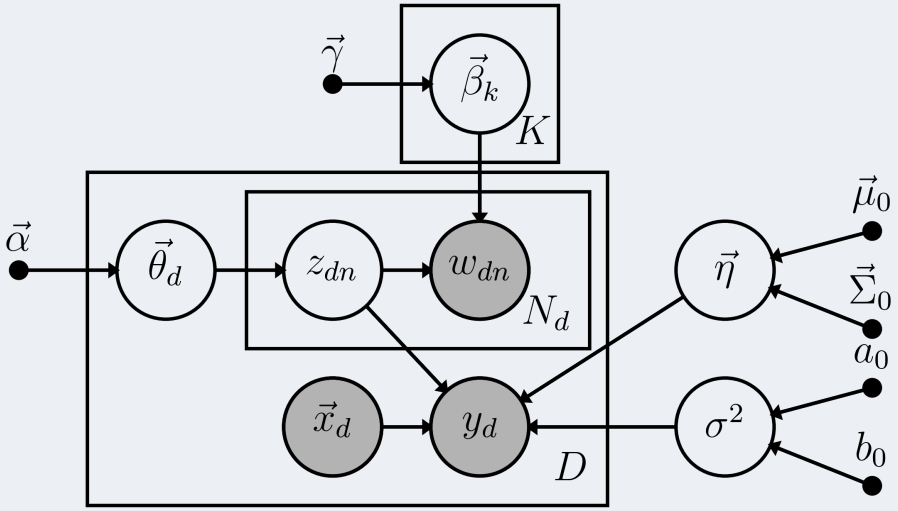
Combining Topic Modeling and Regression

Supervised Topic Modeling with Covariates

Kenneth Tyler Wilcox, Ross Jacobucci, & Zhiyong Zhang

Department of Psychology, University of Notre Dame

IMPS 2020 Spotlight Talk



Text Data in Psychology

- Text is an increasingly popular data source
    - Social media (Schwartz et al., 2013)
    - Free responses (Popping, 2015)
    - Medical health records (Obeid et al., 2019)
- New text mining algorithms are growing in popularity (Finch et al., 2018; Iliev et al., 2015; Kjell et al., 2019; Rohrer et al., 2017)
- Current challenge is to adapt these algorithms to psychological research

Motivating Example

- How do we combine open response items with other measures to study clinical outcomes?
- What can we learn from text that we miss with current scales?
- 827 adults recruited on MTurk
    - Outcome: Beck Hopelessness Scale (Beck et al., 1989)
    - Open response item: "What are your expectations for the future?"
    - Depression Anxiety Stress Scales (Lovibond et al., 1995)
    - Age
- **How do we incorporate the open responses?**

Two Streams

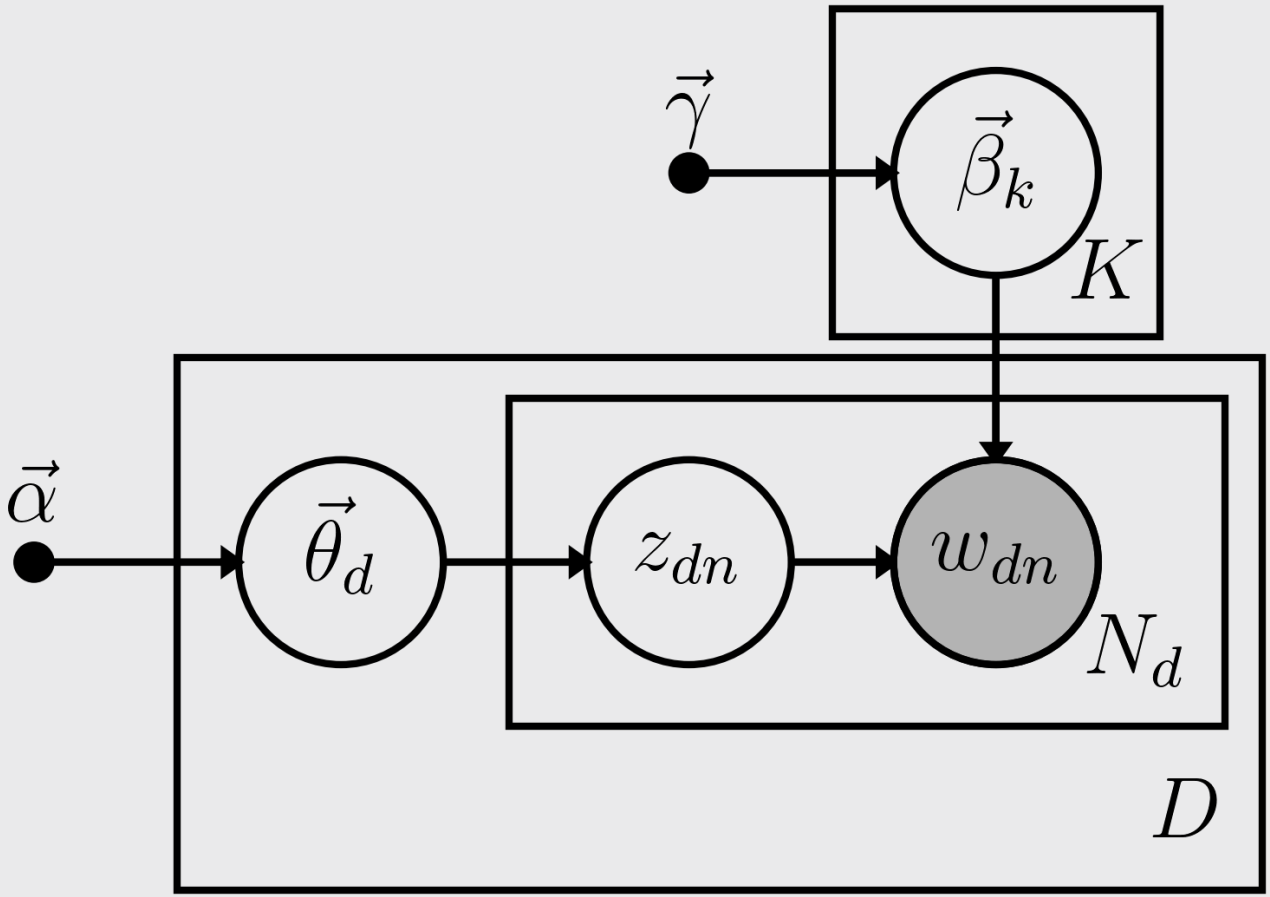
Top Down

- Dictionary methods
  - e.g., LIWC (Tausczik et al., 2010)
  - Define "constructs"
  - Fast, cheap
  - Popular in psychology
  - Dictionaries may not be valid for given data

Bottom Up

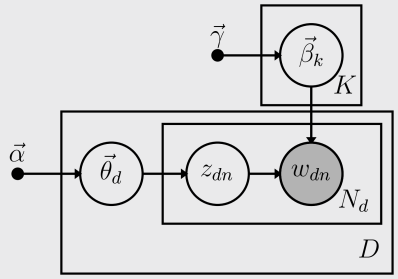
- Qualitative analysis
  - "Gold standard"
  - Time-consuming, expensive
  - Hard to reuse
- Quantitative models
  - e.g., LSI, **topic models**, deep learning
  - Data-driven, fast, cheap
  - Popular outside psychology
  - Reusable

Topic Modeling



Latent Dirichlet Allocation (LDA)

Topic model: probability distributions on words (Blei et al., 2003)



Topics:

$$\vec{\beta}_k = \Pr\left[w_{dn} = m | z_{dn} = k\right]$$
$$\vec{\beta}_k \sim \mathrm{Dir}\left(\vec{\gamma}\right)$$

Topic proportions:

$$\vec{\theta}_d = \Pr\left[z_{dn} = k\right] \sim \mathrm{Dir}\left(\vec{\alpha}\right)$$

$$L(\vec{\Theta}, \vec{B}, \vec{Z}) = \prod_{d=1}^{D} \prod_{n=1}^{N_d} \theta_{d z_{dn}} \beta_{z_{dn} w_{dn}}$$

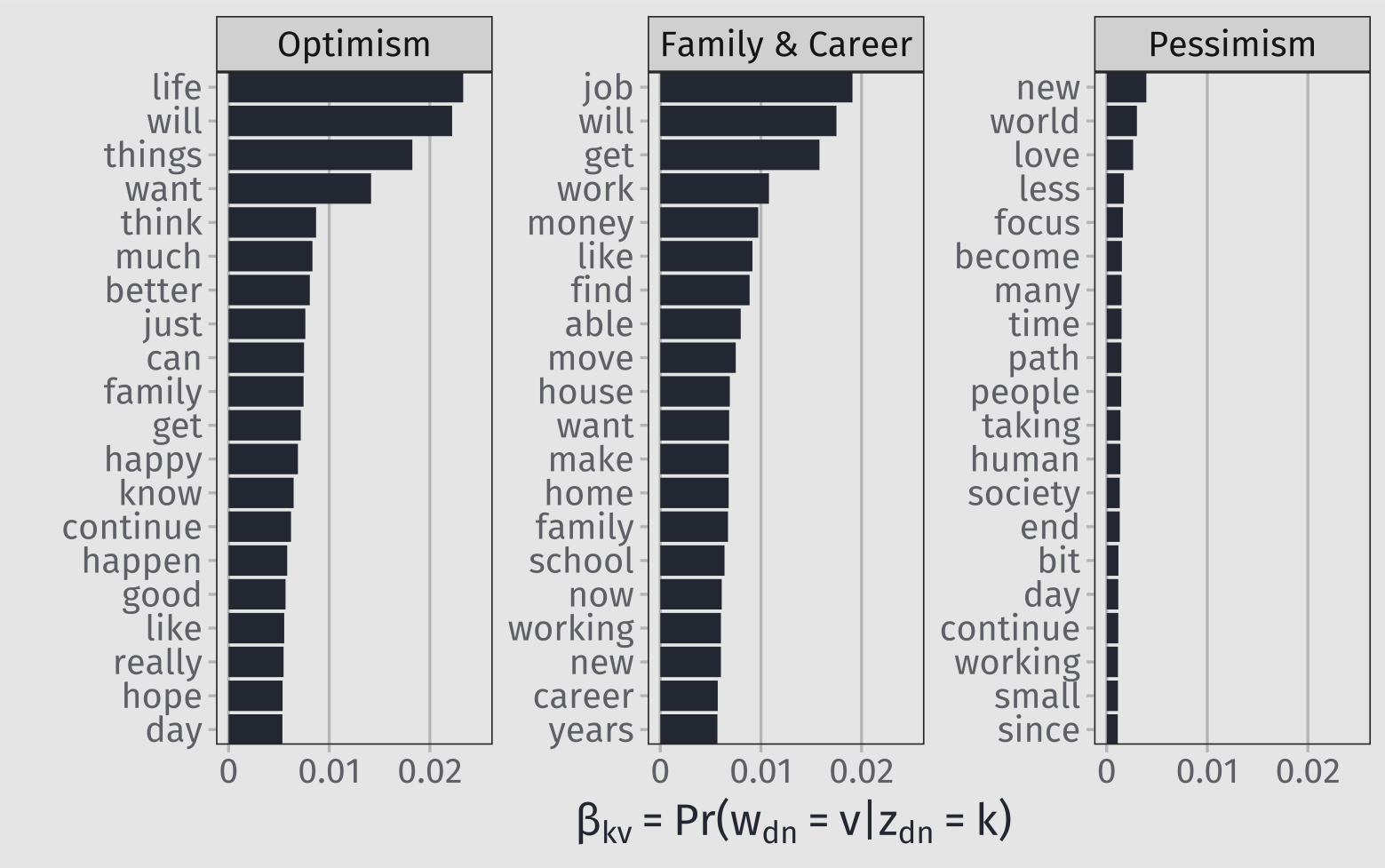
Topic assignments:

$$\left(z_{dn} | \vec{\theta}_d\right) \sim \mathrm{Cat}(\vec{\theta}_d)$$

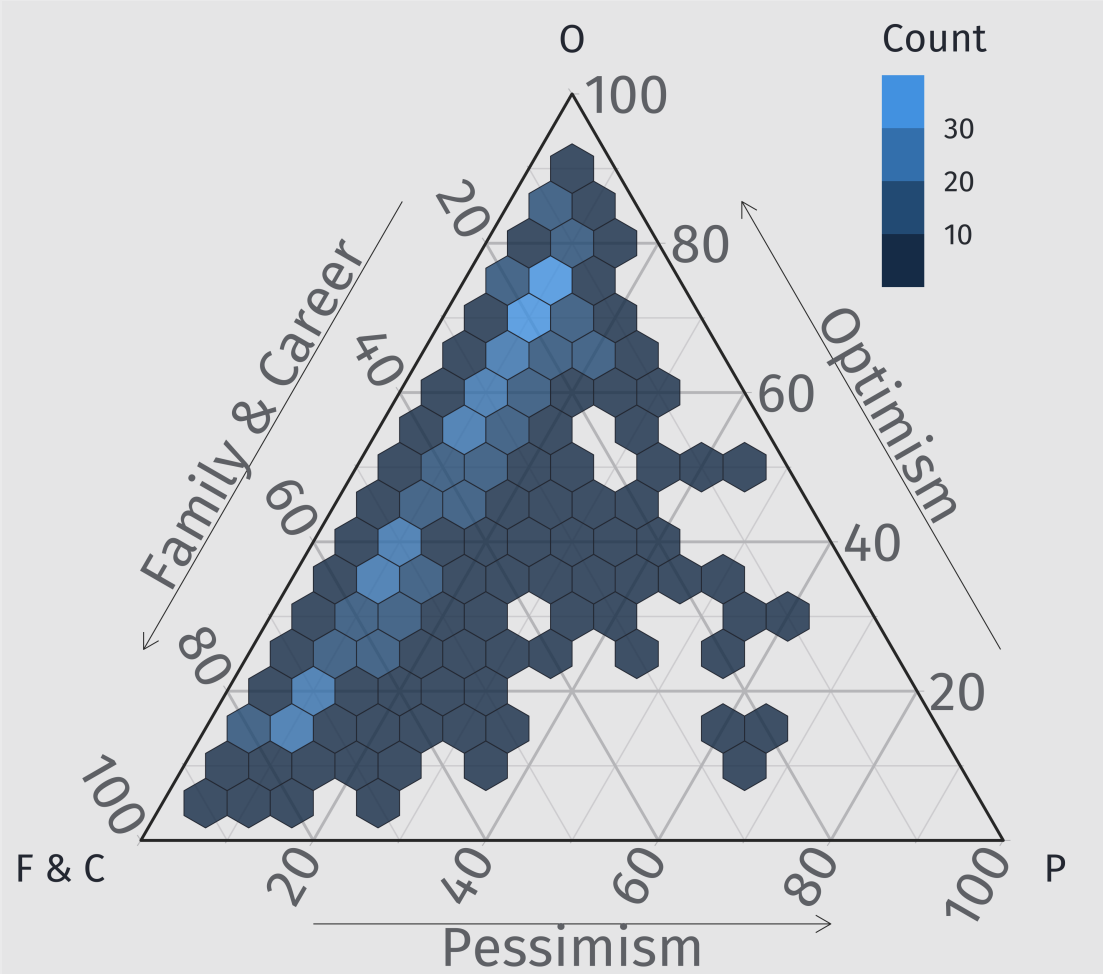
Words:

$$\left(w_{dn} | z_{dn} = k, \vec{\beta}_k\right) \sim \mathrm{Cat}(\vec{\beta}_k)$$

Example of Topics



Example of Topic Proportions



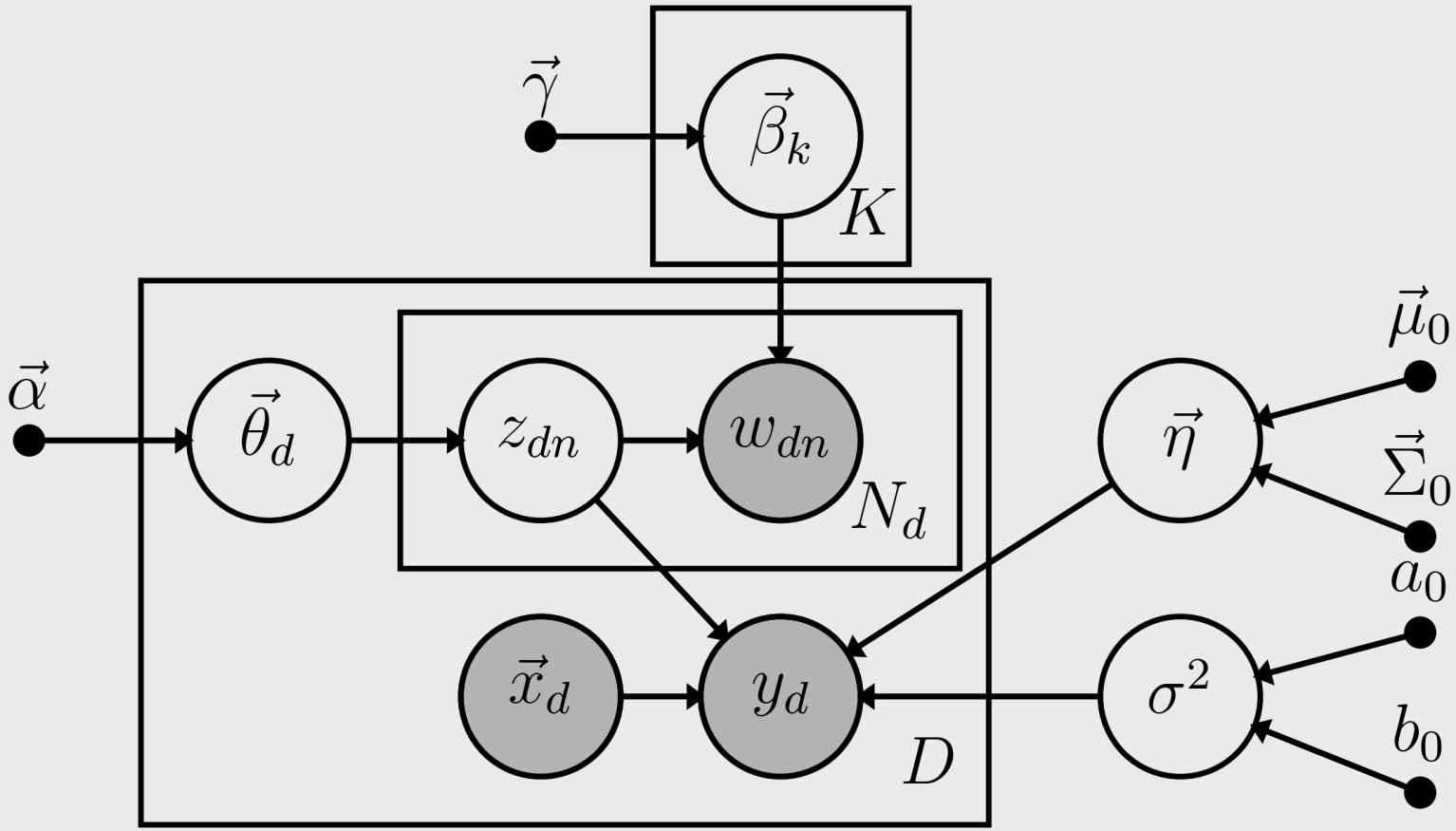
Incorporating Topic Modeling in Regression

Fusing Topic Models and Regression

$$Y = \boxed{\phantom{\vec{\Theta}}}\eta + X\beta + \epsilon$$

- Two-stage approach (Packard et al., 2020; Rohrer et al., 2017)
  - Use estimated $\vec{\Theta}$ to predict Y
  - Could include other manifest predictors $\vec{X}$
- One-stage approach
  - Supervised topic model (SLDA; Blei et al., 2010)
  - Does not include $\vec{X}$
- We propose the SLDAX model
  - One-stage approach
  - Allow topics and manifest predictors of Y

SLDAX


$\vec{\gamma}$
$\vec{\beta}_k$
K
$\vec{\alpha}$
$\vec{\theta}_d$
z_{dn}
w_{dn}
N_d
$\vec{\mu}_0$
$\vec{\eta}$
$\vec{\Sigma}_0$
a_0
$\vec{x}_d$
y_d
σ^2
b_0
D

SLDAX

$$\mathbb{E}\left[Y_d|\vec{X}_d, \vec{\bar{Z}}_d\right] = \sum_{k=1}^{K} \eta_k \bar{Z}_{dk} + \sum_{j=1}^{p} \eta_j X_{dj}$$

- $\bar{z}_{dk} = N_d^{-1} \sum_{n=1}^{N_d} I(z_{dn} = k)$
- Can use generalized linear model framework to extend to non-normal outcomes
- We derived a collapsed Gibbs sampling algorithm for Bayesian estimation
  1. $(Y_d|\cdot) \sim \mathbf{N}(\cdot)$
  2. $(Y_d|\cdot) \sim \mathbf{Ber}(\cdot)$
- As in any mixture model, need to handle label switching (Stephens, 2000)

Inference for Topic Effects

- Because $\vec{\tilde{z}}_d$ is ipsative, inference changes
- η_k represents the conditional mean of Y for topic k alone
- To test the effect of a topic, we test a contrast (Park, 1978; Snee et al., 1976)

$$c_k = \eta_k - \frac{\sum_{k' \neq k}^{K} \eta_{k'}}{K-1} \stackrel{?}{=} 0$$

- We can sample c_k directly from the posterior
- Many applications have incorrectly compared η_k to 0 (Packard et al., 2020; Rohrer et al., 2017; Schwartz et al., 2013)
- Similarly, interpreting the sign of η_k is misleading
  - Interpret the sign of c_k instead

Software

- **psychtm** R package in *early* development
- Features
  - Bayesian estimation of LDA, SLDA, SLDAX in **C++**
    - Normal and dichotomous outcomes supported
  - Visualization of $\vec{\Theta}$ and $\vec{B}$
  - Perform model comparison via WAIC (Watanabe, 2010)
- Available from Github

```
devtools::install_github("ktw5691/psychtm")
fit ← gibbs_sldax(y ~ x1 + x2, data = xy, docs = docs, V = V, K = 2)
```

Do We Need Another Model?

Simulation Study

Goal

- Compare SLDAX with two-stage approach (LDA + OLS regression)
  - SLDAX from our R package **`psychtm`**
  - LDA model from R package **`topicmodels`**
- Conditions
  - # topics K: 2 and 5
  - # documents D: 200, 800, and 1500
  - Mean # words $\bar{N}_d$: 15, 80, and 150
  - Vocabulary V: 500 and 1000
- 100 replications

Simulation Study

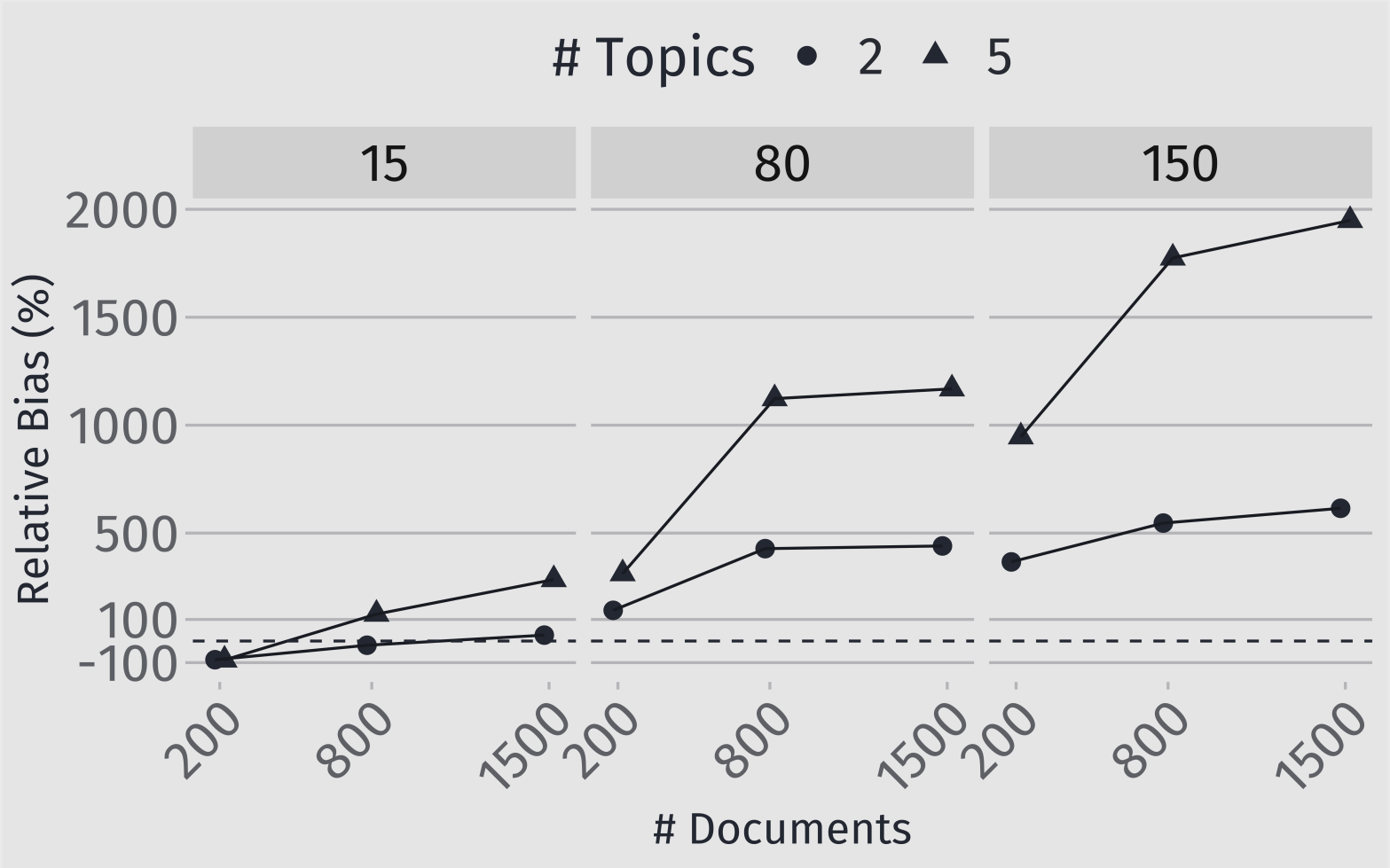
Data Generation

- SLDAX model
  - $X \sim \mathrm{N}(0, 1)$ w/ R^2 = .15
  - $Y \sim \mathrm{N}(\cdot)$
  - K topics w/ joint R^2 = .35

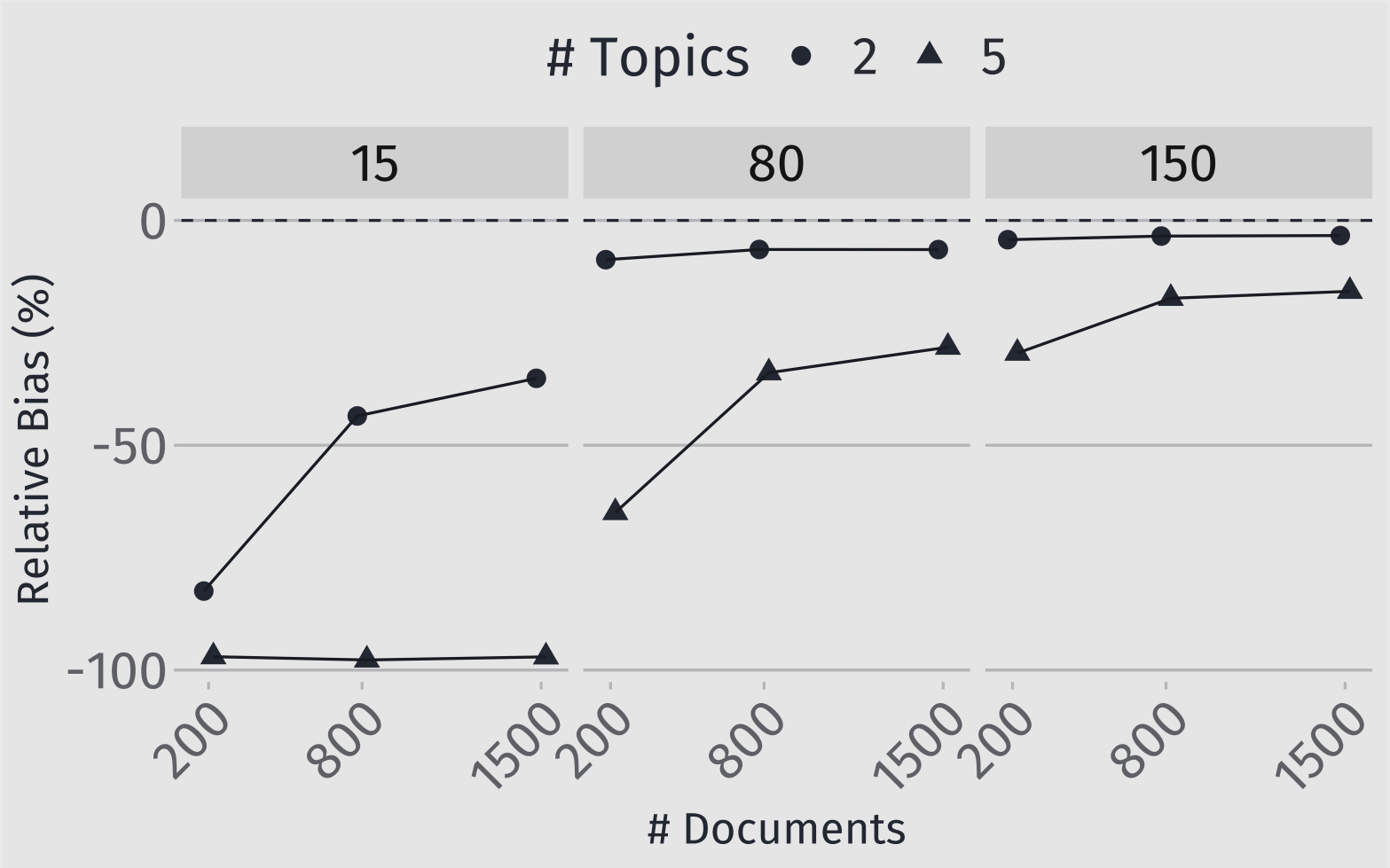
Estimation

- SLDAX with flat priors
- Two-stage
  1. LDA: estimated w/ variational EM (same hyper-parameters)
  2. OLS regression

Two-Stage Estimation Bias for $\eta_{\bar{z}}$



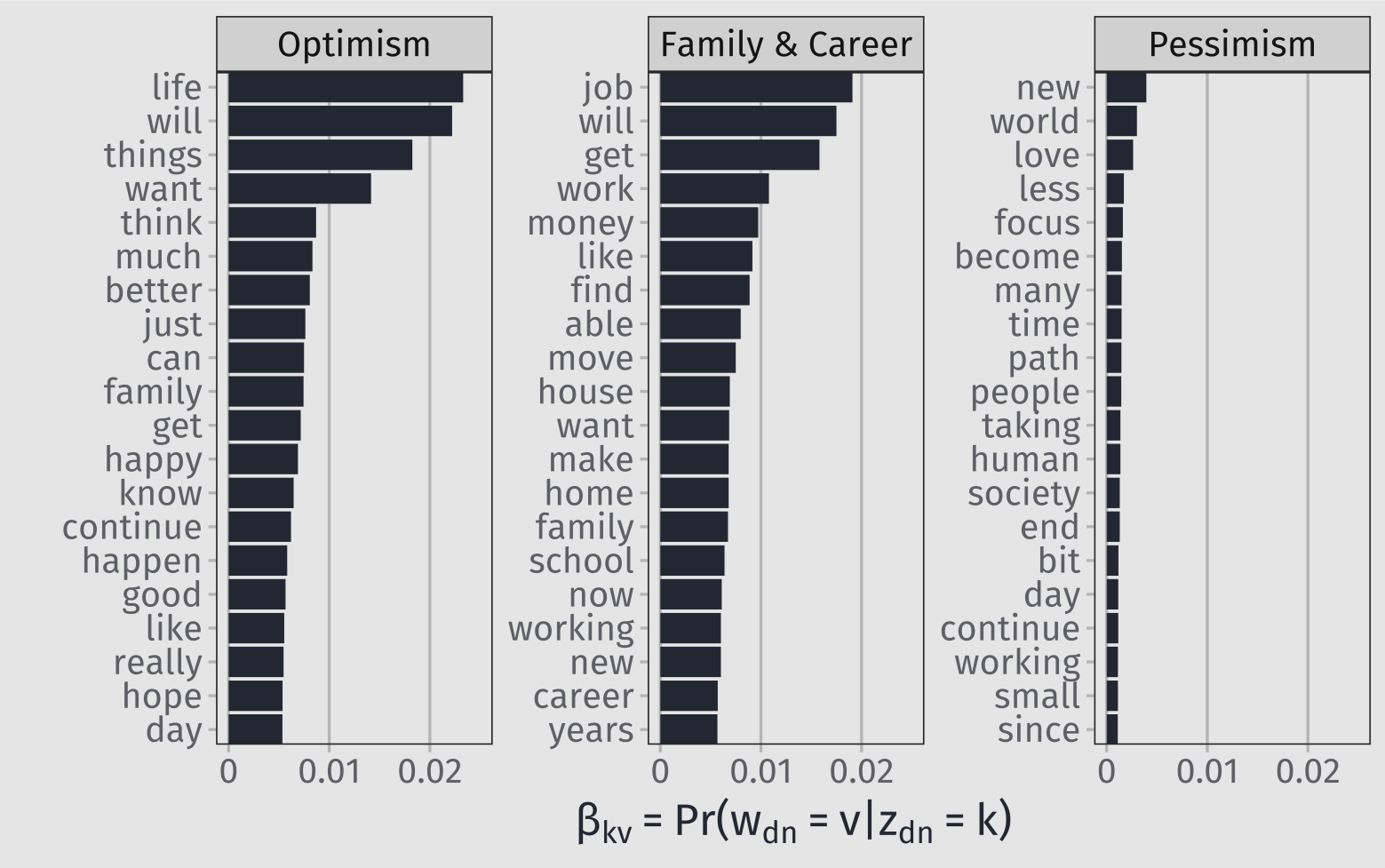
SLDAX Estimation Bias for $\eta_{\bar{z}}$



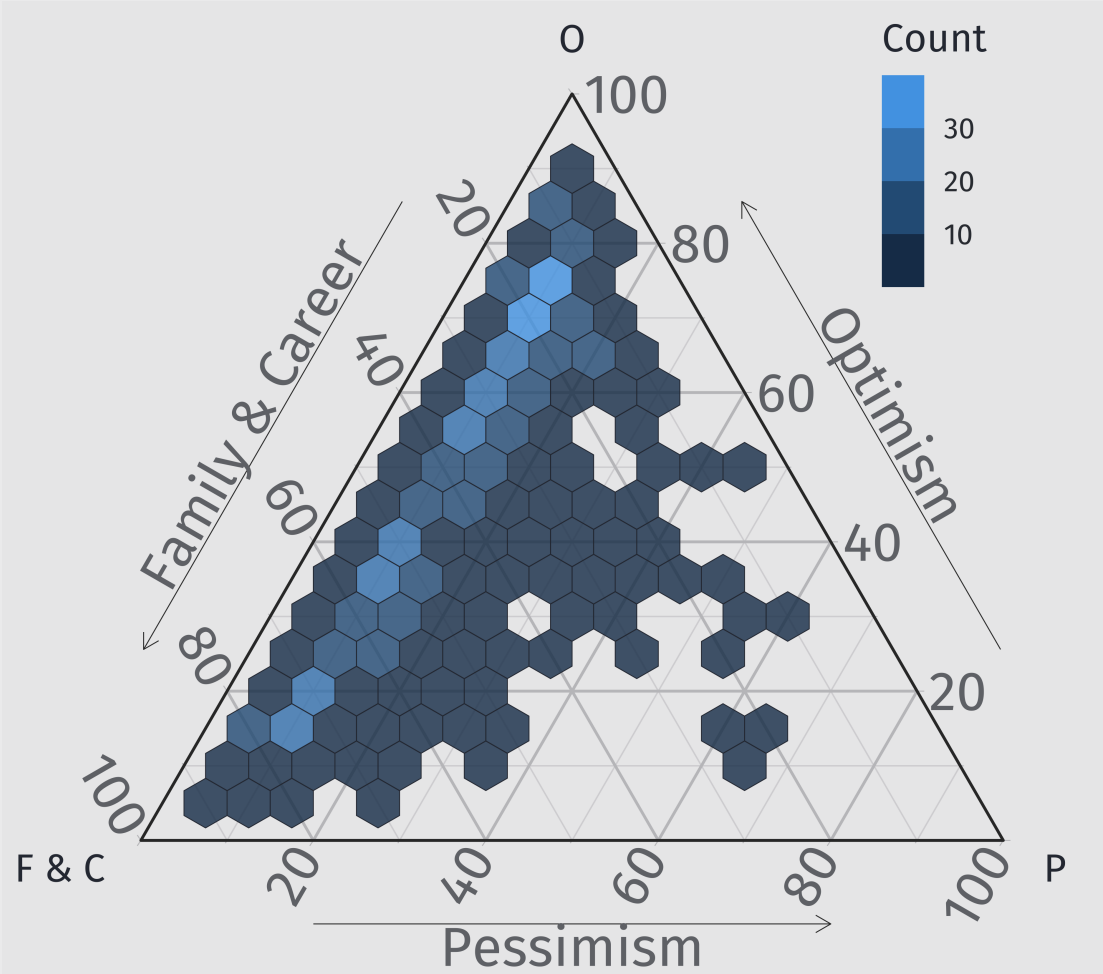
Motivating Example Revisited

- 827 adults
- Outcome: Hopelessness — BHS
- Predictors
  - "What are your expectations for the future?"
  - *M* = 50 words, *SD* = 24, *Range* = 5 – 186
  - After stopword removal:
    - Median = 20 words (*M* = 22, *SD* = 10, *Range* = 3 – 80)
    - Vocabulary of 2,636 words
  - DASS
  - Age (*M* = 33, *SD* = 10, *Range* = 18 – 79)

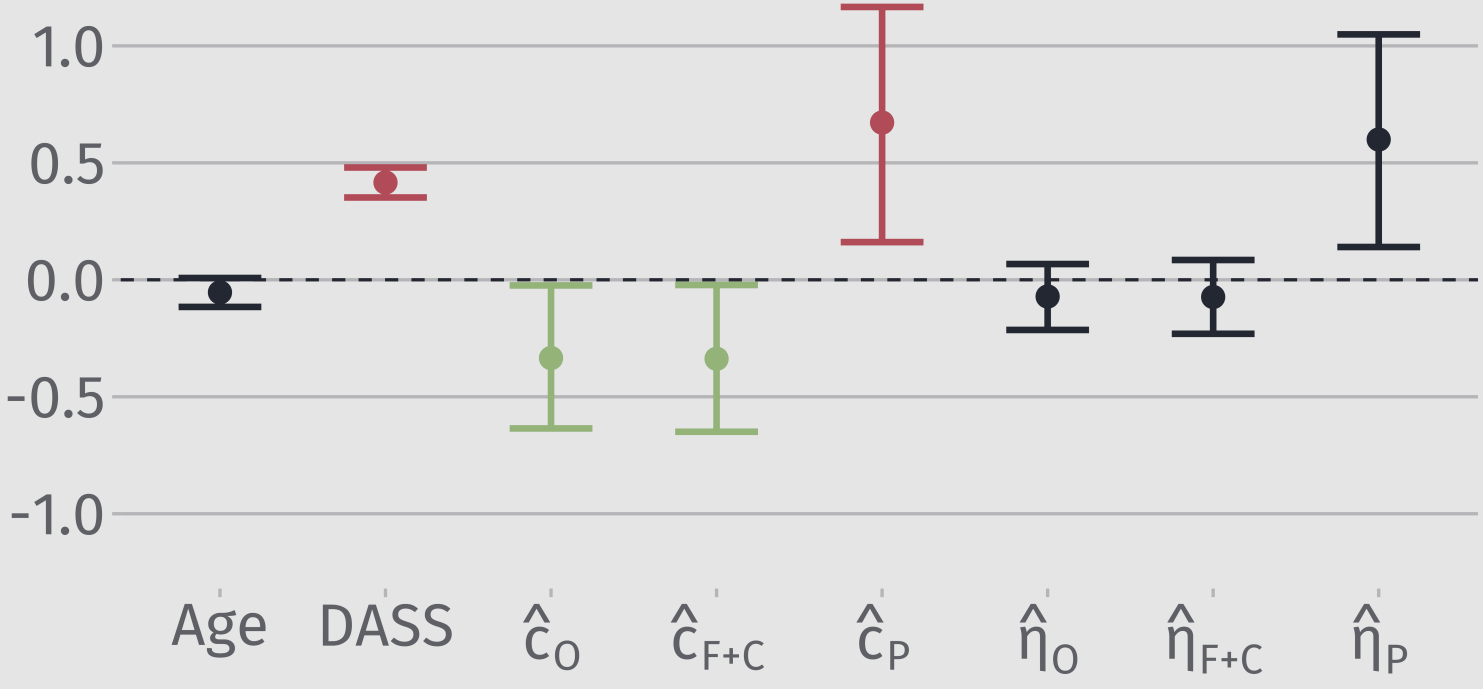
Estimated Topics



Estimated Topic Proportions



Posterior Regression Coefficients



$$c_k = \eta_k - K^{-1} \sum_{(-k)}^{K} \eta_j$$

Conclusions

- Themes in free responses associated with higher & lower hopelessness
    - Convergent validity for topics
    - Text topics associated with BHS above and beyond DASS
    - What are we not measuring?
- Topic effect estimates likely attenuated based on simulation results
    - Large D, small $\bar{N}_d$
- Could predict on new data or update model using new data

Discussion

Key Findings

- We derived MCMC algorithms to estimate SLDAX models
- SLDAX models implemented in open-source **R** package
- The popular two-stage approach yields (severely) biased regression estimates
- SLDAX yields accurate estimates with conservative shrinkage in short-document scenarios

Future Work

- SLDAX framework can be generalized
- Impact of text data quality on performance
- Prior specification with short documents

Thanks!

kwilcox3@nd.edu

ktylerwilcox.netlify.app

@ktw5691

Slides:

https://ktylerwilcox.netlify.app/talk/2020-imps-sldax/

References

Beck AT, Brown G, Steer RA (1989). "Prediction of Eventual Suicide in Psychiatric Inpatients by Clinical Ratings of Hopelessness." *Journal of Consulting and Clinical Psychology*, *57*(2), 309-310. https://doi.org/10.1037/0022-006X.57.2.309.

Blei DM, McAuliffe JD (2010). "Supervised Topic Models." *arXiv*.

Blei DM, Ng AY, Jordan MI (2003). "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, *3*, 993-1022.

Finch WH, Finch MEH, McIntosh CE, Braun C (2018). "The Use of Topic Modeling with Latent Dirichlet Analysis with Open-Ended Survey Items." *Translational Issues in Psychological Science*, 4(4), 403-424. https://doi.org/10.1037/tps0000173.

Iliev R, Dehghani M, Sagi E (2015). "Automated Text Analysis in Psychology: Methods, Applications, and Future Developments." *Language and Cognition*, *7*(2), 265-290. https://doi.org/10.1017/langcog.2014.30.

Kjell ONE, Kjell K, Garcia D, Sikström S (2019). "Semantic Measures: Using Natural Language Processing to Measure, Differentiate, and Describe Psychological Constructs." *Psychological Methods*, *24*(1), 92-115. https://doi.org/10.1037/met0000191.

Lovibond PF, Lovibond SH (1995). "The Structure of Negative Emotional States: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories." *Behaviour Research and Therapy*, *33*(3), 335-343. https://doi.org/10.1016/0005-7967(94)00075-U.

Obeid JS, Weeda ER, Matuskowitz AJ, Gagnon K, Crawford T, Carr CM, Frey LJ (2019). "Automated Detection of Altered Mental Status in Emergency Department Clinical Notes: A Deep Learning Approach." *BMC Medical Informatics and Decision Making*, *19*(1), 164. https://doi.org/10.1186/s12911-019-0894-9.

Packard G, Berger J (2020). "Thinking of You: How Second-Person Pronouns Shape Cultural Success." *Psychological Science*. https://doi.org/10.1177/0956797620902380.

Park SH (1978). "Selecting Contrasts among Parameters in Scheffe's Mixture Models: Screening Components and Model Reduction." *Technometrics*, *20*(3), 273-279. https://doi.org/10.2307/1268136.

Popping R (2015). "Analyzing Open-Ended Questions by Means of Text Analysis Procedures." *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *128*(1), 23-39. https://doi.org/10.1177/0759106315597389.

Roberts ME, Stewart BM, Airoldi EM (2016). "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association*, *111*(515), 988-1003. https://doi.org/10.1080/01621459.2016.1141684.

Rohrer JM, Brümmer M, Schmukle SC, Goebel J, Wagner GG (2017). ""What Else Are You Worried about?" Integrating Textual Responses into Quantitative Social Science Research." *PLoS ONE*, *12*(7), e0182156. https://doi.org/10.1371/journal.pone.0182156.

Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, Shah A, Kosinski M, Stillwell D, Seligman MEP, Ungar LH (2013). "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach." *PloS ONE*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791.

Snee RD, Marquardt DW (1976). "Screening Concepts and Designs for Experiments with Mixtures." *Technometrics*, *18*(1), 19-29. https://doi.org/10.2307/1267912.

Stephens M (2000). "Dealing with Label Switching in Mixture Models." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *62*(4), 795-809.

Tausczik YR, Pennebaker JW (2010). "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology*, *29*(1), 24-54. https://doi.org/10.1177/0261927X09351676.