SUPPLEMENTARY MATERIAL
On the (Un)Predictability of User Watching Behavior with Short Format Videos on TikTok

Carolina Coimbra Vieira[1,2,3,*], Sepehr Mousavi[2,3], Abhisek Dash[2], Krishna P. Gummadi[2], Oshrat Ayalon[4] and Savvas Zannettou[5]

[1]*Max Planck Institute for Demographic Research (MPIDR), Konrad-Zuse-Str. 1, 18057, Rostock, Germany*
[2]*Max Planck Institute for Software Systems (MPI-SWS), Campus E1 4, D-66123, Saarbrücken, Germany*
[3]*Saarland University, Campus, 66123 Saarbrücken, Germany*
[4]*University of Haifa, Abba Khoushy Ave 199, 3498838, Haifa, Israel*
[5]*Delft University of Technology (TU Delft), Mekelweg 5, 2628, Delft, Netherlands*

1. Demographics

Figure 1 shows the age and sex distribution of the 80 participants in our experiment and age-sex distribution of TikTok's U.S. users. Tables 1 and 2 present an overview of the participants' demographics and TikTok usage characteristics, respectively. We report the most prevalent categories for each demographic group and emphasize the most common category in bold.

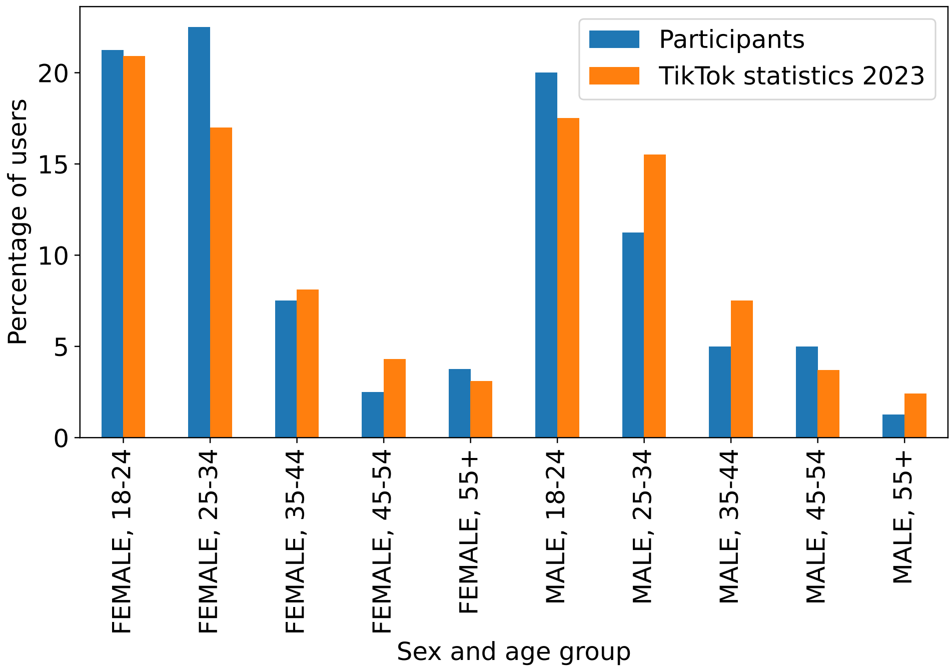


Figure 1: Participants of the experiment (N=80) and TikTok user distributions by age and sex (Statista, Oct 2023).

*Corresponding author.
coimbravieira@demogr.mpg.de (C. C. Vieira)
https://carolcoimbra.github.io/ (C. C. Vieira)
0000-0003-3156-4151 (C. C. Vieira)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Table 1
Demographics of Participants (N = 80)

Group	Categories	Count (%)
Sex	Women	46 (57.50%)
	Men	**34 (42.50%)**
Age	18-24 years old	25 (31.25%)
	25-34 years old	**34 (42.50%)**
	35-44 years old	10 (12.50%)
	45-54 years old	5 (6.25%)
	$\geq$ 55 years old	6 (7.50%)
Race/Ethnicity (multiple choice)	**White**	**36 (45.00%)**
	Asian or Asian American	13 (16.25%)
	Black or African American	11 (13.75%)
	Hispanic or Latino	11 (13.75%)
	...	
Language (multiple choice)	**English**	**80 (100.00%)**
	Spanish	14 (17.50%)
	Chinese	5 (6.25%)
	French	4 (5.00%)
	...	
Education	High school diploma	2 (2.50%)
	Some college, no degree	16 (20.00%)
	Associate's degree	10 (12.50%)
	Bachelor's degree	**39 (48.75%)**
	Advanced degree	13 (16.25%)
Employment Status (multiple choice)	**Employed full-time**	**34 (42.50%)**
	Not employed	12 (15.00%)
	Student	11 (13.75%)
	Employed part-time	10 (12.50%)
	...	
Political Affiliation	**Democrat**	**56 (70.00%)**
	Independent-Democrat	12 (15.0%)
	Independent-Republican	5 (6.25%)
	Republican	3 (3.75%)
	Strong Republican	3 (3.75%)
	No preference, closer to Democrat	1 (1.25%)
Annual Income	**< $5,000**	**11 (13.75%)**
	$5,000–$10,000	9 (11.25%)
	$10,000–$20,000	7 (8.75%)
	$20,000–$30,000	5 (6.26%)
	$30,000–$40,000	8 (10.00%)
	$40,000–$50,000	10 (12.50%)
	$50,000–$65,000	10 (12.50%)
	...	

Table 2
TikTok Usage Characteristics of Participants

Group	Categories	Count (%)
TikTok Usage Duration	Less than a month	3 (3.75%)
	1-6 months	6 (7.50%)
	6-12 months	12 (15.00%)
	1–3 years	**40 (50.00%)**
	More than 3 years	19 (23.75%)
Frequency of Use	Almost constantly	5 (6.25%)
	Several times a day	**37 (46.25%)**
	About once a day	13 (16.25%)
	Several times a week	18 (22.50%)
	Less often	7 (8.75%)
Engagement per Session	Most videos (almost all videos)	8 (10.00%)
	Many (more than every other video)	6 (7.50%)
	Half (every other video)	7 (8.75%)
	Moderate (few to half)	**37 (46.25%)**
	Few (1–2 videos)	19 (23.75%)
	None of them	3 (3.75%)
Daily Usage Time	< 10 minutes/day	15 (18.75%)
	10–30 minutes/day	**23 (28.75%)**
	31–60 minutes/day	17 (21.25%)
	1–2 hours/day	17 (21.25%)
	2–3 hours/day	4 (5.00%)
	More than 3 hours per day	4 (5.00%)
User Type	**Content consumer**	**70 (87.50%)**
	Equally consumer and creator	9 (11.25%)
	Content creator	1 (1.25%)
Account Type	**Personal account**	**73 (91.25%)**
	Both personal and business	5 (6.25%)
	Business account	2 (2.50%)
TikTok Access Context (multiple choice)	**When bored**	**16 (19.42%)**
	Before bed	13 (15.65%)
	During work breaks	10 (12.46%)
	While waiting briefly	9 (11.59%)
	In the restroom	8 (10.15%)
	While eating	6 (7.54%)
	Getting up	5 (6.67%)
	While family watches other content	5 (6.38%)
	While traveling	5 (6.38%)
	While commuting	3 (3.19%)
	...	

2. Experiment setup

Playlist: Figure 2 shows the cumulative duration of the playlist created for our experiment.

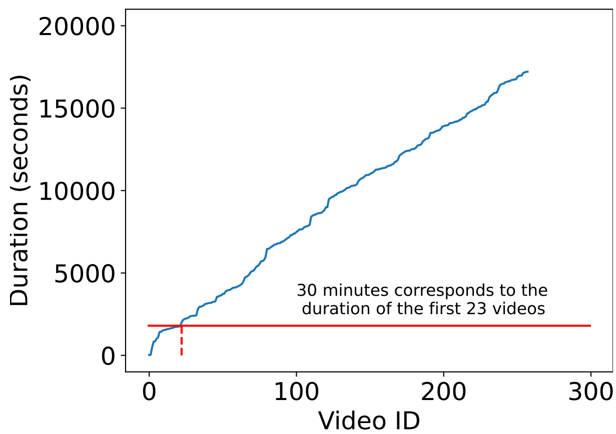
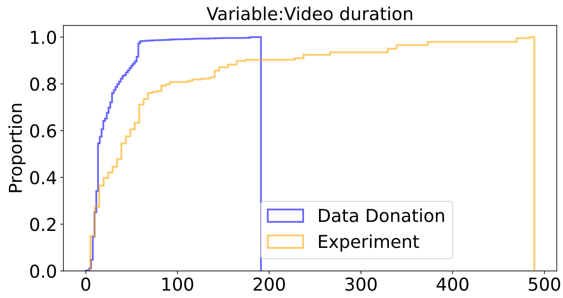


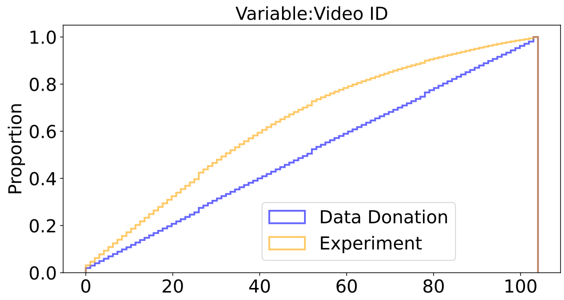
Figure 2: Cumulative duration of the playlist created for the controlled experiment.

3. Comparison between experimental and real-world datasets

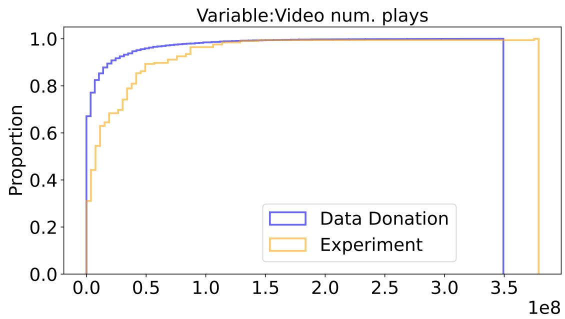
Figure 3 shows the variable comparison between our experimental dataset and the subset of the real-world dataset in North/Central America.



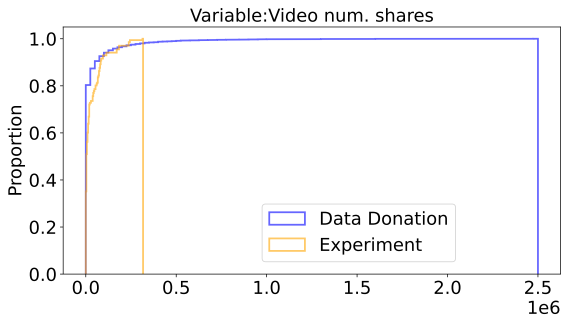
(a) The duration in seconds of videos.



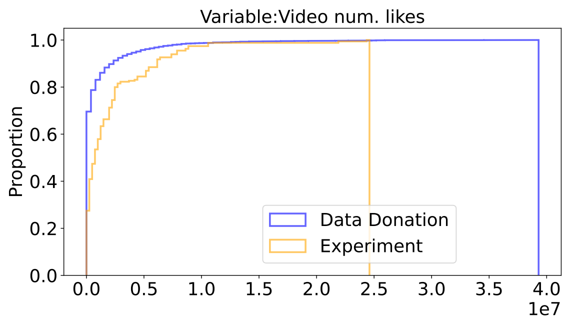
(b) The order in which the videos were watched.



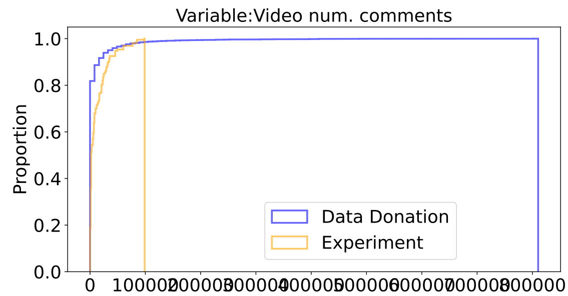
(c) The total number of times the video was played



(d) The total number of times the video was shared



(e) The total number of times the video was liked



(f) The total number of times the video received comments

Figure 3: CDF of the video duration, the order in which the video was watched, and the total number of times the video was played, shared, liked, and received comments.

4. Model

Features: Table 3 lists all the features used in our models, as well as their type and brief description.

Model specifications: Below, we report the seeds, models, and hyperparameters used by the classification models reported in our study. First, we present the Python code used to split the dataset into train and test sets. Next, we created a pipeline to normalize the data and randomly search to select the hyperparameters that optimize the performance of the classification model. The list of classifiers as well as the hyperparameters tested are shown in Table 4.

Table 3
Description of the features used in our models.

Feature name	Type	Description
Video ID	Numerical	Identifier for the video considering the order of its inclusion in the playlist (in the experimental setting) or the order in which the video is watched by the user (in the real-world setting).
Video duration	Numerical	Length of the video in seconds.
Video num. likes	Numerical	Number of likes the video has received.
Video num. shares	Numerical	Number of times the video has been shared.
Video num. comments	Numerical	Number of comments the video has received.
Video num. plays	Numerical	Number of times the video has been played.
Year born	Numerical	Birth year of the user.
Gender	Categorical	Gender reported by the user.
Race/Ethnicity	Categorical	Race/ethnicity reported by the user.
Language (e.g., English)	Numerical	For each language the value represents the proficiency level on a scale of 1 to 5, where 1 represents basic proficiency and 5 represents native.
Highest level of school	Numerical	Highest level of school reported by the user.
Employment status	Categorical	Employment status reported by the user.
Political leaning: Republican	Numerical	The value represents the degree of Republican-leaning.
Political leaning: Democrat	Numerical	The value represents the degree of Democratic-leaning.
Income	Numerical	The user's income level.
Interest Similarity	Numerical	Similarity (measured as a variation of the Jaccard Similarity) between the video's topics and the participants' topics of interest.
How long use TikTok	Numerical	Duration in months of how long the user has a TikTok account.
How often access TikTok	Numerical	Frequency of accessing TikTok.
How many videos engage with	Numerical	Number of videos with which the user interacts.
Avg time PER DAY in the past week using TikTok	Numerical	Average daily usage time in the past week the user spent on TikTok.
TikTok viewer	Categorical	Whether the user views content on TikTok (Yes/No).
TikTok creator	Categorical	Whether the user creates content on TikTok (Yes/No).
TikTok personal	Categorical	Whether the user uses TikTok for personal purposes (Yes/No).
TikTok business	Categorical	Whether the user uses TikTok for business purposes (Yes/No).
When accessing TikTok	Categorical	Moment when the participants watch TikTok.

```
from sklearn.model_selection import RandomizedSearchCV, train_test_split
from sklearn.pipeline import make_pipeline

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

rand_search = make_pipeline(StandardScaler(),
RandomizedSearchCV(classifier,
param_distributions = parameters,
n_iter=10,
cv=5,
random_state=42,
refit=True))
```

Table 4
Classification models' specifications.

	classifier	parameters
Logistic Regression	LogisticRegression(random_state=0)	penalty: [l2], solver: [lbfgs, liblinear, newton-cg, newton-cholesky, sag, saga], C: np.arange(0.025, 1, 0.25), class_weight: [balanced]
KNN	KNeighborsClassifier()	n_neighbors: range(3,30)
SVM	SVC(random_state=4)	kernel: [linear, poly, rbf, sigmoid], C: np.arange(0.025, 1, 0.25), gamma: [auto, scale], degree: range(1,6,1), class_weight: [balanced]
Decision Tree	DecisionTreeClassifier(random_state=4)	max_depth: range(1,50), min_samples_leaf: range(1,20), class_weight: [balanced]
Random Forest	RandomForestClassifier(random_state=4)	max_depth: range(1,100), min_samples_leaf: range(2,20), n_estimators: range(10,100,10), class_weight: [balanced]
MLP	MLPClassifier(random_state=4, max_iter=500)	hidden_layer_sizes: range(6,len(X.columns)-2), learning_rate: [constant], alpha:np.arange(0.0001, 0.001, 0.0001)

Model evaluation: Table 5 shows the performance of each model.

Table 5
Evaluation of models' performance on our experimental dataset using all the features.

Model	F1 Score	Accuracy	Precision	Recall
Logistic Regression	0.70	0.71	0.55	0.75
K Nearest Neighbors	0.68	0.74	0.66	0.44
SVM	0.71	0.74	0.58	0.68
Decision Tree	0.69	0.70	0.54	0.72
Random Forest	**0.74**	**0.77**	**0.66**	**0.63**
MLP	0.73	0.77	0.66	0.60