Title: Dictionary validation appendix for "Lies and presidential debates: How political misinformation spread across media streams during the 2020 election"
Authors: Jaren Haber (1), Lisa Singh (1,2), Ceren Budak (3,4), Josh Pasek (5), Meena Balan (1), Ryan Callahan (1), Rob Churchill (2), Brandon Herren (1), Kornraphop Kawintiranon (2)
Date: December 17th, 2021
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix F: Dictionary validation

An essential step in workflows involving dictionaries (conceptually related word or phrase lists) is to check whether they mean what we think they mean—that is, to validate them in the study context (Grimmer & Stewart, 2013). The question here is how effectively our dictionaries capture media discussion of our misinformation-related topics, as opposed to unrelated conversation about a different topic. While there are different approaches for testing the validity of dictionaries, our approach is to manually annotate a random set of posts identified by the dictionaries as being misinformation-related.[1]

Our validation procedure involved hand-coding a sample of 101 social media posts (tweets) for three of the most common misinformation-related topics (*climate change*, *Biden personal attacks*, and *election integrity*) and three of the least common (*healthcare*, *taxes*, and *military*) on social media. We randomly selected sample tweets from those tweets that matched a phrase in one of the myth dictionaries. For each topic, we selected a single phrase list related to a specific false claim: respectively, these are claims about Hunter Biden, his laptop or the Ukraine scandal; about the election being "rigged" or issues with mail-in ballots; that the California forest fires were caused by forest mismanagement (not climate change); that prices of insulin and other prescription drugs were falling; that Trump paid millions of dollars vs. $750 in federal income tax in 2016 and 2017; and that Biden called military members "stupid bastards." For the most frequent topics, each tweet was triple coded by independent coders; 18 coders each labeled approximately 50 tweets. For the least common topics, each tweet was double coded, and the first author resolved any disagreements. In both cases, coders answered two questions to determine the high-level topic and the specific topic of the post. These questions and the response options are shown in Table 1.

[1] Here we measure precision of our dictionary as opposed to coverage. Given that our dictionaries are non-exhaustive and may leave out phrases relevant to our myths, we have likely undercounted the misinformation conversation.

Table 1. *Coding options for our two dictionary validation exercises.*

Question	Options for frequent topics validation	Options for infrequent topics validation
Q1. Which high-level topic is the post about?	*Biden's family/personal life*[2]	*Healthcare*
	Trump's family/personal life	
	Election integrity	*Taxes*
	Climate change	
	Health	*Military*
	Economy	
	None of the above	
Q2. Which specific topic is the post related to?	*The election being "rigged"/issues with mail-in ballots*	*Prices of insulin and other prescription drugs are falling*
	Hunter Biden, his laptop or the Ukraine scandal	
	Forest mismanagement as a cause of the wildfires in CA	*Trump paid millions of dollars vs. $750 in federal income tax in 2016/2017*
	The US having the greatest economy in the history of the country	
	Serious people (like Fauci) saying that masks are not important	*Biden called military members "stupid bastards"*
	None of the above	

Based on the hand-coding results, we measured our dictionaries' accuracy, or the proportion of tweets flagged as pertaining to a misinformation-related topic that were actually about that topic. For the most frequent topics, we also measured their task-based agreement, or the proportion of coders that agreed on the most common label for a given tweet; and their inter-rater reliability, or the overall consistency between coders (as measured by Krippendorff's alpha). While task-based agreement measures reliability at the task level, accuracy and alpha are computed at the question level—that is, they compare across topics and across myths. Table 2 shows our validation results for the most frequent topics using all of these metrics.

[2] Cells in the right-side columns in *italics* were used to select tweets for validation purposes—in other words, we expected these topics to be represented in our sample. The false claims not in italics were included to increase the number of options given to coders, making the test more rigorous.

Table 2. *Dictionary validation results for most frequent topics.*

Level of measurement	Task-based agreement	Alpha	Accuracy
Topic (Q1)	0.932	0.863	0.970
Myth (Q2)	0.908	0.818	0.954

The task-based agreement for both questions is over 0.9, and the alpha score for both questions is over 0.81, indicating that independent coders tend to assign the same topic and myth to a given tweet. These results suggest high reliability in our phrase-based method for identifying misinformation-related social media content. Moreover, the very high accuracies (over 0.95) demonstrate that the hand-selected topic and myth for a given tweet typically match the topic and myth predicted by phrase matching, evidencing the validity of our dictionary-based method for the most frequent topics. Table 3 shows the resulting accuracies for each of the least common topics we validated.

Table 3. *Dictionary validation accuracies for least frequent topics.*

Level of measurement	***Healthcare* topic**	***Taxes* topic**	***Military* topic**
Topic (Q1)	0.989	1.00	0.593
Myth (Q2)	0.851	0.842	0.593

These accuracies are also generally high, reaching full or nearly full general agreement about the *taxes* and *healthcare* topics and about 0.84 for their specific myth. However, both accuracies for the *military* topic were 0.59, due almost entirely to two overly broad phrases indicating names of political importance: "Andrew Bates" and "Karen Johnson." The former was a campaign official who sought to contextualize Biden's comments, while the latter was mentioned specifically in Biden's speech to service members—but tweets selected by matching these phrases rarely relate to the relevant myth (about Biden's "stupid bastards" comment). Nonetheless, these results overall suggest that such imprecise phrases were rare and that our less common topics are also effective in capturing misinformation discussion in the media.