# Building Knowledge-Guided Lexica to Model Cultural Variation

**Anonymous ACL submission**

## Abstract

Cultural variation exists between regions (e.g., the United States vs. China), but also *within* regions (e.g., California vs. Texas, Los Angeles vs. San Francisco). Measuring this regional cultural variation can illuminate how and why people think and behave differently. Historically, it has been difficult to computationally model cultural variation due to a lack of training data and scalability constraints. In this work, we introduce a new research problem for the NLP community: *How do we measure variation in cultural constructs across regions using language?* We then provide a scalable solution: building knowledge-guided lexica to model cultural variation, encouraging future work at the intersection of NLP and cultural understanding.

## 1 Introduction

People think and behave differently around the world. This is partly due to *cultural variation*, or the differences among individuals that exist due to some form of social learning (Cohen, 2001). Having a computational method that utilizes language to measure cultural variation could help us better understand humans (Tsai et al., 2006; Oishi et al., 2009), build more culturally-aware NLP systems (Hovy and Yang, 2021), and advance interdisciplinary research in anthropology, cultural psychology, etc. However, due to a lack of data and scalability constraints, few such methods exist.

In this paper, we present *measuring regional variation in culture* as a problem of interest for the NLP community and build a knowledge-guided lexical model as a scalable solution. Specifically, we focus on measuring *individualism and collectivism*[1] across the United States (US) using geolocated Tweets.

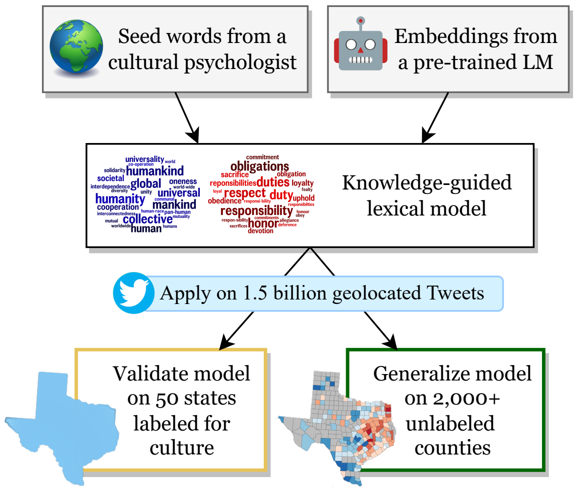


Figure 1: We build knowledge-guided lexica to model cultural variation using two types of domain knowledge: seed words based on cultural psychology theory and embeddings from a pre-trained language model.

Historically, measuring cultural dimensions across regions has been mostly done through questionnaires, such as the World Values Survey (WVS) (Haerpfer et al., 2020). However, questionnaires are time-consuming and heavily restricted in scope; the most recent WVS wave required 4 years and averaged 52 participants per US state. Recent work probes language models (LMs) for cultural values (Arora et al., 2023), but these LMs do not reflect all cultures equally (Havaldar et al., 2023).

The overhead of traditional survey-based approaches and inconsistent cultural awareness of existing LMs motivates scalable, computational methods that use *existing language data* to measure cultural variation instead. For the example problem addressed in this paper, we seek to measure individualism and collectivism across US counties using the following resources:

- Domain expertise from cultural psychologists.
- An open-source corpus (see Appendix A) of 1.5 billion geolocated Tweets from 6 million

[1] Cultural psychologists have quantified axes on which culture differs, also called *cultural dimensions*. A key cultural dimension that influences behaviors like voting, donating, etc. is *individualism vs. collectivism* (Hofstede, 2011). Collectivism stresses the importance of the community, while individualism focuses on each person's rights and concerns.

US users (Giorgi et al., 2018).
- Individualism and collectivism scores for fifty US states (Vandello and Cohen, 1999).

Pre-LM era solutions to measure culture use single words (Giorgi et al., 2020) or manually curated lexica (Graham et al., 2009), thus relying on a small number of highly specific words. A more modern NLP solution would take the form of either (1) training a model, or (2) prompting an LM. However, to classify 1.5 billion Tweets, (1) requires a sizable amount of labeled training data, and (2) is not computationally scalable. For instance, running this corpus through GPT-4 would cost roughly $900,000 (see Appendix B).

Additionally, building a Tweet-level deep learning model to predict culture is impractical. Most of an individual's language does not indicate their cultural beliefs; therefore, it is prohibitively expensive to label enough Tweets to train an adequate model.

Our method builds upon a line of work in NLP called lexicon induction (Araque et al., 2020; Buechel et al., 2020; Geng et al., 2022), which analyzes massive corpora in NLP without solely relying on deep learning. Past work mainly builds lexica for sentiment, emotion, etc. We uniquely focus on the domain where *little training data exists* and *not every utterance can be relevantly labeled.*

**Leveraging domain knowledge.** Our proposed method to model cultural variation utilizes both domain expertise from cultural psychology (via a set of expert-curated seed words) and knowledge implicit in LMs (via word embeddings) to build scalable lexical models.

We validate our method against past collectivism research at the US state-level (Pelham et al., 2022; Vandello and Cohen, 1999) and extend the analysis of individualism and collectivism across US counties, allowing for a more fine-grained spatial analysis (i.e., understanding how large areas like states are culturally heterogeneous).

We also show how county-level analyses of culture can obtain new insights into existing populations, via a taxonomy of *communities* (sociodemographic clusters of counties) from the American Communities Project (Chinni and Gimpel, 2011). This taxonomy has been previously utilized to understand differences in health behaviors and outcomes (Aggarwal et al., 2023; Guntuku et al., 2021), and we use it to better understand how communities vary in individualism and collectivism.

## 2 Building Knowledge-Guided Lexica

Lexica, or sets of curated words, are a highly scalable method for analyzing large datasets. However, building lexica linked to cultural theory from the ground up is also a time-intensive process.

To mitigate this, we propose a method that combines two types of domain knowledge to efficiently create lexica that can measure cultural variation. We first ask an expert psychologist to generate two small sets of seed words that capture individualism and collectivism respectively based on their knowledge of cultural psychology. We next leverage knowledge implicitly present in language models (e.g., word associations, word similarity, etc.) to expand these small sets of seed words into high-validity lexical models.

Using this method, we can measure regional variation for any cultural construct using language from those regions. Our approach has two components: Expansion and Purification. Figure 2 details this approach for our example problem – measuring individualism and collectivism across US counties.

**Step 1: Expansion** Given a set of seed words from a cultural psychologist (see Appendix for seed words), we utilize word embeddings[2] to expand the set of seed words in two ways: we locate all words that are similar to each individual seed word (*synonym expansion*), as well as locate the words that are similar to the overall construct described by the complete set of seed words (*concept expansion*).

For synonym expansion, we find the nearest neighbors for each individual seed word in embedding space and add these neighboring words to our lexica. For concept expansion, we average the embeddings of each seed word set (e.g. individualism) to find the *centroid embeddings*. We then find the nearest neighbors of each centroid embedding. By using embedding space to expand our lexica, we additionally calculate a weight for each expanded word, i.e., the cosine similarity between the expanded word and the corresponding seed word or centroid embedding. The weight for each seed word is 1.

This method is highly tunable – any embeddings can be used, and the number of nearest neighbors returned during expansion can be adjusted based on desired length of the final lexicon.

[2]We use FastText (Bojanowski et al., 2017) due to its fixed vocabulary size, efficient nearest neighbors functionality, and ability to find synonyms in context-free scenarios, but our methods are more general and agnostic to embedding type.

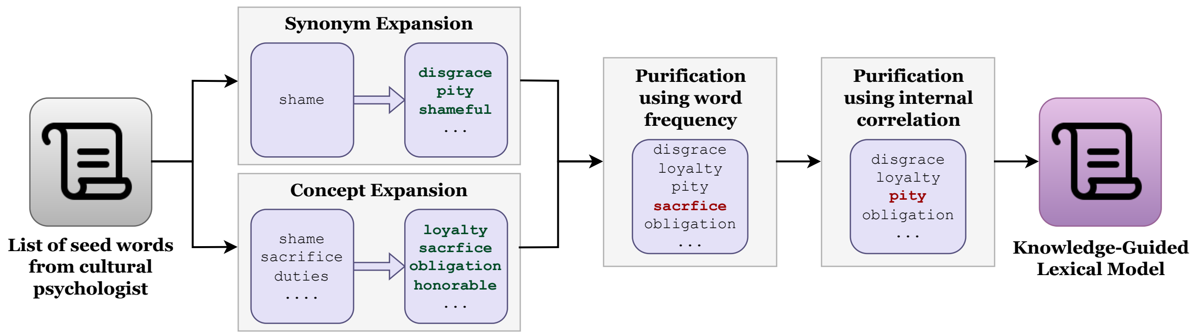


Figure 2: Our knowledge-guided lexica creation method. The first stage, *Expansion*, consists of synonym expansion and concept expansion, done in parallel. The second stage, *Purification*, includes frequency-based and correlation-based pruning, done sequentially.

**Step 2: Purification** Upon aggregating the words returned from both expansion types, we want to ensure that the resulting lexica are both pertinent and internally correlated.

To ensure pertinence, we filter out rare words, or any words below a given usage frequency (Bojanowski et al., 2017). Next, we ensure internal correlation. We apply our lexica to our US Twitter Corpus and compute the weighted frequencies for each word at two granularities: county-level and state-level. This produces scores that reflect the individualism and collectivism tendencies of every US county and state. We then check how each individual word's frequency correlates with the corresponding overall individualism/collectivism score. Any word that doesn't show a significant positive correlation (product-moment correlation coefficient $r < 0.15$) is removed from the lexica. This step ensures that every word contributes correctly to measuring the relevant cultural dimension.

Figure 5 visualizes our knowledge-guided individualism and collectivism lexica.

## 3 Validation

Upon expanding and purifying the lexica, we validate our results using the collectivism scores from Vandello and Cohen (1999) for each of the 50 US states. We see a significant positive correlation of our collectivism scores with Vandello & Cohen's collectivism scores (Table 1).

We also use relevant collectivism indicators from the Global Collectivism Index (Pelham et al., 2022) – religiosity, living arrangements (i.e. grandparent living in the household), and in-group bias. Using corresponding questions from the 2017 U.S. census and the 2017 wave of the World Values Survey (Haerpfer et al., 2020), we get data for all of these indicators at the US state level. We also see a significant positive correlation with all of these indicators (Table 1). Further details on validation are given in Appendix C.

| | **Individualism Lexicon Score** | **Collectivism Lexicon Score** |
|---|---|---|
| **Vandello & Cohen's Collectivism Scores** | -0.374 | 0.388 |
| **Living Arrangements** | -0.291 | 0.200 |
| **Religiosity** | -0.658 | 0.400 |
| **Ingroup Bias** | -0.513 | 0.464 |

Table 1: Pairwise product-moment correlations between our individualism and collectivism lexica applied to our Twitter Corpus and validation variables. We use Vandello & Cohen's Collectivism Scores and GCI indicators, at the US state-level, for validation. All correlations are significant ($p < 0.05$).

## 4 Results

We apply our validated knowledge-guided lexica to county-level geolocated Tweets, to gain a more fine-grained understanding of how individualism and collectivism vary regionally.[3]

Figure 4 illustrates this variation, plotting the difference between individualism and collectivism score. The deep south shows high levels of collectivism (dark red) and low levels of individualism (light blue). Conversely, the West coast and the Northeast show low levels of collectivism (light red) and high levels of individualism (dark blue). Counties with under 100 users are poorly represented (Giorgi et al., 2018) and are colored gray.

[3] We release our lexica, county-level and state-level scores, and relevant code at https://anonymous.4open.science/r/knowledge_driven_lexica-E8EE/

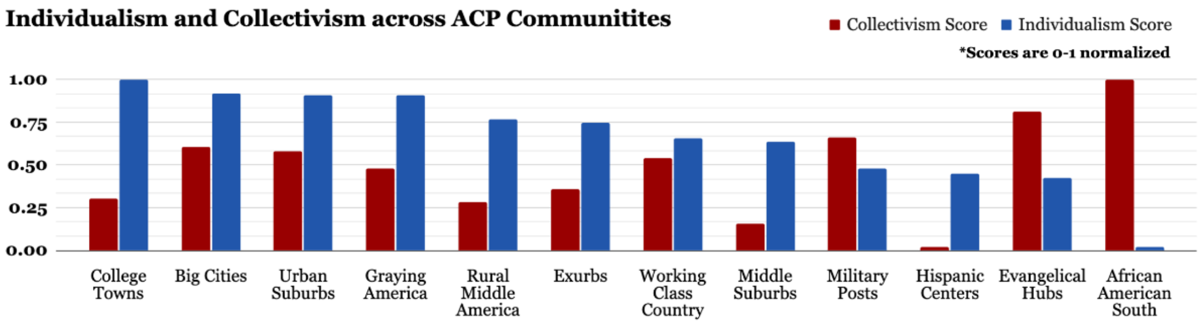


Figure 3: A comparison of collectivism (red) and individualism (blue) scores across communities defined by the American Communities Project, ordered from most individualistic (left) to least individualistic (right). We only analyze communities with over 40 included counties. Scores are 0-1 normalized.

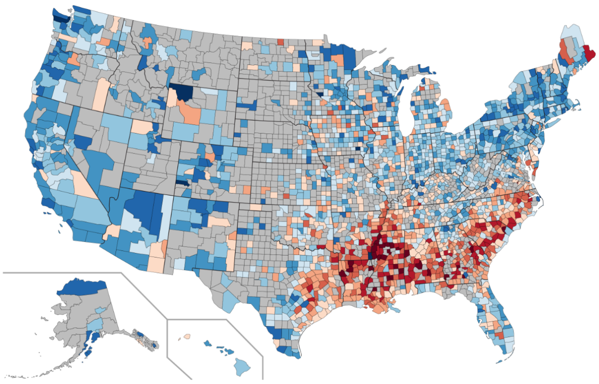

Figure 4: Collectivism (red) and individualism (blue) across US counties. Dark red = higher collectivism and dark blue = higher individualism. Gray counties have insufficient Tweets to estimate a score.

**Community-level insights.** Cultural similarity is not always based on geographical proximity; two cities hundreds of miles apart may be more similar than a city and a rural farm a few miles away (Guntuku et al., 2021). To show how county-level analyses of culture can help us better understand *communities*, we additionally use 15 community types (e.g., College Towns, Urban Suburbs) identified by the American Communities Project (ACP). The ACP identified these communities based on socio-demographic attributes, not spatial clusters of counties. Previous studies have used these community types to identify cultural variation in excessive alcohol consumption (Giorgi et al., 2020) and self-reported physical and mental health (Aggarwal et al., 2023; Mangalik et al., 2023).

Figure 3 shows county-level individualism and collectivism scores grouped into their corresponding ACP community (see Table 3 for counts.) These results provide novel insights into how culture varies regionally. For example, College Towns and Big Cities are highly individualist. These areas are also more affluent and have higher rates of education (ACP, 2023). This fits with prior research finding that people who are wealthy or educated tend to be more individualistic (Binder, 2019). In contrast, the data shows that Evangelical Hubs and the African American South are highly collectivist. These communities are tight-knit and religious areas (ACP, 2023), which have been linked to collectivism (Pelham et al., 2022). Military Posts are also more collectivist, which fits with the tight ties in military service and "duty to one's troop." This insight is helpful because we know of no cultural psychology research comparing military communities with civilian communities.

## 5 Conclusion & Future Work

We present a method to efficiently measure cultural variation by leveraging domain knowledge from cultural psychology and language models to create knowledge-guided lexica. These lexica, applied to social media language, can estimate cultural differences at fine-grained geographic levels, such as states, counties, and communities.

Future work could build on this method to get deeper insights into communities and cultures. For example, our method could be used to identify more types of Tweets that mark cultural differences; we encourage researchers to build more sophisticated models on these identified Tweets. Additionally, our method is easily extendable to other cultural dimensions, such as tightness/looseness, future orientation,etc. This method could also measure cultural variation globally, which requires analyzing different languages. Since our method is language agnostic, it can easily extend to non-English settings by leveraging multilingual embeddings.

## 6 Limitations

While we label each county for individualism and collectivism, we note that regions do not have a single culture. Within all regions, there is heterogeneity of cultural values and beliefs. Since we use an open-source Twitter corpus, we also have poor coverage of counties with little to no Twitter data. Additionally, not all aspects of culture are revealed in language – we are limited to analyzing only what people say online.

In our analyses, we do not control for race, income, or other demographic variables. We know cultural values are correlated to some demographic variables. For example, collectivism and individualism vary with income. Future work can improve upon these estimates by accounting for individual demographics. Additionally, it is unclear if this method of measuring cultural variation will work for all cultural dimensions. For example, power distance (Hofstede, 2011) involves the relationship dynamics of two people, which might make it difficult to capture with lexica.

## 7 Ethical Considerations

The goal of studying cultural variation is to better understand cultures, not individuals. Nonetheless, the characterization of culture has the danger of stereotyping individuals. Individuals within each culture vary greatly. Studying culture can help us understand differences in psychology, but we should not assume that a cultural average will definitely apply to a particular individual from that culture.

All data used in this study are publicly available. While geolocated Twitter data is used, only aggregated spatial-level data is reported. That is, no person-level identifiable information is used or released for this study.

## References


ACP. 2023. Home - american communities project - americancommunities.org. https://www.americancommunities.org/. [Accessed 14-08-2023].

Arnav Aggarwal, Sunny Rai, Salvatore Giorgi, Shreya Havaldar, Garrick Sherman, Juhi Mittal, and Sharath Chandra Guntuku. 2023. A cross-modal study of pain across communities in the united states. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1050–1058.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, 191:105184.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Carola Conces Binder. 2019. Redistribution and the individualism–collectivism dimension of culture. *Social Indicators Research*, 142(3):1175–1192.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Sven Buechel, Susanna Rücker, and Udo Hahn. 2020. Learning and evaluating emotion lexicons for 91 languages. *arXiv preprint arXiv:2005.05672*.

Dante Chinni and James Gimpel. 2011. *Our patchwork nation: The surprising truth about the" real" America*. Penguin.

Dov Cohen. 2001. Cultural variation: considerations and implications. *Psychological bulletin*, 127(4):451.

Yilin Geng, Zetian Wu, Roshan Santhosh, Tejas Srivastava, Lyle Ungar, and João Sedoc. 2022. Inducing generalizable and interpretable lexica. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4430–4448.

Salvatore Giorgi, Daniel Preoţiuc-Pietro, Anneke Buffone, Daniel Rieman, Lyle Ungar, and H. Andrew Schwartz. 2018. The remarkable benefit of user-level aggregation for lexical-based population-level predictions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1167–1172, Brussels, Belgium. Association for Computational Linguistics.

Salvatore Giorgi, David B Yaden, Johannes C Eichstaedt, Robert D Ashford, Anneke EK Buffone, H Andrew Schwartz, Lyle H Ungar, and Brenda Curtis. 2020. Cultural differences in tweeting about drinking across the us. *International journal of environmental research and public health*, 17(4):1125.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Sharath Chandra Guntuku, Alison M. Buttenheim, Garrick Sherman, and Raina M. Merchant. 2021. Twitter discourse reveals geographical and temporal variation in concerns about covid-19 vaccines in the united states. *Vaccine*, 39(30):4034–4038.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bi Puranen, et al. 2020. World values survey: Round seven–country-pooled datafile. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat*, 7:2021.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Geert Hofstede. 2011. Dimensionalizing cultures: The hofstede model in context. *Online readings in psychology and culture*, 2(1):8.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Siddharth Mangalik, Johannes C Eichstaedt, Salvatore Giorgi, Jihu Mun, Farhan Ahmed, Gilvir Gill, Adithya V Ganesan, Shashanka Subrahmanya, Nikita Soni, Sean AP Clouston, et al. 2023. Robust language-based mental health assessments in time and space through social media. *arXiv preprint arXiv:2302.12952*.

Shigehiro Oishi, Ed Diener, Richard E Lucas, and Eunkook M Suh. 2009. Cross-cultural variations in predictors of life satisfaction: Perspectives from needs and values. *Culture and well-being: The collected works of Ed Diener*, pages 109–127.

Brett Pelham, Curtis Hardin, Damian Murray, Mitsuru Shimizu, and Joseph Vandello. 2022. A truly global, non-weird examination of collectivism: The global collectivism index (gci). *Current Research in Ecological and Social Psychology*, 3:100030.

Jeanne L Tsai, Brian Knutson, and Helene H Fung. 2006. Cultural variation in affect valuation. *Journal of personality and social psychology*, 90(2):288.

Joseph A Vandello and Dov Cohen. 1999. Patterns of individualism and collectivism across the united states. *Journal of personality and social psychology*, 77(2):279.

# A Open-Source Twitter Corpus

We use the County Tweet Lexical Bank, an open source data set of features extracted from a corpus of 1.5 billion tweets from approximately 6 million US county mapped users (Giorgi et al., 2018). While the full details of the dataset can be found in the original paper, we give a high-level summary to aid the reader. The dataset is built from a larger corpus which is a 10% sample of Twitter from 2009-2015 (over 30 billion tweets). These tweets are then mapped to US counties via latitude and longitude coordinates associated with the tweets or self-reported location information in the Twitter user's profile (a free text field). A Twitter user is included in this data set if they have posted at least 30 or more English tweets, and a county is included if at least 100 such users are mapped to that respective county. This process resulted in 1.5 billion tweets mapped to over 2000 US counties.

# B Scalability Calculations

We outline the proposed costs of using various LM-based techniques to label our corpus of 1.5 billion Tweets:

**Proposed cost of GPT-4** As of August 2023, the OpenAI API rate for GPT-4 is $0.06 cents per 1,000 tokens. Assuming 10 tokens per Tweet, we get:

$$1.5e9\,\text{Tweets} \times \frac{10\ \text{Tokens}}{\text{Tweet}} \times \frac{\$0.06}{1,000\ \text{Tokens}} \quad (1)$$

This yields a total cost of $900,000.

**Proposed cost of GPT-3.5** As of August 2023, the OpenAI API rate for GPT-3.5 is $0.002 cents per 1,000 tokens. Assuming 10 tokens per Tweet, we get:

$$1.5e9\,\text{Tweets} \times \frac{10\ \text{Tokens}}{\text{Tweet}} \times \frac{\$0.002}{1,000\ \text{Tokens}} \quad (2)$$

This yields a total cost of $30,000.

# C Validation: Additional Details

All six variables in the Global Collectivism Index – total fertility rate, living arrangements (% households with people over 60 and children under 14), stability of marriage (divorce rate to marriage rate ratio), religiosity, collective transportation, and in-group bias (approximated by compatriotism due to lack of state-level data) – are replicable at the state-level using US census data and WVS data.

| | |
|---|---|
| **Collectivism Seed Words** | duties, responsibilities, role, fit in, community, sacrifice, shame, required, rules, honor, support, rely, loyal, respect, obedience |
| **Individualism Seed Words** | humans, humanity, worldwide, universal, mankind, everyone, collective, global, equity, imagination, cooperate, cooperation, shared, joint, identity, guilt, diversity |

Table 2: Seed words hypothesized to identify individualism and collectivism on social media, provided by a domain expert in cultural psychology.

| ACP Community | Num Counties |
|---|---:|
| Exurbs | 207 |
| Graying America | 164 |
| African American South | 252 |
| Evangelical Hubs | 269 |
| Working Class Country | 159 |
| Military Posts | 70 |
| Urban Suburbs | 103 |
| College Towns | 151 |
| Big Cities | 46 |
| Hispanic Centers | 87 |
| Rural Middle America | 403 |
| Middle Suburbs | 77 |

Table 3: Number of included counties for each ACP community included in the analysis in Figure 3.


Figure 5: Word clouds visualizing our individualism lexica (blue, top) and collectivism lexica (red, bottom). Larger words have a higher weight, while smaller words have a lower weight.

Note that when aggregating US census data from county-level to state-level, we treat each county as being weighted equally, due to disproportionate amounts of data coming from big cities.

In order to determine which of these six replicated variables also measure collectivism within the United States, we sample subsets of the six variables and use Cronbach's alpha to measure internal consistency. We limit the subsets to size three or larger, following Pelham and colleagues' (Pelham et al., 2022) validation of three collectivism indicators per nation. The set of living arrangements, religiosity, and compatriotism yielded the highest Cronbach's alpha (0.702), so we chose these three variables as a validation metric.

Table 1 shows the correlations between each of the three validation variables, the collectivism lexicon score, and the individualism lexicon score for US states. Collectivism word use positively correlates with all validation outcomes, and individualism word use correlates negatively. We further validate against median income at the state-level. Prior research has found that income is negatively correlated with collectivism (Pelham et al., 2022) Similarly, income was negatively correlated with our collectivism lexicon scores (-0.273) and positively with our individualism lexicon scores (0.424). We also observe a strong negative correlation (-0.470) between our individualism and collectivism scores at the US state level.

We also validate against Vandello and Cohen's collectivism scores (Vandello and Cohen, 1999). We see a positive correlation (0.388) with our collectivism lexicon scores and a negative correlation (-0.374) with our individualism lexicon scores. This suggests that our lexica measurements are indeed tapping into real cultural differences.