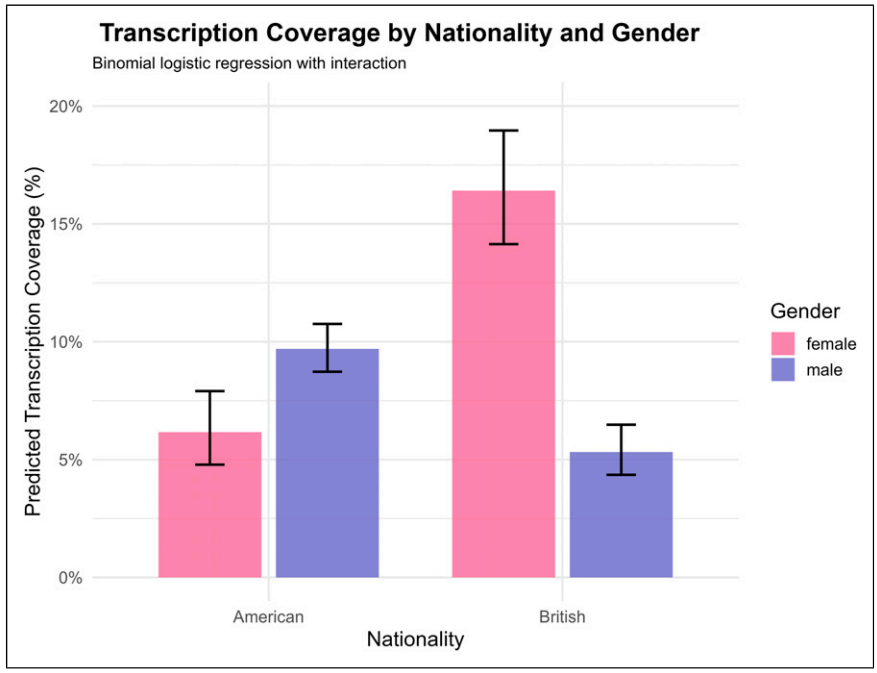


Figure 11. Transcription rate of research API

these limitations will require methodological designs that integrate complementary sources, apply rigorous bias detection, and ensure transparency in reporting.

Multimodal Data Collection: Capabilities and Constraints (RQ1)

Quantitative Superiority and Scalability Advantages. The analysis indicates that the TikTok Research API is highly effective for large-scale data acquisition, outperforming purchased datasets and manual collection in quantitative terms. Linear mixed-effects models confirm that it retrieves more videos overall, though results remain sensitive to sample composition. Meanwhile, the study suggests that API-sourced data systematically capture higher viewing, commenting, and sharing activity. Crucially, its advantage extends beyond volume: the API grants access to unique variables – including transcriptions, effect identifiers, and precise metadata fields – that remain unavailable through alternative approaches. These affordances enable rapid assembly of large longitudinal datasets, facilitating both quantitative modelling and qualitative content analysis.

This scalability marks a significant advance for computational social science. It enables the rapid assembly of datasets that would be prohibitively resource-intensive to collect manually. The ability to retrieve historical content spanning multiple years further enhances its value for longitudinal studies, particularly those examining temporal dynamics during crisis periods such as the COVID-19 pandemic. These capabilities support a wide range of analytical approaches, from quantitative modelling of engagement metrics to qualitative content analysis.

Technical Constraints Undermining Data Validity. Yet these strengths are tempered by technical shortcomings that compromise validity. Unstable API performance, consistent with Ruz et al. (2023), produces inconsistent data volumes across collection periods. Content visibility and access issues arise through deletion or privacy settings, rendering content inaccessible via the

API (TikTok for Developers, 2024), whilst manual calibration remain limited to publicly available videos. The outlier '*@spencewuah*' illustrates marked discrepancies in video counts between API and manual sources.

Further distortions occur because the API includes archived videos, such as drafts, inflating counts relative to platform-visible data. More data returned by the API does not necessarily represent better quality but may compromise results when researchers focus on platform-accessible content. Transcription coverage is particularly sparse; voice-to-text data exist for around 10% of videos, predominantly from 2021 onwards. This temporal skew excludes much of the early pandemic period from linguistic analysis, potentially obscuring shifts in discourse during initial lockdowns.

Metadata inconsistencies, notably the API's failure to decode emojis, produce mojibake (Kita et al., 2022) and require manual correction (Faust, 2017/2024). As Cohn's foundational work on the visual language of sequential images (2013) demonstrates, pictorial elements, whether in comics or digital media, operate within structured semiotic systems. Subsequent research on emoji sequencing (Cohn et al., 2019) extends this framework, showing that emoji, like other visual modalities, often lacks fully developed grammatical structures yet still interacts systematically with accompanying text. Such issues challenge Tromble's (2021) assertion that minimal programming suffices for API research, as specialist intervention is often essential.

Additionally, the Research API cannot return or store video or audio data, which must be obtained through third-party services (see a list of possible services in Appendix 7). Prosody, captured through speech rhythm, intonation, and emphasis, offers cues to stance and affect otherwise absent from purely textual analysis (Arvaniti, 2020; Mauchand et al., 2020). In multimodal contexts, prosodic contours synchronise with gesture, jointly signalling pragmatic meanings such as information status, stance, and (im)politeness (Brown & Prieto, 2021; Kendon, 2004; McNeill, 2005). This integration supports the view that prosody and gesture operate as sister systems in sociopragmatic marking (Ambrazaitis & House, 2017; Prieto & Espinal, 2020). In short-form video, these systems work in concert with visual symbols including emoji, to form a tightly integrated semiotic ensemble.

Ethical Constraints and Access Limitations. Regional restrictions shape content availability through government bans, platform moderation, and algorithmic filtering (Zhao, 2024). In this study, some U.S.-origin videos were inaccessible from the United Kingdom, reducing dataset completeness and constraining cross-national comparison. These constraints expose the API's limitations for longitudinal analysis and point to the need for multi-method calibration to safeguard validity. The API can increase data volume, but doing so demands substantial computational resources for processing and verification.

Technical and ethical constraints compound these issues. Strict retrieval limits and the requirement for formal ethical approval impede reproducibility, echoing earlier critiques of API-based research (Freelon, 2018). Access policies that privilege university-affiliated researchers raise equity concerns and appear misaligned with the EU's *Digital Services Act* commitment to transparency (Burnat & Davidson, 2025). Challenges in data dissemination and retention derived from ToS still haunt open science (Bak-Coleman, 2023; Venkatagiri, 2023). Collectively, these barriers not only narrow the empirical scope but also shape the wider conditions under which regulated digital research is conducted.

Algorithmic and Demographic Biases: Compromising Research Equity (RQ2)

The analysis identifies pronounced disparities within TikTok Research API that compromise research equity. Transcription coverage systematically varies across demographic groups, with

American and male creators demonstrating higher rates than their U.K. and female counterparts. Whilst the choice to enable TikTok's Automatic Captions feature rests with individual creators, the observed patterns suggest that adoption may be influenced by broader structural factors.

Several mechanisms may account for these disparities. American creators may have adopted auto-captioning functionality at higher rates than British creators, potentially reflecting earlier exposure to the feature, stronger platform literacy, or differing community norms around accessibility. Similarly, gendered differences in transcription availability may stem from variations in technical confidence, awareness of accessibility tools, or expectations regarding content presentation. Research on digital inequalities has consistently demonstrated that platform feature adoption is rarely neutral but often reflects pre-existing social stratifications (Feng et al., 2021; Markl, 2022; Sari et al., 2021).

Moreover, even when creators do enable captions, the accuracy of automated speech recognition systems may vary systematically across demographic groups. ASR models are typically trained on datasets that over-represent certain dialects, accents, or demographic groups, leading to differential performance (Wassink et al., 2022). Creators whose speech patterns align with dominant training data enjoy greater transcription accuracy, whilst those outside these norms may experience lower quality outputs, potentially discouraging further use of the feature. This creates a feedback loop wherein initial technical disparities become reinforced through user behaviour.

These demographic skews have significant implications for the fairness and validity of societal trend analyses. Unrepresentative samples risk distorting interpretations of cultural phenomena, particularly in policy-relevant contexts where findings might inform interventions. In such cases, conclusions may disproportionately reflect the perspectives of privileged groups, limiting generalisability in global or gender-diverse settings. The marked attenuation of effects following outlier removal further indicates that extreme observations can disproportionately shape perceived differences between data collection methods. This underscores the importance of examining distributional properties and identifying influential cases before drawing inferences about data quality.

Multi-Method Integration: Mitigating Inconsistencies for Enhanced Validity (RQ3)

In light of the challenges identified in RQ1 and RQ2, the analysis demonstrates that integrating multiple data collection methods, combining API-derived data with manual calibration and third-party analytics, offers a robust means of improving validity. This triangulated approach mitigates inconsistencies such as video-count discrepancies and enhances completeness through cross-validation of engagement metrics. In doing so, it addresses both technical limitations and bias-related distortions, thereby strengthening the evidential base for subsequent interpretation.

The comparative analysis reveals systematic differences in how engagement metrics are captured across methods. Data source exerts a significant influence on reported values: videos obtained via the API exhibit lower log-transformed view counts than those gathered manually, yet higher comment and share counts. This asymmetry suggests that platform-specific definitions or calculation procedures for engagement metrics may differ between collection routes. In other words, the same video can appear to perform differently depending on the method used, a non-trivial complication for comparative research.

Multi-method integration also proves critical for examining temporal engagement patterns. The year of posting emerged as a consistent predictor across metrics, with videos from 2021 recording markedly higher engagement than 2020 content. It is reasonable that established creators benefit from a larger reach over time, also known as 'cumulative advantage' (De Oliveira Santini et al., 2020). These temporal shifts underscore the importance of accounting for contextual factors when interpreting engagement trends. Encouragingly, as of 15 August 2025, TikTok Research API

includes a batch compliance task[10] feature (TikTok for Developers, 2024), allowing researchers to submit lists of video or comment identifiers for validation. This process enables verification of whether each identifier remains publicly accessible, supporting accurate reporting of dataset availability at the time of analysis.

By incorporating alternative transcription services such as Happy Scribe (see a list of possible services in Appendix 7) and cross-checking outputs, this methodological pluralism reduces the impact of incomplete coverage and demographic skew. Crucially, it enables researchers to distinguish findings that are robust across methods from those that are artefacts of a particular collection strategy. In periods of rapid social change, when data quality assumes heightened importance, such integration is not merely desirable but essential for producing credible, equitable, and generalisable insights.

Data Accessibility, Methodological Transparency, and Ethical Considerations

The empirical patterns identified in RQ1–RQ3 reinforce the existing literature on the tension between data legitimacy and research reproducibility in computational social science (Davidson et al., 2023), including demographic and algorithmic biases, metric inconsistencies across collection methods, and the benefits of multi-method integration. While the API enables structured, large-scale data retrieval, its utility is constrained by platform-imposed rate limits and post-hoc deletions by content creators. Such constraints, consistent with Ruz et al. (2023) and Entrena Serrano et al. (2025), undermine data completeness, particularly in longitudinal studies of temporally sensitive phenomena such as pandemic-related humour. Echoing Pfeffer et al. (2018), sample-streaming endpoints also exhibit non-random omissions, manipulation, and popularity bias, necessitating corrective measures such as stratified sampling or weighting to preserve representativeness.

Discrepancies in engagement metrics further complicate interpretation. API-sourced videos tend to report lower view counts, but higher comment and share counts than manually collected equivalents, aligning with Pearson et al. (2025), who show that both data source and year significantly influence TikTok viewership. Notably, no significant effects emerged for data source in like-count analyses, suggesting that likes may be more consistently measured across methods. These asymmetries highlight the need for methodological transparency when comparing engagement metrics derived from different pipelines.

Ethical considerations compound these methodological challenges. The API's exclusivity disadvantages independent scholars and civil society groups, reinforcing structural inequities in research access. Moreover, its granular engagement data, including comment histories, raise privacy risks such as potential deanonymisation, even within TikTok's consent frameworks (Mimizuka et al., 2025). These risks are magnified in high-stakes contexts such as public health or political discourse, where the digital divide exacerbates disparities in who can produce, access, and analyse data (Bezjak et al., 2018; Rossi & Lenzini, 2020). Robust ethical protocols and transparent governance are therefore imperative to balance analytical utility with accountability.

By leveraging a large, balanced dataset spanning two nations and genders, this study addresses the limitations of prior TikTok research, which often relied on smaller, non-systematic samples (e.g. Cervi & Divon, 2023). It complements emerging machine-learning applications of the TikTok Research API in political and behavioural domains (Corso et al., 2024; Pearson et al., 2025), while applying a critical social science lens to issues of methodological rigour and representational equity. The findings affirm the API's potential as a scalable, compliant tool for computational social science, but also underscore the necessity of multi-method strategies to mitigate its shortcomings. Balancing its strengths (volume, unique variables) against its

weaknesses (quality, bias) offers a framework for best practice in ensuring validity and fairness, contributing to the ongoing discourse on platform-based methodologies.

Study Limitations

Whilst this study offers valuable insights into TikTok Research API for social science research, several limitations must be acknowledged. First, the restricted sample of creators poses a notable constraint. With data drawn from only 14 creators based in the United States and the United Kingdom, the findings may not fully represent the diverse TikTok creator population. This limitation restricts the generalisability of the results, particularly for creators from other regions or with varying audience demographics (Herring, 2009; Lomborg & Bechmann, 2014).

Second, the discrepancies between API and manual coding are unlikely to be attributable solely to time accumulation. Although the API data were collected several months after the manual coding, potentially inflating shares and comments by allowing more time to build up, uploads within a fixed past period are not changed, yet the API still reports more uploads (suggesting manual under-coverage rather than a timing artefact). Moreover, despite this timing advantage, the API shows fewer views, indicating that view differences are not explained by timing alone and likely reflect differences in coverage or measurement.

Additionally, the TikTok Research API introduced challenges affecting data quality. Transcriptions were available for approximately 10% of videos, limiting the scope of linguistic analysis. Inconsistent metadata accuracy, such as discrepancies in engagement metrics, further complicates the reliability of the findings. Finally, methodological constraints warrant consideration. The study's timeframe (2020–2022) may not reflect longer-term trends, whilst the integration of manual calibration with API data risks introducing inconsistencies. These limitations suggest caution in interpreting the results and highlight avenues for future research, such as expanding the sample diversity and refining data collection methods.

Conclusion

This study set out to interrogate the reliability, representativeness, and ethical dimensions of TikTok Research API, using a large, balanced dataset to examine how methodological choices shape the validity of social media research. By systematically addressing RQ1–RQ3, it has shown that while the API offers unprecedented opportunities for structured, large-scale analysis, its outputs are neither neutral nor complete. Algorithmic and demographic biases, metric inconsistencies, and coverage gaps are not peripheral flaws but structural features of the data environment, features that, if left unexamined, risk distorting scholarly narratives and policy-relevant insights. Yet these limitations are tractable: multi-method integration, bias-aware modelling, and transparent reporting can materially improve both the equity and reproducibility of platform-based research.

This critical assessment highlights the API's dual character. On the one hand, it provides compliant and scalable access to multimodal data, and offers a legally sanctioned means of integrating quantitative and qualitative analyses (Nguyen & Diederich, 2023). On the other hand, it exhibits technical instability, systematic biases towards American and male creators, and restrictive ethical constraints. Whilst the API enhances computational social science by enabling diverse analytical approaches, its shortcomings in data quality, transparency, and accessibility necessitate a multi-method strategy to ensure valid and equitable insights. By doing so, this research contributes to ongoing debates on platform governance and methodological pluralism, demonstrating how comparative strategies can offset the deficiencies of single-source datasets.

Enhancing the reliability and representativeness of social media data remains essential for robust societal understanding and for fostering a more equitable digital research ecosystem.

Future work should adopt integrated approaches, combining API data with manual scraping or third-party platforms to cross-validate findings and address coverage gaps. Documenting discrepancies enhances transparency (Bezjak et al., 2018), while post-weighting adjustments can counter demographic biases (Ohme et al., 2024). Platform developers, for their part, should increase retrieval limits, expand transcription coverage, clarify metric definitions, and broaden access beyond academic institutions to promote inclusivity.

Finally, these findings open pathways for comparative API studies, for example, across Douyin, YouTube Shorts, and Instagram Reels, and for deeper examination of transcription biases and algorithmic opacity, particularly in relation to misinformation and political mobilisation. Leveraging automated tools or language models offers further potential, provided ethical standards are upheld. In this way, social media research can evolve to meet contemporary challenges while remaining methodologically rigorous, transparent, and fair.

ORCID iDs

Dan Bai https://orcid.org/0009-0000-1376-7655
Yan Gu https://orcid.org/0000-0001-6093-3919

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Liu, S., Sloan, L., Al Baghal, T., Williams, M., Jessop, C., & Serôdio, P. (2025). Linking survey with Twitter data: Examining associations among smartphone usage, privacy concern and Twitter linkage consent. *International Journal of Social Research Methodology*, *28*(1), 71–85. https://doi.org/10.1080/13645579.2023.2299482
2. https://apify.com/clockworks/tiktok-scraper.
3. https://tikapi.io/.
4. Vombatkere, K., Mousavi, S., Zannettou, S., Roesner, F., & Gummadi, K. P. (2024). Tiktok and the art of personalization: Investigating exploration and exploitation on social media feeds. *Proceedings of the ACM Web Conference 2024*, 3789–3797. https://doi.org/10.1145/3589334.3645600.
5. https://tiktok-audit.com/about/ and https://tiktok-audit.com/blog/2023/the-TikTok-research-API-falls-woefully-short/.
6. Krotov, V., Johnson, L., Murray State University, Silva, L., & University of Houston. (2020). Legality and ethics of web scraping. *Communications of the Association for Information Systems*, *47(1)*, 539–563. https://doi.org/10.17705/1CAIS.04724.
7. Divon, T., & Ebbrecht-Hartmann, T. (2023). PERFORMING DEATH AND TRAUMA? PARTICIPATORY MEM(E)ORY AND THE HOLOCAUST IN TIKTOK #POVCHALLENGES. *AoIR Selected Papers of Internet Research*, *2022*. https://doi.org/10.5210/spir.v2022i0.12995

8. Official TikTok business partners: https://partners.tiktok.com/directory/pc/en?rid=xm39skrb37&cspecialties=304%2C505%2C504%2C501%2C503%2C502%2C103%2C104.
9. https://anonymous.4open.science/r/r-tiktok-research-api-4E35/single_user_video.
10. How to create a batch compliance task: https://developers.tiktok.com/doc/batch-compliance-apis?enter_method=left_navigation.

References

Acker, A., & Kreisberg, A. (2020). Social media data archives in an API-driven world. *Archival Science*, *20*, 105–123. https://doi.org/10.1007/s10502-019-09325-9

Agrawal, E. (2024). *Going viral: An analysis of advertising of technology products on TikTok*. arXiv. https://doi.org/10.48550/ARXIV.2402.00010

Alexandre, I., Jai-sung Yoo, J., & Murthy, D. (2022). Make tweets great again: Who are opinion leaders, and what did they tweet about Donald Trump? *Social Science Computer Review*, *40*(6), 1456–1477. https://doi.org/10.1177/08944393211008859

Ambrazaitis, G., & House, D. (2017). Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings. *Speech Communication*, *95*, 100–113. https://doi.org/10.1016/j.specom.2017.08.008

Arora, A., Bansal, S., Kandpal, C., Aswani, R., & Dwivedi, Y. (2019). Measuring social media influencer index-insights from Facebook, Twitter and Instagram. *Journal of Retailing and Consumer Services*, *49*(C), 86–101. https://doi.org/10.1016/j.jretconser.2019.03.012

Arvaniti, A. (2020). The phonetics of prosody. In A. Arvaniti (Ed.), *Oxford research encyclopedia of linguistics* (pp. 86–101). Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.411

Bak-Coleman, J. (2023, February 22). *TikTok's API guidelines are a minefield for researchers | TechPolicy.Press*. *Tech Policy Press*. https://techpolicy.press/tiktoks-api-guidelines-are-a-minefield-for-researchers.

Bezjak, S., Clyburne-Sherin, A., Conzett, P., Fernandes, P., Görögh, E., Helbig, K., Kramer, B., Labastida, I., Niemeyer, K., Psomopoulos, F., Ross-Hellauer, T., Schneider, R., Tennant, J., Brinken, H., Heller, L., Verbakel, E., et al. (2018). Open Science Training Handbook. *Zenodo*. https://zenodo.org/records/1212496.

Brown, L., & Prieto, P. (2021). Gesture and prosody in multimodal communication. In M. Haugh, D. Z. Kádár, & M. Terkourafi (Eds.), *The Cambridge handbook of sociopragmatics* (1st ed., pp. 430–453). Cambridge University Press. https://doi.org/10.1017/9781108954105.023

Bruns, Axel (2019). After the 'APIcalypse': social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Burnat, F. A. D., & Davidson, B. I. (2025). *The accountability paradox: How platform API restrictions undermine AI transparency mandates*. arXiv. https://doi.org/10.48550/ARXIV.2505.11577

Cervi, L., & Divon, T. (2023). Playful activism: Memetic performances of Palestinian resistance in Tiktok #challenges. *Social Media + Society*, *9*(1), 20563051231157607. https://doi.org/10.1177/20563051231157607

Chang, A. (2018). The Facebook and Cambridge Analytica scandal, explained with a simple diagram. *Vox*. https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram.

Chen, Z., & Zhang, Q. (2021). A survey study on successful marketing factors for Douyin(TikTok). In F. F.-H. Nah & K. Siau (Eds.), *HCI in business, government and organizations* (pp. 22–42). Springer International Publishing. https://doi.org/10.1007/978-3-030-77750-0_2

Cohn, N. (2013). *The visual language of comics: Introduction to the structure and cognition of sequential images*. Bloomsbury Academic, An Imprint of Bloomsbury Pub. Plc.

Cohn, N., Engelen, J., & Schilperoord, J. (2019). The grammar of emoji? Constraints on communicative pictorial sequencing. *Cognitive Research: Principles and Implications*, *4*(1), 33. https://doi.org/10.1186/s41235-019-0177-0

Corso, F., Pierri, F., & De Francisci Morales, G. (2024). What we can learn from TikTok through its Research API. *Companion Proceedings of the 16th ACM Web Science Conference*, 110–114. https://doi.org/10.1145/3630744.3663611

Davidson, B. I., Wischerath, D., Racek, D., Parry, D. A., Godwin, E., Hinds, J., van der Linden, D., Roscoe, J. F., Ayravainen, L., & Cork, A. G. (2023). Platform-controlled social media APIs threaten open science. *Nature Human Behaviour*, *7*(12), 2054–2057. https://doi.org/10.1038/s41562-023-01750-2

De Oliveira Santini, F., Ladeira, W. J., Pinto, D. C., Herter, M. M., Sampaio, C. H., & Babin, B. J. (2020). Customer engagement in social media: A framework and meta-analysis. *Journal of the Academy of Marketing Science*, *48*(6), 1211–1228. https://doi.org/10.1007/s11747-020-00731-5

Entrena-Serrano, C. (2025). Watch, scroll, repeat: How interface design shapes consumptive curation affordances on Tiktok. *Social Media + Society*, *11*(3), Article 20563051251358529. https://doi.org/10.1177/20563051251358529

Entrena-Serrano, C., Degeling, M., Romano, S., & Çetin, R. B. (2025). *TikTok's research API: Problems without explanations*. arXiv. https://doi.org/10.48550/ARXIV.2506.09746

European Centre for Algorithmic Transparency. (2025). *FAQs: DSA data access for researchers - European Commission*. European Commission. Available from: https://algorithmic-transparency.ec.europa.eu/news/faqs-dsa-data-access-researchers-2025-07-03_en

Faust, W. (2024). *Wolfgang42/ftfyweb* [smarty]. (Original work published 2017). https://github.com/wolfgang42/ftfyweb

Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021). *Quantifying bias in automatic speech recognition*. arXiv. https://doi.org/10.48550/ARXIV.2103.15122

Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication*, *35*(4), 665–668. https://www.tandfonline.com/doi/full/10.1080/10584609.2018.1477506.

Haime, Z., & Biddle, L. (2025). Exploring mental health content moderation and well-being tools on social media platforms: Walkthrough analysis. *JMIR Human Factors*, *12*, e69817. https://doi.org/10.2196/69817

Herring, S. C. (2009). Web content analysis: Expanding the paradigm. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International handbook of internet research* (pp. 233–249). Springer. https://doi.org/10.1007/978-1-4020-9789-8_14

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kirkpatrick, C. E., & Lawrie, L. L. (2024). Tiktok as a source of health information and misinformation for young women in the united states: Survey study. *JMIR Infodemiology*, *4*, e54663. https://doi.org/10.2196/54663

Kita, H., Kitamura, Y., Hioki, H., Sakai, H., & Lin, Donghui. (2022). *The Practice of Basic Informatics 2022* (pp. 1–236). Japan: Kyoto University. http://hdl.handle.net/2433/289713.

Klug, D., Qin, Y., Evans, M., & Kaufman, G. (2021). Trick and please. A mixed-method study on user assumptions about the tiktok algorithm. *13th ACM Web*. *Science Conference*, *2021*, 84–92. https://doi.org/10.1145/3447535.3462512

Lachief. (2023). *Analisa.io*. Lachief. https://www.lachief.io/tool-posts/analisa-io

Lazer, D., Hargittai, E., Freelon, D., Gonzalez-Bailon, S., Munger, K., Ognyanova, K., & Radford, J. (2021). Meaningful measures of human society in the twenty-first century. *Nature*, *595*(7866), 189–196. https://doi.org/10.1038/s41586-021-03660-7

Leedham, M., Lillis, T., & Twiner, A. (2020). Exploring the core 'preoccupation' of social work writing: A corpus-assisted discourse study. *Journal of Corpora and Discourse Studies*, *3*(0), 1. https://doi.org/10.18573/jcads.26

Liang, LX. (2021). *Research on how to perceive their behavior for international high school students based on using tiktok with semi-structured interview*. 796–799. https://doi.org/10.2991/assehr.k.210407.151

Literat, I. (2021). "Teachers act like we're robots": TikTok as a window into youth experiences of online learning during COVID-19. *AERA Open*, *7*(1), 1–15, 233285842199553. https://doi.org/10.1177/2332858421995537

Lomborg, S., & Bechmann, A. (2014). Using APIs for data collection on social media. *The Information Society*, *30*(4), 256–265. https://doi.org/10.1080/01972243.2014.915276

Lu, Y., & Shen, C. (2023). Unpacking multimodal fact-checking: Features and engagement of fact-checking videos on Chinese TikTok (Douyin). *Social Media + Society*, *9*(1). https://doi.org/10.1177/20563051221150406

Markl, N. (2022). Language variation and algorithmic bias: Understanding algorithmic bias in British English automatic speech recognition. *2022 ACM conference on Fairness*. *Accountability, and Transparency*, 521–534. https://doi.org/10.1145/3531146.3533117

Mauchand, M., Vergis, N., & Pell, M. D. (2020). Irony, prosody, and social impressions of affective stance. *Discourse Processes*, *57*(2), 141–157. https://doi.org/10.1080/0163853X.2019.1581588

McNeill, D. (2005). *Gesture and thought*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226514642.001.0001

Mimizuka, K., Brown, M. A., Yang, K.-C., & Lukito, J. (2025). *Post-Post-API Age: Studying Digital Platforms in Scant Data Access Times*. arXiv. https://doi.org/10.48550/ARXIV.2505.09877

Moir, A. (2023). The Use of TikTok for Political Campaigning in Canada: The Case of Jagmeet Singh. *Social Media + Society*, *9*(1). https://doi.org/10.1177/20563051231157604 (Original work published 2023)

Mosnar, M., Skurla, A., Pecher, B., Tibensky, M., Jakubcik, J., Bindas, A., Sakalik, P., & Srba, I. (2025). Revisiting algorithmic audits of tiktok: Poor reproducibility and short-term validity of findings. *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3357–3366. https://doi.org/10.1145/3726302.3730293

Nguyen, H., & Diederich, M. (2023). How civil are comments on tiktok's educational videos? Insights for learning at scale. *Proceedings of the Tenth ACM Conference on Learning @ Scale*, 292–296. https://doi.org/10.1145/3573051.3596174

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology: CB*, *21*(19), 1641–1646. https://doi.org/10.1016/j.cub.2011.08.031

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B. B., & Robinson, T. N. (2024). Digital trace data collection for social media effects research: APIs, data donation, and (screen) tracking. *Communication Methods and Measures*, *18*(2), 124–141. https://doi.org/10.1080/19312458.2023.2181319

Parikh, N. (2025, December 19). *How to increase tiktok engagement: 16 expert insights*. Socialinsider Blog; Socialinsider. https://www.socialinsider.io/blog/how-to-get-more-engagement-on-tiktok/

Parisi, L., Mulargia, S., Comunello, F., Bernardini, V., Bussoletti, A., Nisi, C. R., Russo, L., Campagna, I., Lanfranchi, B., Croci, I., Grassucci, E., & Gesualdo, F. (2023). Exploring the vaccine conversation on TikTok in Italy: Beyond classic vaccine stances. *BMC Public Health*, *23*(1), 880. https://doi.org/10.1186/s12889-023-15748-y

Pascual, C. (2020). R API tutorial: Getting started with APIs in R. *Dataquest*. https://www.dataquest.io/blog/r-api-tutorial/

Pearson, G. D. H., Silver, N. A., Robinson, J. Y., Azadi, M., Schillo, B. A., & Kreslake, J. M. (2025). Beyond the margin of error: A systematic and replicable audit of the TikTok research API. *Information, Communication & Society*, *28*(3), 452–470. https://doi.org/10.1080/1369118x.2024.2420032

Pearson, J. (2019). The human imagination: The cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*, *20*(10), 624–634. https://doi.org/10.1038/s41583-019-0202-9

Pearson, J., Naselaris, T., Holmes, E. A., & Kosslyn, S. M. (2015). Mental imagery: Functional mechanisms and clinical applications. *Trends in Cognitive Sciences*, *19*(10), 590–602. https://doi.org/10.1016/j.tics.2015.08.003

Perriam, J., Birkbak, A., & Freeman, A. (2020). Digital methods in a post-API environment. *International Journal of Social Research Methodology*, *23*(3), 277–290. https://doi.org/10.1080/13645579.2019.1682840

Pfeffer, J., Mayer, K., & Morstatter, F. (2018). Tampering with twitter's sample API. *EPJ Data Science*, *7*(1), 50. https://doi.org/10.1140/epjds/s13688-018-0178-0

Pinto, G., Bickham, C., Salkar, T., Luceri, L., & Ferrara, E. (2024a). *Tracking the 2024 us presidential election chatter on tiktok: A public multimodal dataset*. arXiv. https://doi.org/10.48550/ARXIV.2407.01471

Pinto, G., Burghardt, K., Lerman, K., & Ferrara, E. (2024b). *Get-tok: A genai-enriched multimodal tiktok dataset documenting the 2022 attempted coup in Peru*. arXiv. https://doi.org/10.48550/ARXIV.2402.05882

Prieto, P., & Espinal, M. T. (2020). Negation, prosody, and gesture. In V. Déprez & M. T. Espinal (Eds.), *The Oxford handbook of negation* (1st ed., pp. 677–693). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198830528.013.34

Reisach, U. (2021). The responsibility of social media in times of societal and political manipulation. *European Journal of Operational Research*, *291*(3), 906–917. https://doi.org/10.1016/j.ejor.2020.09.020

Rejeb, A., Rejeb, K., Appolloni, A., Treiblmaier, H., Iranmanesh, M., et al. (2024). Mapping the scholarly landscape of TikTok (Douyin): A bibliometric exploration of research topics and trends. *Digital Business*, *4*(1), 1–23. https://doi.org/10.1016/j.digbus.2024.100075

Rogers, R., & Zhang, X. (2024). The Russia–Ukraine war in Chinese social media: LLM analysis yields a bias toward neutrality. *Social Media + Society*, *10*(2), 1–12. https://doi.org/10.1177/20563051241254379

Rossi, A., & Lenzini, G. (2020). Transparency by design in data-informed research: A collection of information design patterns. *Computer Law & Security Review*, *37*(105402), 1–22. https://doi.org/10.1016/j.clsr.2020.105402

Ruz, Santiago Sordo, Degeling, Martin, Meßmer, Kathy, et al. (2023). The Research API falls woefully short | auditing TikTok. *auditing TikTok blog*. https://tiktok-audit.com/blog/2023/the-TikTok-research-API-falls-woefully-short/. Access on December 31 2025.

Salganik, M. J. (2018). *Bit by bit: Social research in the digital age*. Princeton University Press.

Sari, L., Hasegawa-Johnson, M., & Yoo, C. (2021). Counterfactually fair automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3515–3525. https://doi.org/10.1109/taslp.2021.3126949

Sato, M. (2023, February 21). *Researchers will get access to TikTok data-Pending company approval*. The Verge. https://www.theverge.com/2023/2/21/23604737/tiktok-research-api-expansion-public-user-data-transparency

Sharma, K., Ferrara, E., & Liu, Y. (2022). Characterizing online engagement with disinformation and conspiracies in the 2020 U.S. presidential election. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*(1), 908–919. https://doi.org/10.1609/icwsm.v16i1.19345

Stevens, G., O'Donnell, V. L., & Williams, L. (2015). Public domain or private data? Developing an ethical approach to social media research in an inter-disciplinary project. *Educational Research and Evaluation*, *21*(2), 154–167. https://doi.org/10.1080/13803611.2015.1024010

Sweney, M. (2022, October 4). TikTok reports $1bn turnover across international markets. *The Guardian*. https://www.theguardian.com/technology/2022/oct/04/tiktok-reports-1bn-turnover-across-international-markets

Thole, A. N. (2022). *Tiktok forensic scraper to retrieve user video details*. 3998505 Bytes. https://doi.org/10.25394/PGS.21676727.V1

TikTok. (2024a). Digital services act: Publishing our second transparency report on content moderation in Europe. *Newsroom|TikTok*. https://newsroom.tiktok.com/en-eu/digital-services-act-second-transparency-report

TikTok. (2024b). *Terms of service*. TikTok. https://www.tiktok.com/legal/page/eea/terms-of-service/en

TikTok for Developers. (2024). *Research API*. TikTok for Developers. https://developers.tiktok.com/products/research-api/

TikTok Newsroom. (2021, April 6). Introducing auto captions. *Newsroom|TikTok; ByteDance*. https://newsroom.tiktok.com/introducing-auto-captions?lang=en-GB

Transparency Centre. (2024). Publishing with Meta content library and content library API. *Meta for Developers*. https://developers.facebook.com/docs/content-library-and-api/citations/

Tromble, R. (2021). Where have all the data gone? A critical reflection on academic digital research in the Post-API age. *Social Media + Society*, *7*(1), 1–8. https://doi.org/10.1177/2056305121988929

Venkatagiri, S. (2023, February 23). Researcher beware: Four red flags with the TikTok API's terms of service [substack newsletter]. *Technomoral*. https://technomoral.substack.com/p/researcher-beware-four-red-flags

Wassink, A. B., Gansen, C., & Bartholomew, I. (2022). Uneven success: Automatic speech recognition and ethnicity-related dialects. *Speech Communication*, *140*, 50–70. https://doi.org/10.1016/j.specom.2022.03.009

X Developer Platform. (2024). Non-commercial use of the X API. X. https://developer.x.com/en/developer-terms/commercial-terms

Yang, C. (2022). *Bias in short-video recommender systems: User-centric evaluation on TikTok*. [Master's paper, University of North Carolina]. https://cdr.lib.unc.edu/concern/masters_papers/v405sm32s

Zeitzoff, T. (2011). Using social media to measure conflict dynamics: An application to the 2008–2009 Gaza conflict. *Journal of Conflict Resolution*, *55*(6), 938–969. https://doi.org/10.1177/0022002711408014

Zhao, Y. (2024). Tiktok and researcher positionality: Considering the methodological and ethical implications of an experimental digital ethnography. *International Journal of Qualitative Methods*, *23*, 1–11. https://doi.org/10.1177/16094069231221374

Author Biographies

Dan Bai is a PhD student in ISER at University of Essex. Her research examines digital media platforms, linguistic genre dynamics, and the methodological challenges of large-scale social data. She specialises in statistical modelling, corpus-based analysis, and the operationalisation of qualitative concepts for quantitative research.

Dr Yan Gu is a Lecturer in the Department of Psychology at the University of Essex. He received his PhD in Psycholinguistics from Tilburg University. Previously, he has been a postdoctoral researcher at University College London, where he now serves as an Honorary Research Fellow. His research examines time perception, multimodal language use, and the cognitive and social consequences of linguistic behaviour. He works across speech, text, gesture, sign, and gaze to investigate how language shapes thought, learning, and decision-making. He also conducts interdisciplinary research with economists and social scientists on topics including ageing, migration, financial behaviour, and well-being. His work has been published in leading journals across psychology, linguistics, and cognitive science.