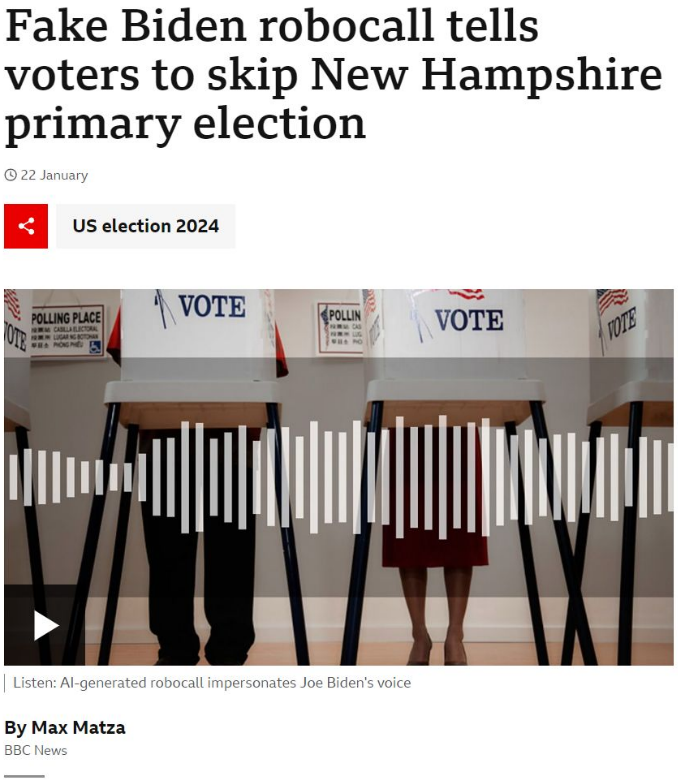
Sociolinguistically-informed educational trainings for audio deepfake discernment

Chloe Evered[1], Christine Mallinson[2], Lavon Davis[2], Vandana Janeja[2], Noshaba Basir Bhalli[2], Zahra Khanjani[2], Nehal Naqvi[2], Kifekachukwu Nwosu[3]

[1]Georgetown University, Washington, DC, USA
[2]University of Maryland, Baltimore County, Baltimore, MD, USA
[3]Rochester Institute of Technology, Rochester, NY, USA

BAAL 2024 | September 6, 2024
HuMaLa SIG



Fake Biden robocall tells voters to skip New Hampshire primary election

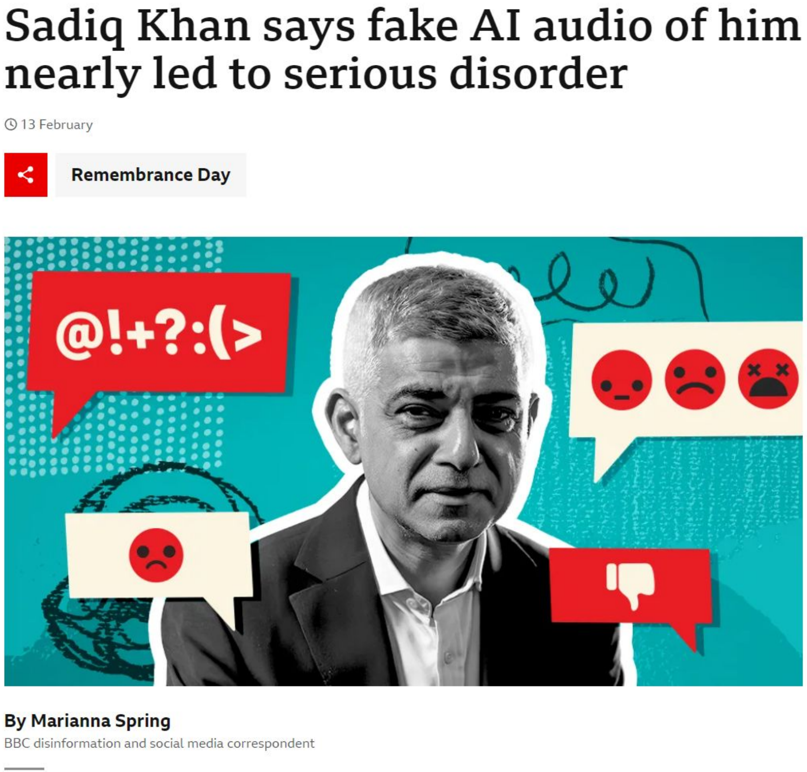
22 January

US election 2024

Listen: AI-generated robocall impersonates Joe Biden's voice

By Max Matza
BBC News

(Matza, 2024)



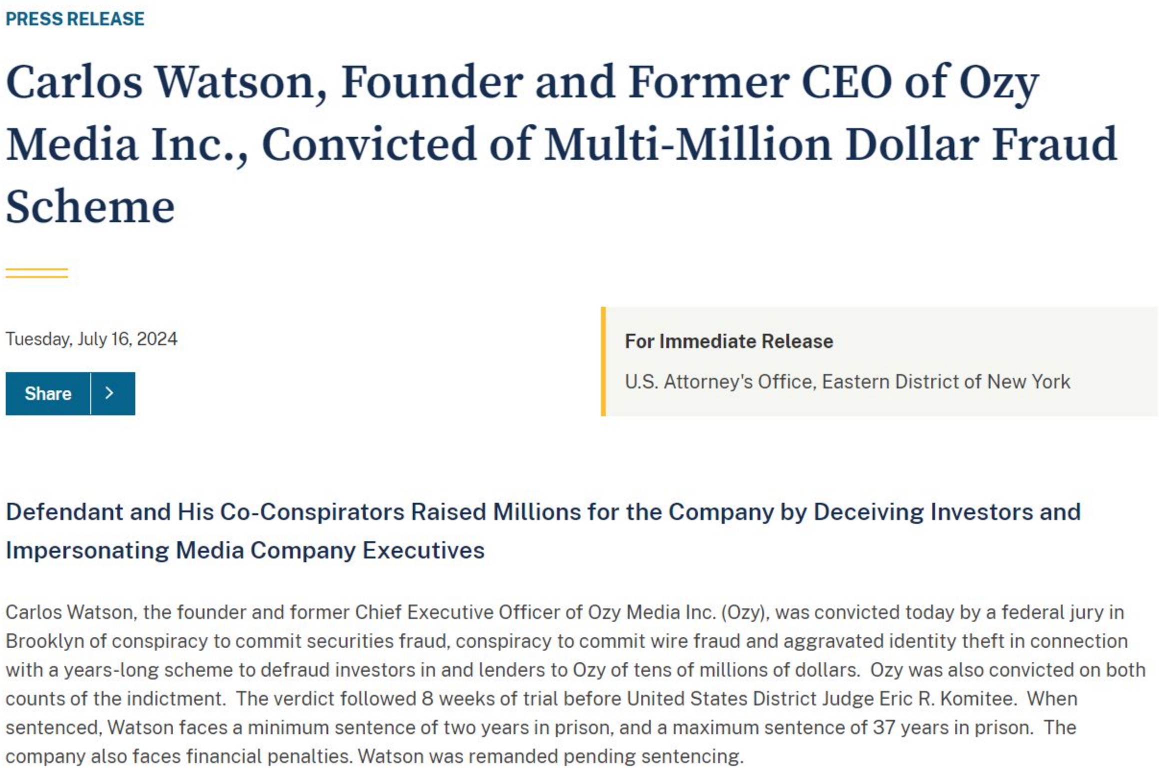
Sadiq Khan says fake AI audio of him nearly led to serious disorder

13 February

Remembrance Day

By Marianna Spring
BBC disinformation and social media correspondent

(Spring, 2024)



PRESS RELEASE

Carlos Watson, Founder and Former CEO of Ozy Media Inc., Convicted of Multi-Million Dollar Fraud Scheme

Tuesday, July 16, 2024

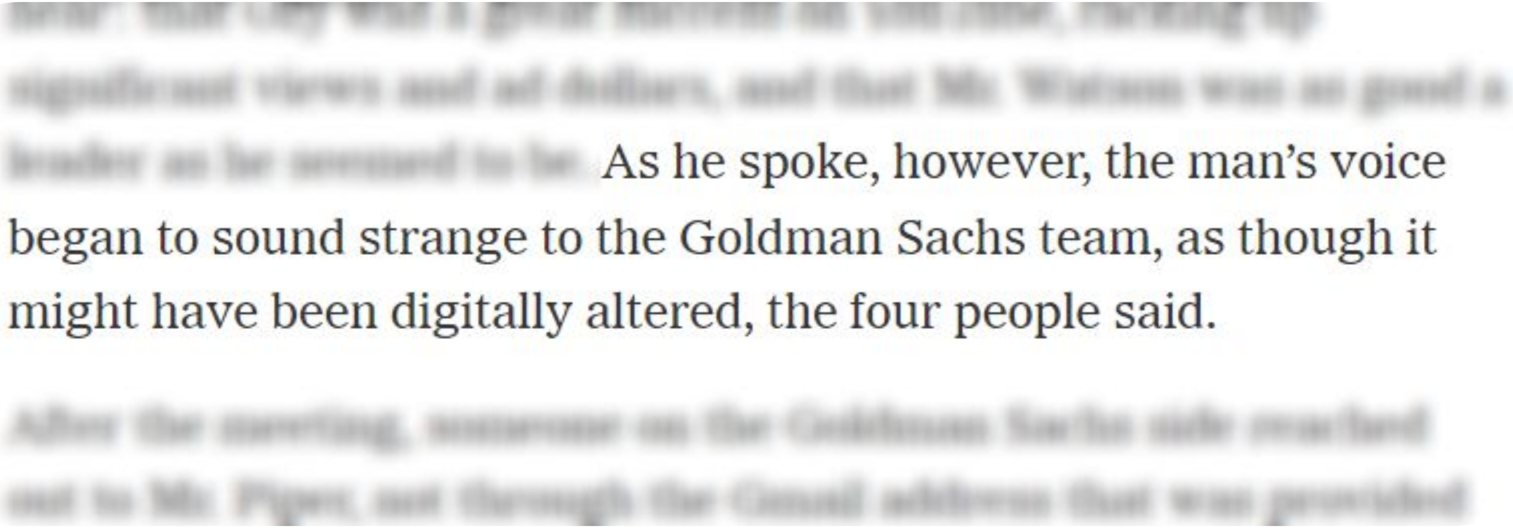
For Immediate Release

U.S. Attorney's Office, Eastern District of New York

Defendant and His Co-Conspirators Raised Millions for the Company by Deceiving Investors and Impersonating Media Company Executives

Carlos Watson, the founder and former Chief Executive Officer of Ozy Media Inc. (Ozy), was convicted today by a federal jury in Brooklyn of conspiracy to commit securities fraud, conspiracy to commit wire fraud and aggravated identity theft in connection with a years-long scheme to defraud investors in and lenders to Ozy of tens of millions of dollars. Ozy was also convicted on both counts of the indictment. The verdict followed 8 weeks of trial before United States District Judge Eric R. Komitee. When sentenced, Watson faces a minimum sentence of two years in prison, and a maximum sentence of 37 years in prison. The company also faces financial penalties. Watson was remanded pending sentencing.

(U.S. Attorney's Office, Eastern District of New York, 2024)



hear: that Ozy was a great success on YouTube, racking up significant views and ad dollars, and that Mr. Watson was as good a leader as he seemed to be. As he spoke, however, the man’s voice began to sound strange to the Goldman Sachs team, as though it might have been digitally altered, the four people said.

After the meeting, someone on the Goldman Sachs side reached out to Mr. Piper, not through the Gmail address that was provided

(Smith, 2021)

Typical Approaches to Deepfake Detection

- Spoofed audio countermeasures typically rely on improving algorithms to catch fakes, leading to a vicious cycle as audio deepfake generation methods become more sophisticated

(Chesney & Citron, 2019)

- Many different types of deepfake generation methods

(Khan et al., 2023)

- Require specialized knowledge or advanced computing skills to implement

Background

- Listeners have a highly attuned capacity to hear variation in others’ language and use it to pick up on social information

(Thomas, 2002; Purnell, Isardi, & Baugh, 1999)

- However, they can’t always pinpoint what specific features they are picking up on: some linguistic features are highly salient, while others fall below listeners’ conscious awareness

(Labov, 1972)

Background

- Linguistic bias can be mitigated through education

(See e.g. Baese-Berk, 2019; Boduch-Grabka & Lev-Ari, 2021; Godley et al., 2006; Mallinson & Charity Hudley, 2014, 2016; Reaser, 2006; Rickford & Rickford, 2007; Sweetland, 2006)

- Perceptual acuity can be honed with training

(Linebaugh & Roche, 2013, 2015)

Discovery Phase

- Inductive approach led by main author and students with training in (variationist) sociolinguistics
- Qualitatively noted distinguishing features for real, fake clips
- Collaboratively compared notes, resolved any divergences, and agreed upon feature selection and strategy of annotation

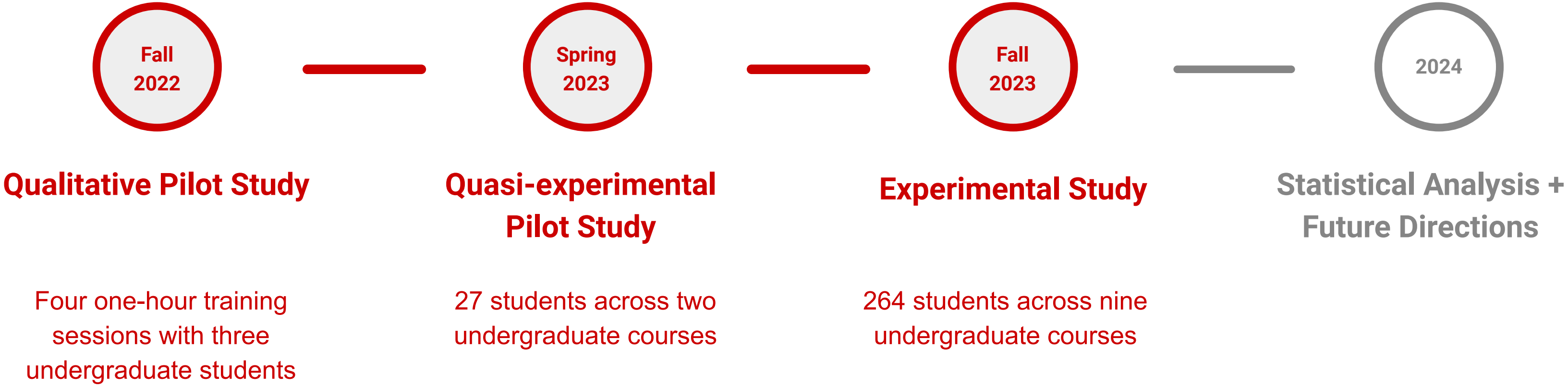
- Features must be:
  - Frequent
  - Discernible
  - Definable
- Sample of 344 genuine and spoof clips (Khanjani et al., 2023)

Expert-Defined Linguistic Features (EDLFs)

Pitch	Relative high or low tone of a speech sample
Pause	Break in speech production within a speech sample
Word Initial/Final Consonant Bursts	Release bursts of consonant stops /p/, /b/, /t/, /d/, /k/, and /g/
Intake/Outtake of Breath	Presence or absence of any audible intake or outtake of breath
Audio Quality	Overall qualitative estimation of the audio quality of a speech sample

EDLFs include commonly occurring, variable, and distinguishing phonetic and phonological characteristics of spoken English. *For each sample, the sociolinguist team members perceptually identified and identified the* ***presence or absence of these features and annotated any anomalies in their production****. As such, the labels indicate potential linguistic characteristics of real versus fake audio.*

EDLF Deepfake Discernment Training Phases



Fall 2022

Qualitative Pilot Study

Four one-hour training sessions with three undergraduate students

Spring 2023

Quasi-experimental Pilot Study

27 students across two undergraduate courses

Fall 2023

Experimental Study

264 students across nine undergraduate courses

2024

Statistical Analysis + Future Directions

Qualitative Pilot Study

Fall 2022

Four one-hour training sessions with three undergraduate students with no background in linguistics

- Students were able to **listen with a deeper intention** and **explain concepts from the training** to peers with minimal understanding

Qualitative Pilot Study

Fall 2022

Four one-hour training sessions with three undergraduate students with no background in linguistics

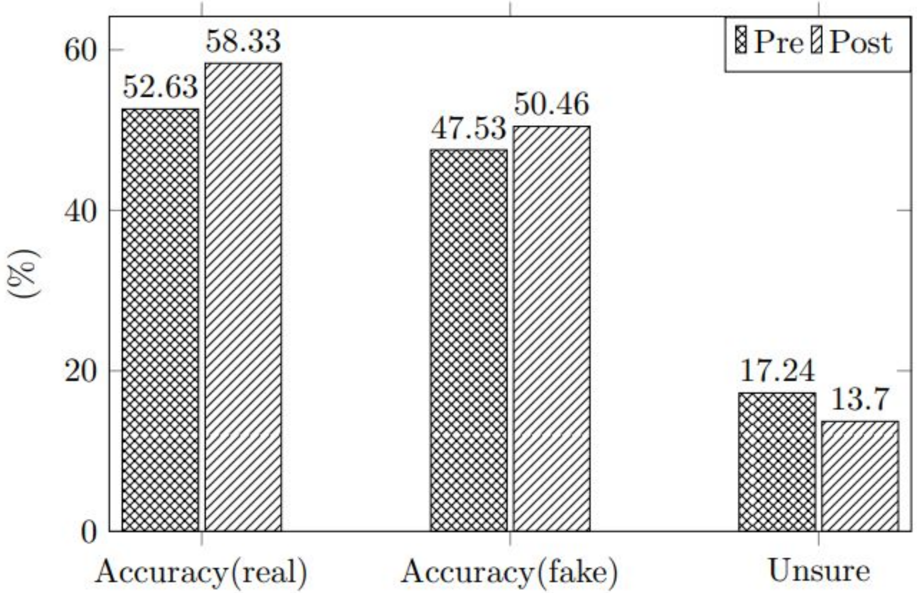
- Students were able to **listen with a deeper intention** and **explain concepts from the training** to peers with minimal understanding

“After the training I am confident to be able to distinguish [anomalous EDLFs] in an audio clip, listen much more carefully, considering the context of audio recordings, speaker background, additional noise etc., and approach this task without jumping straight to assumptions.”

“I learned about some of the formal [linguistic] indicators for a deepfake..., as well as training myself when to and when not to form a conclusion [about] the authenticity of an audio file.”

Quasi-experimental Pilot Study (Spring 2023)

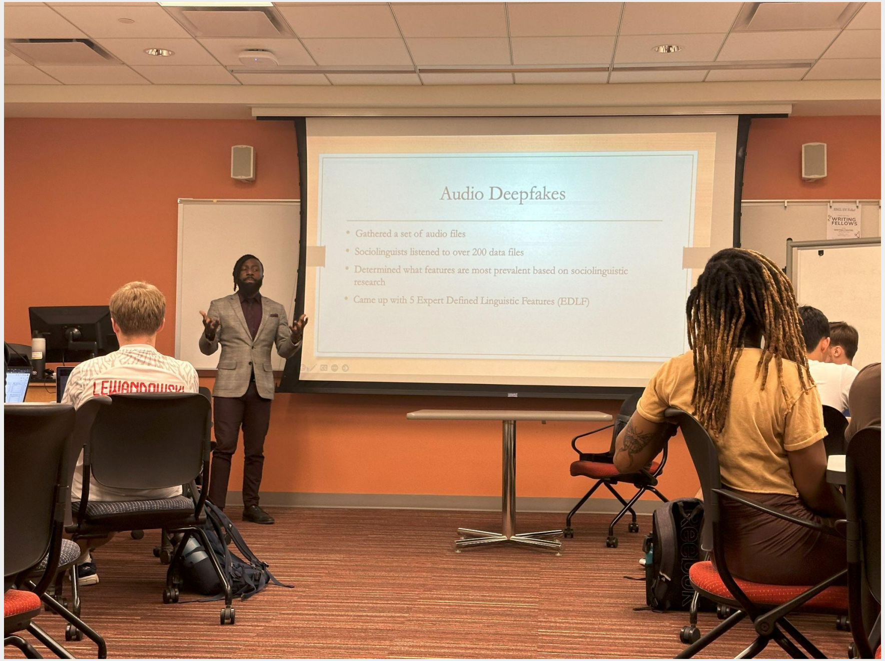
- 27 students across two introductory undergraduate courses
- Pre-survey
  - 20 audio clips (half real, half fake)
  - Real, fake, or unsure?
  - Open-ended questions
- 20-minute training session
  - Based on longer training session from Fall 2022
- Post-survey
  - Administered one month after the pre-survey
- Debrief



Results: Quasi-experimental Pilot Study

Experimental Study

Fall 2023
264 students



Audio Deepfakes

- Gathered a set of audio files
- Sociolinguists listened to over 200 data files
- Determined what features are most prevalent based on sociolinguistic research
- Came up with 5 Expert Defined Linguistic Features (EDLF)

Findings

Results from the pre- and post-tests revealed that training **increased confidence** for some students, yet this decrease in unsurety **did not always come with an similar increase deepfake discernment accuracy**

- While a significant number of students showed improvement, this wasn't significant in magnitude
- Female students in the experimental group showed a much greater decrease in unsurety, which disproportionately drives the overall trend we observe in the experimental group
- Students in the experimental group who reported English as their first language showed beneficial decrease in unsurety, but students' majors and fluency in other languages did not significantly predict performance after receiving training

Findings

Training led students to be **more skeptical** of genuine speech samples, leading them to label real clips as fake

- Of the 353 clips marked as “unsure” in the pre-test, 85% of such clips which represented a fake clip were correctly identified as fake in the post-test. However, only about 20% of such real clips were correctly identified.

Findings

The control group, who received a short reading about audio deepfakes, **also showed improvement**

- The control group showed significant improvement in their accuracy of identifying clips, with 4% improvement overall on all clips and real clips driving this difference.
- Students in the control group were significantly more accurate, by almost 10%, in their ability to correctly identifying short clips (<2 sec) as real or fake
- Non-computing majors in the control group conformed to the overall trend of significant improvement for all clips, while computing majors showed significant improvement for real clips only

Future directions for deepfake discernment training

- Longer training module
- Robust, holistic approach that incorporates digital media literacy education in tandem with perceptual sociolinguistic training
- Public-facing training accessible online (cisaad.umbc.edu)

References

Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review 107*(6), 1753–1760. https://doi.org/10.15779/Z38RV0D15J

Khan, A., Malik, K. M., Ryan, J., & Saravanan, M. (2023). Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures. In *Artificial Intelligence Review* (Vol. 56, Issue S1, pp. 513–566). Springer Science and Business Media LLC. https://doi.org/10.1007/s10462-023-10539-8

Khanjani, Z., Davis, L., Tuz, A., Nwosu, K., Mallinson, C., & Janeja, V. (2023). *Learning to listen and listening to learn: Spoofed audio detection through linguistic data augmentation*. Intelligence and Security Informatives: IEEE International Conference on Intelligence and Security Informatives, ISI 2023 (The 2023 IEEE International Conference in Intelligence and Security Informatics). Charlotte, NC: Oct. 2-3.

Linebaugh, G., and T. B. Roche. (2013). Learning to hear by learning to speak: The effect of articulatory training on Arab learners' English phonemic discrimination. *Australian Review of Applied Linguistics, 36*(2), 146–159. https://doi.org/10.1075/aral.36.2.02lin

Linebaugh, G., & Roche, T. B. (2015). Evidence that L2 production training can enhance perception. *Journal of Academic Language and Learning*, *9*(1), A1-A17. Retrieved from https://journal.aall.org.au/index.php/jall/article/view/326

Mallinson, C., & Charity Hudley, A.H. (2014). Partnering through science: Developing linguistic insight to address educational inequality for culturally and linguistically diverse students in U.S. STEM education. *Language and Linguistics Compass, 8*(1), 11–23. https://doi.org/10.1111/lnc3.12060

Matza, M. (2024). Fake Biden robocall tells voters to skip New Hampshire primary election. *BBC News*. https://www.bbc.co.uk/news/world-us-canada-68064247

Purnell, T., Idsardi, W., & Baugh, J. (1999). Perceptual and Phonetic Experiments on American English Dialect Identification. *Journal of Language and Social Psychology, 18*(1), 10-30. https://doi.org/10.1177/0261927X99018001002

Smith, B. (2021). Goldman Sachs, Ozy Media and a $40 million conference call gone wrong. *The New York Times*. Retrieved from https://www.nytimes.com/2021/09/26/business/media/ozy-media-goldman-sachs.html

Spring, M. (2024, February 13). Sadiq Khan says fake AI audio of him nearly led to serious disorder. *BBC News.* https://www.bbc.co.uk/news/uk-68146053

Thomas, E. R. (2002). Sociophonetic applications of speech perception experiments. *American Speech, 77*(2), 115-147. https://doi.org/10.1215/00031283-77-2-115

U.S. Attorney's Office, Eastern District of New York. (2024, July 16). *Carlos Watson, Founder and Former CEO of Ozy Media Inc., Convicted of Multi-Million Dollar Fraud Scheme.*. https://www.justice.gov/usao-edny/pr/carlos-watson-founder-and-former-ceo-ozy-media-inc-convicted-multi-million-dollar

Thank you & Questions

Community Infrastructure to Strengthen AI for Audio Deepfake Analysis (CISAAD)

This work was supported in part by Award #2210011. This work also included Data Science scholars supported through additional grants #1915714 and #2118285.

We also thank the students who participated in our study, their instructors Antione Tomlin and Mark Berczynski, and doctoral student Lavon Davis from the Language, Literacy & Culture Program for facilitating training sessions.

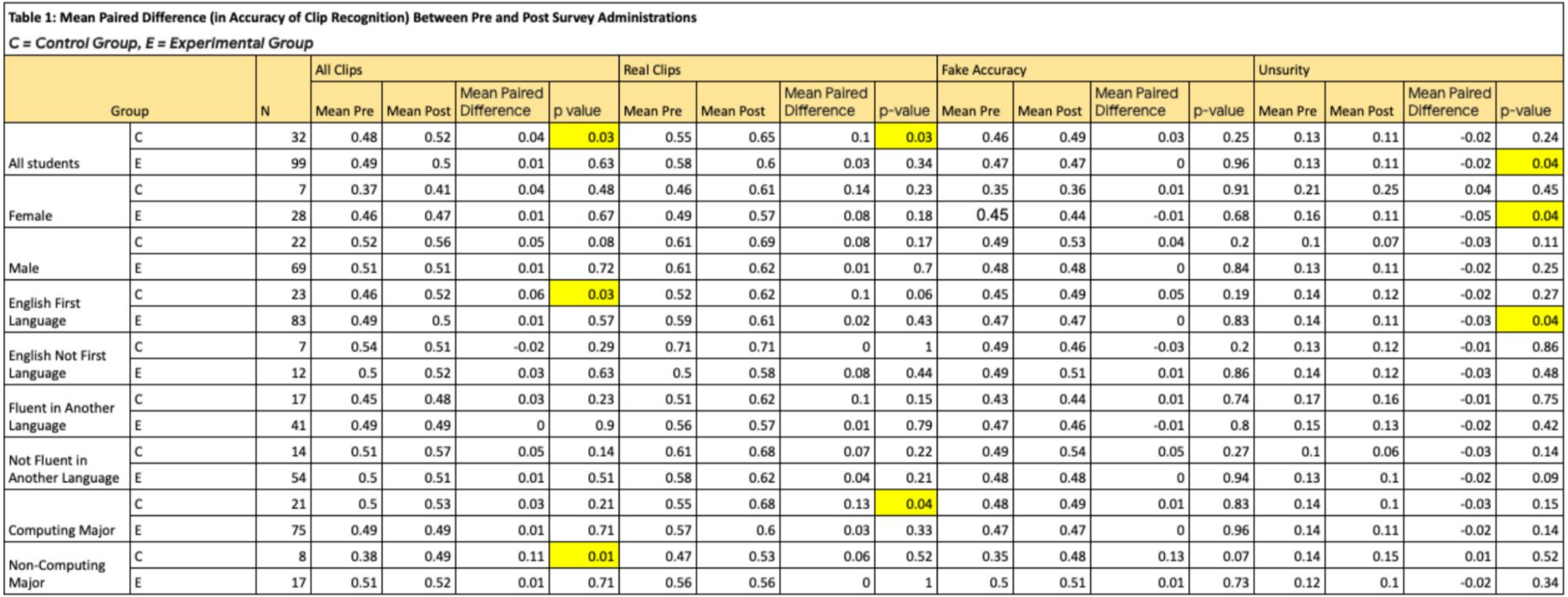


Table 1: Mean Paired Difference (in Accuracy of Clip Recognition) Between Pre and Post Survey Administrations

C = Control Group, E = Experimental Group

Group		N	All Clips				Real Clips				Fake Accuracy				Unsurity			
			Mean Pre	Mean Post	Mean Paired Difference	p value	Mean Pre	Mean Post	Mean Paired Difference	p-value	Mean Pre	Mean Post	Mean Paired Difference	p-value	Mean Pre	Mean Post	Mean Paired Difference	p-value
All students	C	32	0.48	0.52	0.04	0.03	0.55	0.65	0.1	0.03	0.46	0.49	0.03	0.25	0.13	0.11	-0.02	0.24
	E	99	0.49	0.5	0.01	0.63	0.58	0.6	0.03	0.34	0.47	0.47	0	0.96	0.13	0.11	-0.02	0.04
Female	C	7	0.37	0.41	0.04	0.48	0.46	0.61	0.14	0.23	0.35	0.36	0.01	0.91	0.21	0.25	0.04	0.45
	E	28	0.46	0.47	0.01	0.67	0.49	0.57	0.08	0.18	0.45	0.44	-0.01	0.68	0.16	0.11	-0.05	0.04
Male	C	22	0.52	0.56	0.05	0.08	0.61	0.69	0.08	0.17	0.49	0.53	0.04	0.2	0.1	0.07	-0.03	0.11
	E	69	0.51	0.51	0.01	0.72	0.61	0.62	0.01	0.7	0.48	0.48	0	0.84	0.13	0.11	-0.02	0.25
English First Language	C	23	0.46	0.52	0.06	0.03	0.52	0.62	0.1	0.06	0.45	0.49	0.05	0.19	0.14	0.12	-0.02	0.27
	E	83	0.49	0.5	0.01	0.57	0.59	0.61	0.02	0.43	0.47	0.47	0	0.83	0.14	0.11	-0.03	0.04
English Not First Language	C	7	0.54	0.51	-0.02	0.29	0.71	0.71	0	1	0.49	0.46	-0.03	0.2	0.13	0.12	-0.01	0.86
	E	12	0.5	0.52	0.03	0.63	0.5	0.58	0.08	0.44	0.49	0.51	0.01	0.86	0.14	0.12	-0.03	0.48
Fluent in Another Language	C	17	0.45	0.48	0.03	0.23	0.51	0.62	0.1	0.15	0.43	0.44	0.01	0.74	0.17	0.16	-0.01	0.75
	E	41	0.49	0.49	0	0.9	0.56	0.57	0.01	0.79	0.47	0.46	-0.01	0.8	0.15	0.13	-0.02	0.42
Not Fluent in Another Language	C	14	0.51	0.57	0.05	0.14	0.61	0.68	0.07	0.22	0.49	0.54	0.05	0.27	0.1	0.06	-0.03	0.14
	E	54	0.5	0.51	0.01	0.51	0.58	0.62	0.04	0.21	0.48	0.48	0	0.94	0.13	0.1	-0.02	0.09
Computing Major	C	21	0.5	0.53	0.03	0.21	0.55	0.68	0.13	0.04	0.48	0.49	0.01	0.83	0.14	0.1	-0.03	0.15
	E	75	0.49	0.49	0.01	0.71	0.57	0.6	0.03	0.33	0.47	0.47	0	0.96	0.14	0.11	-0.02	0.14
Non-Computing Major	C	8	0.38	0.49	0.11	0.01	0.47	0.53	0.06	0.52	0.35	0.48	0.13	0.07	0.14	0.15	0.01	0.52
	E	17	0.51	0.52	0.01	0.71	0.56	0.56	0	1	0.5	0.51	0.01	0.73	0.12	0.1	-0.02	0.34

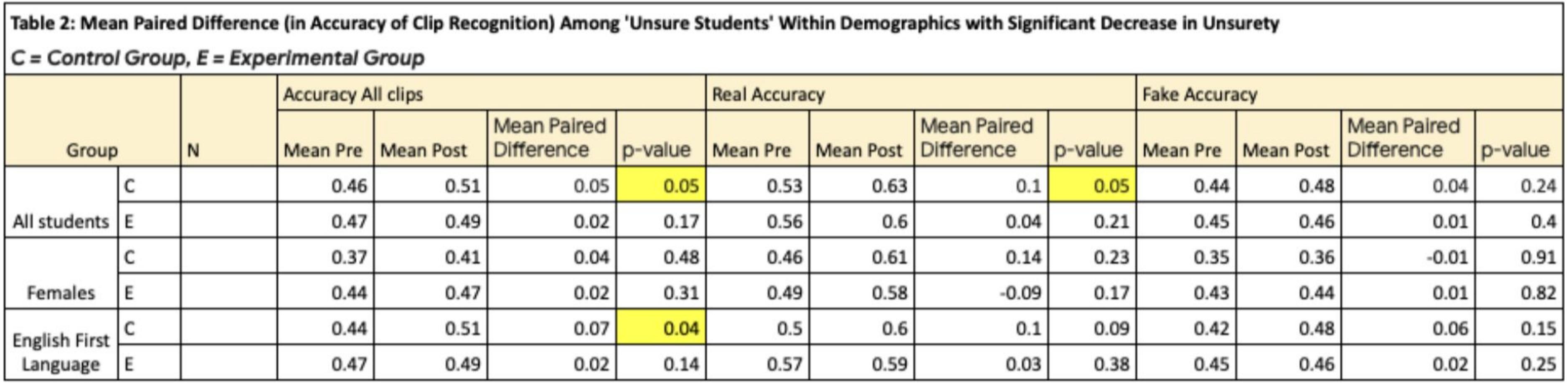


Table 2: Mean Paired Difference (in Accuracy of Clip Recognition) Among 'Unsure Students' Within Demographics with Significant Decrease in Unsurety

C = Control Group, E = Experimental Group

Group		N	Accuracy All clips				Real Accuracy				Fake Accuracy			
			Mean Pre	Mean Post	Mean Paired Difference	p-value	Mean Pre	Mean Post	Mean Paired Difference	p-value	Mean Pre	Mean Post	Mean Paired Difference	p-value
All students	C		0.46	0.51	0.05	0.05	0.53	0.63	0.1	0.05	0.44	0.48	0.04	0.24
	E		0.47	0.49	0.02	0.17	0.56	0.6	0.04	0.21	0.45	0.46	0.01	0.4
Females	C		0.37	0.41	0.04	0.48	0.46	0.61	0.14	0.23	0.35	0.36	-0.01	0.91
	E		0.44	0.47	0.02	0.31	0.49	0.58	-0.09	0.17	0.43	0.44	0.01	0.82
English First Language	C		0.44	0.51	0.07	0.04	0.5	0.6	0.1	0.09	0.42	0.48	0.06	0.15
	E		0.47	0.49	0.02	0.14	0.57	0.59	0.03	0.38	0.45	0.46	0.02	0.25