IATs, and WEATs, and WEFATs, oh my!

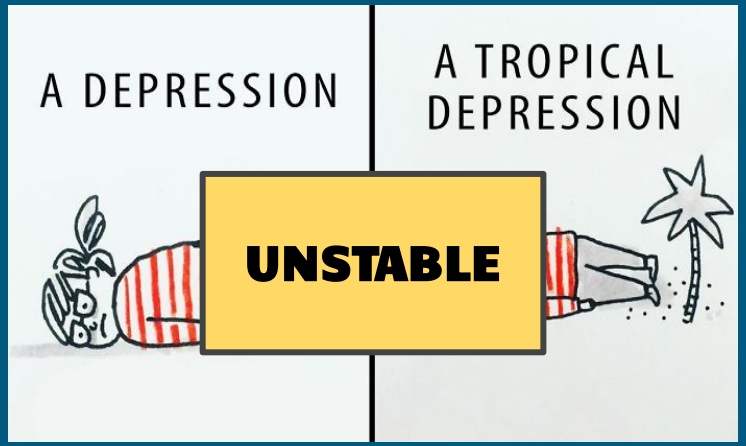
Karie Moorman, CIS UC Merced
Paul Smaldino, CIS UC Merced
2nd Year Presentation, 4 May 2018

outline:

1. implicit bias and IATs (Greenwald et al, 1998)
2. word vectors, WEAT and WEFAT (Caliskan et al, 2017)
3. application of these measures

1. We remember IATs, right?

(Greenwald et al., 1998; Nosek et al., 2002)



How are these biases represented in “big data”?

2. WEAT's a WEFAT? (Caliskan et al., 2017)

"Machines learn what people know implicitly."

The effect size is

$$\frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std-dev}_{w \in X \cup Y} s(w, A, B)}$$

WEAT permutation test statistic: probability that random permutation >= the difference of the sample means

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

WEFAT permutation test statistic: "predict the property given the vector"

$$s(w, A, B) = \frac{\text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})}{\text{std-dev}_{x \in A \cup B} \cos(\vec{w}, \vec{x})}$$

Pretrained model: glove word vectors using Common Crawl corpus (840 billion tokens) (glove.840B.300dcasedCommoncrawl.txt)

3. In what ways are speakers of English biased against "immigrants"?

Previous work using IATs:

Vanderbilt: Efrén Pérez (2010)
White vs. Latino immigrant surnames & Good vs. Bad words

Rice: James Hedrick & Aleks Ksiazkiewicz (2012)
White vs. Latino immigrant surnames & Positive vs. Negative words
White vs. Latino immigrant surnames & High vs. Low-skilled jobs

Yale: Lydia Keating (2017)
Immigrant vs. Non-immigrant words & Positive vs. Negative words

Replicating IAT effects using WEFAT

Studies	Lexemes	IAT (Cohen's d)	WEFAT (Cohen's d)
Vanderbilt (*N*=44; 337) (10x10) Surnames & Good/Bad words	honest, joy, love, peace, wonderful, honor, pleasure, glorious, laughter, happy agony, prison, terrible, horrible, nasty, evil, awful, failure, hurt, poverty	**1.65**	**1.58**
Rice (*N*=49) (4x4) Surnames & +/- words High/Low-skilled jobs	wonderful, pleasure, glorious, happy terrible, horrible, nasty, awful doctor, engineer, professor, scientist laborer, busboy, janitor, maid	**0.69** **-0.13**	**1.61** **1.78**
Yale (*N*= 67) (6x7) Immigrant/Non-Immigrant & +/- words	lovely, pleasure, glorious, beautiful, marvelous, wonderful, joyful humiliate, terrible, painful, nasty, horrible, agony, tragic	**-0.29**	**-1.19**

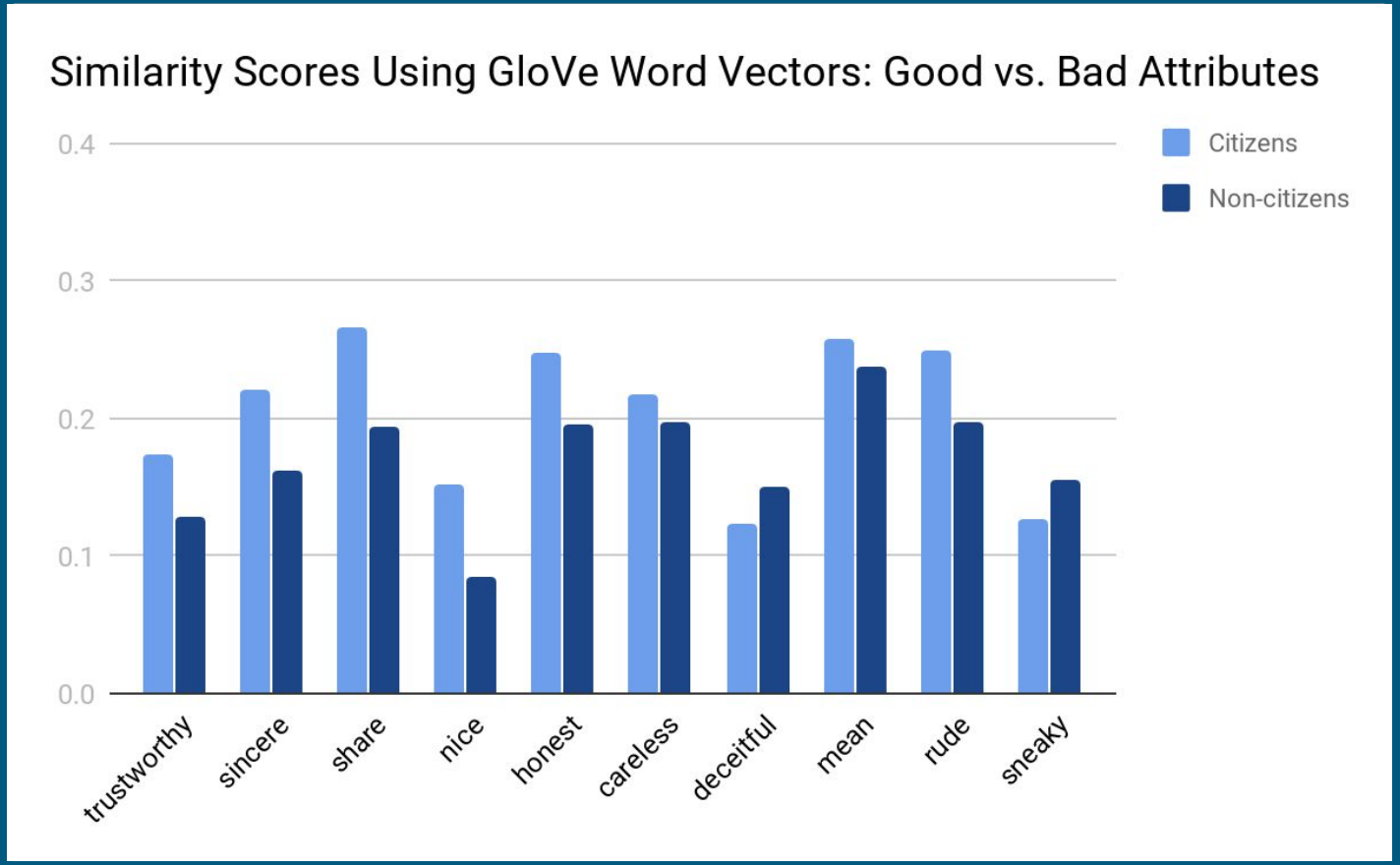
Other interesting word sets we can compare:

Citizens: citizens, locals, residents, natives, inhabitants

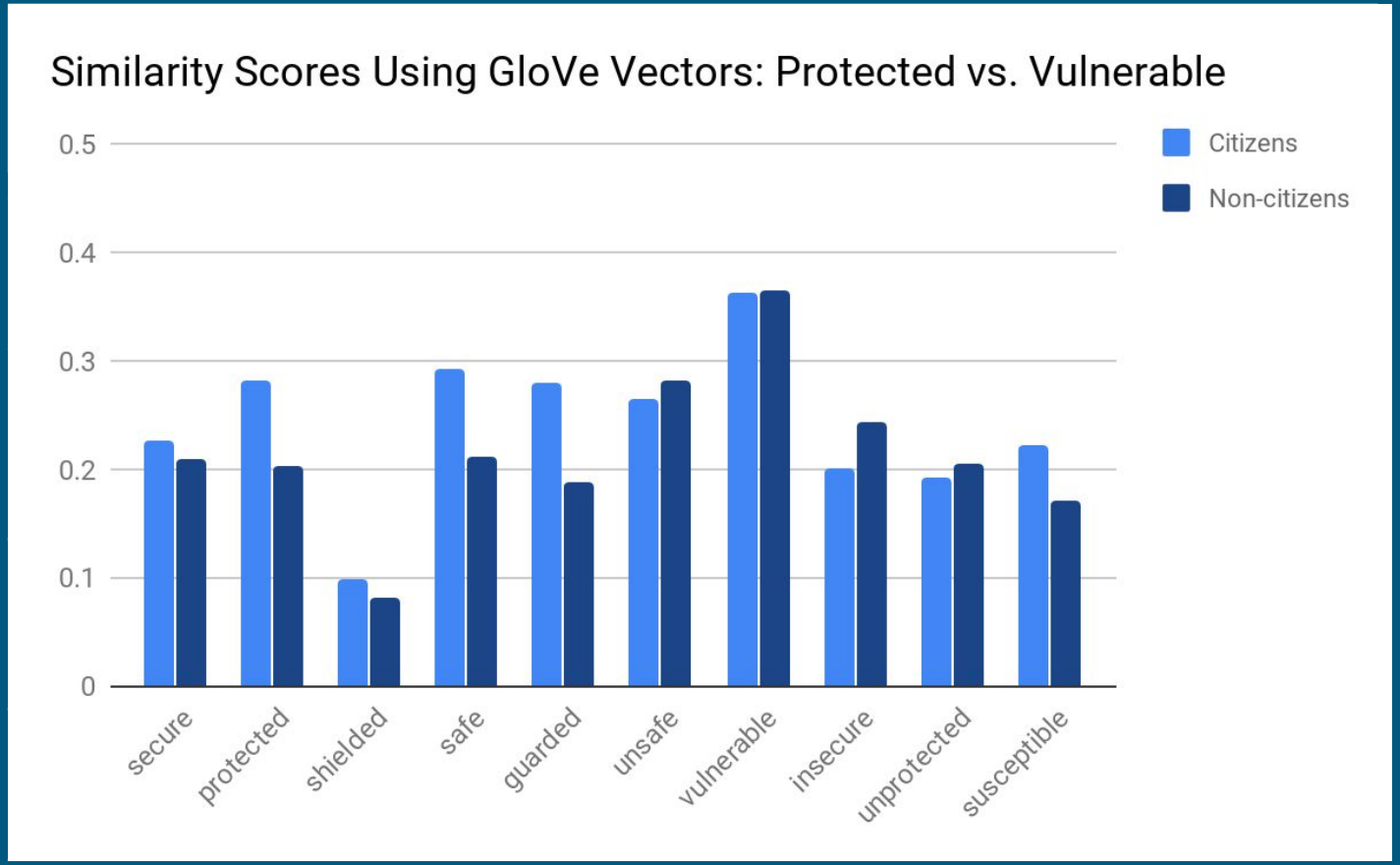
Non-citizens: immigrants, illegals, foreigners, undocumented, refugees

	Citizens/ Non-citizens	Lexemes
Positive/ Negative Attributes	***d*=0.82**	sincere, honest, nice, trustworthy, reliable sneaky, deceitful, mean, rude, careless
Protected/ Vulnerable words	***d*=1.50**	safe, guarded, secure, protected, shielded unsafe, susceptible, vulnerable, insecure, unprotected
Metaphors	***d*=1.09**	come, enter, move, walk, go steal, cheat, hide, harm, burden

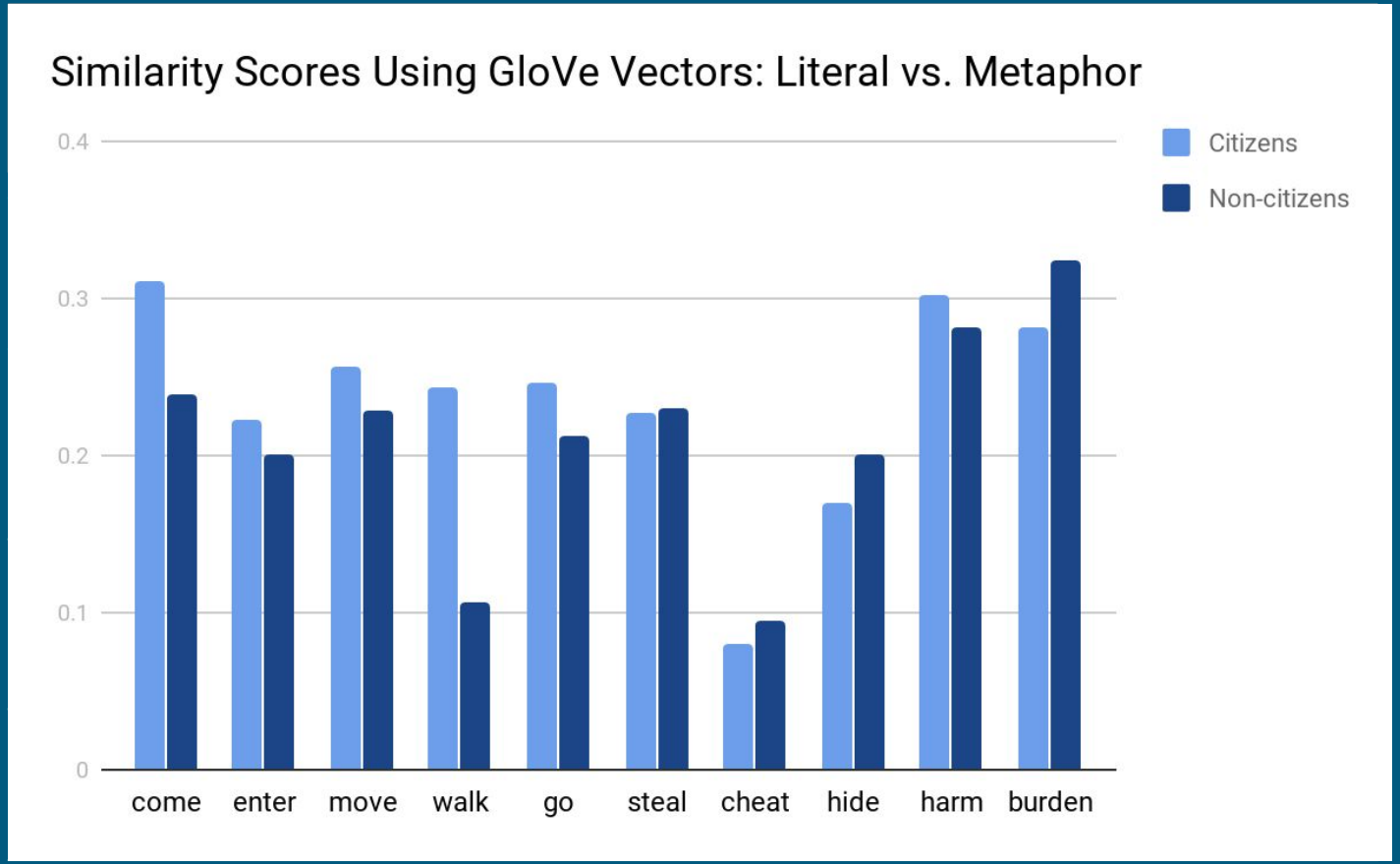
Other interesting word sets we can compare:



Other interesting word sets we can compare:



Other interesting word sets we can compare:



“immigrants” vs. “refugees” by Countries

Mexico
Venezuela
Guatemala
Honduras
Cuba
Haiti

Syria
Turkey
Greece
Afghanistan
Yemen
Iran

Cohen’s $d = 1.41$

IAT (pilot): Citizen vs. Non-citizen words and Good vs. Bad Attribute words

citizens, locals, residents, natives, inhabitants

immigrants, illegals, undocumented, refugees, foreigners

sincere, honest, nice, trustworthy, reliable

sneaky, deceitful, mean, rude, careless

- Is implicit bias present?
- Is participant bias really "implicit"? (Nosek et al., 2002; Howell, 2017)
- How do these results compare to "big data" findings?
- Does social/ethnic group affiliation influence bias? (Pérez, 2010)
- Does political affiliation influence bias?

IAT: Citizens vs. Non-citizens and Good vs. Bad Attributes

N ≈ 400-500 (60-70 participants per group)

(i.e., race/ethnicity, gender, self-identification as “immigrant” or “refugee”)

Procedure:

(1) IAT (5x5)

(2) Explicit preference for “non-citizens” (5pt. Likert scale: Extremely good/bad) (Nosek et al., 2002)

(3) Expectation of feedback (Howell, 2017)

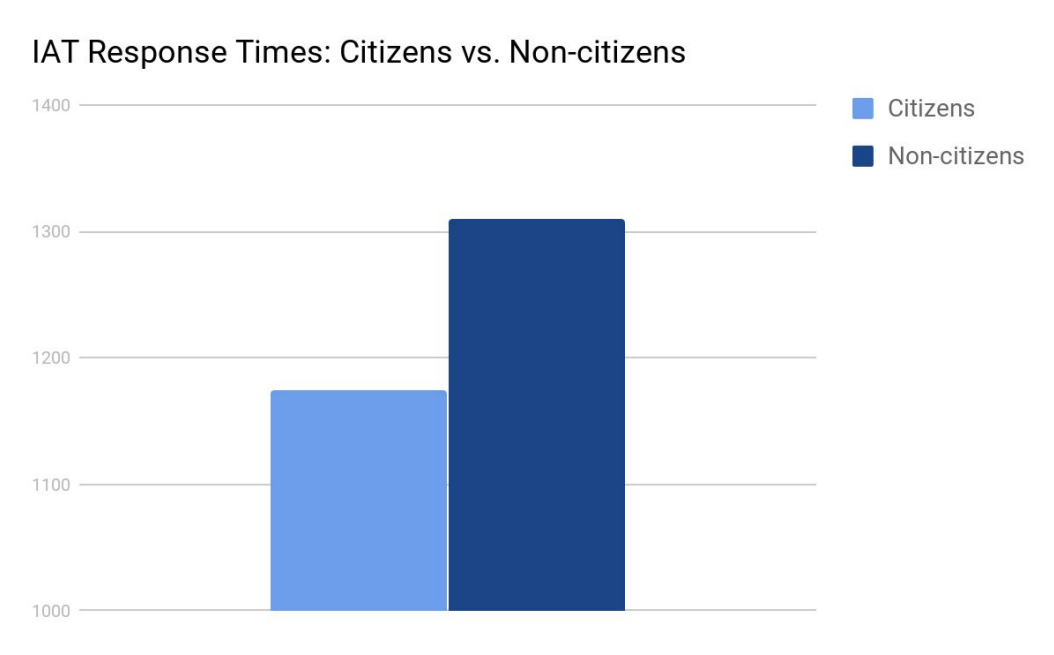
(7pt. Likert scale: Strong Preference for Non-citizens/Citizens)

(4) Self-identification as “immigrant” or “refugee” (Y/N)

(5) Political Affiliation Questions (Wilson-Patterson Conservatism Scale)

(6) Demographics Questions

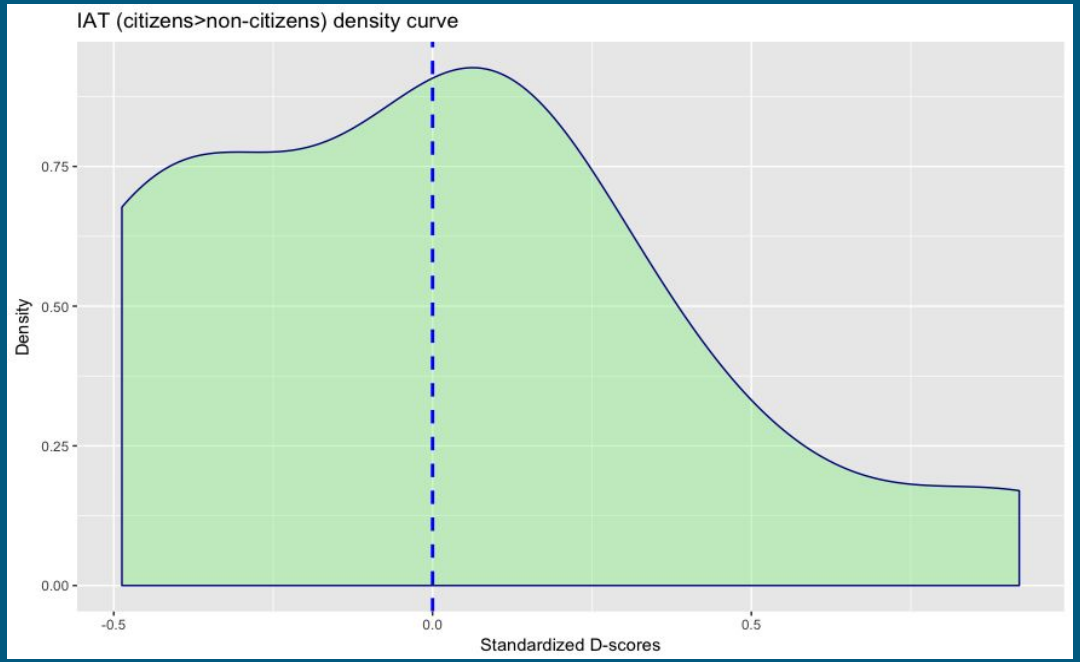
IAT pilot results: Citizens vs. Non-citizens and Good vs. Bad Attributes



(*N*=13)

(Mean D-score = 0.29, t(12)=2.607, *p*=0.02, Cohen's *d*=0.72, α=0.89)

IAT pilot results: Citizens vs. Non-citizens and Good vs. Bad Attributes



(*N*=13)

(Mean D-score = 0.29, t(12)=2.607, *p*=0.02, Cohen's *d*=0.72, α=0.89)

Future Work

Finalize experimental design and complete pre-registration.

Rewrite WEFAT in python (and/or R).

Future-Future Work

Investigate bias in other social groups

(e.g., gender/sexuality, political affiliation, diet, mode of transportation).

How does bias vary across corpora?

- Pre-trained models (e.g., Google News, WordNet)
- News and Forums (e.g., Reddit, Latino/Asian news outlets)

Directionality of Metaphor using RTs from IATs?

References

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

Carpenter, T., Pogacar, R., Pullig, C., Kouril, M., LaBouff, J.,… Chakroff, A. (2018, April 3). Conducting IAT Research within Online Surveys: A Procedure, Validation, and Open Source Tool. http://doi.org/10.17605/OSF.IO/6XDYJ

Hedrick, J., & Ksiazkiewicz, A. (2012). Implicit Attitudes toward Highly Skilled and Low-skilled Immigration.

Howell, J. L., Redford, L., Pogge, G., & Ratliff, K. A. (2017). Defensive Responding to IAT Feedback. *Social Cognition*, *35*(5), 520-562.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464.

Keating, L. (2017). *A Cross Cultural Analysis of Implicit and Explicit Xenophobia* (Doctoral dissertation, Yale University).

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101.

Pérez, E. O. (2010). Explicit evidence on the import of implicit attitudes: The IAT and immigration policy judgments. *Political Behavior*, *32*(4), 517-545.

thanks.

Paul Smaldino, Colin Holbrook,
Alexia Galati, Jordan Ackerman,
Kyle Hamilton