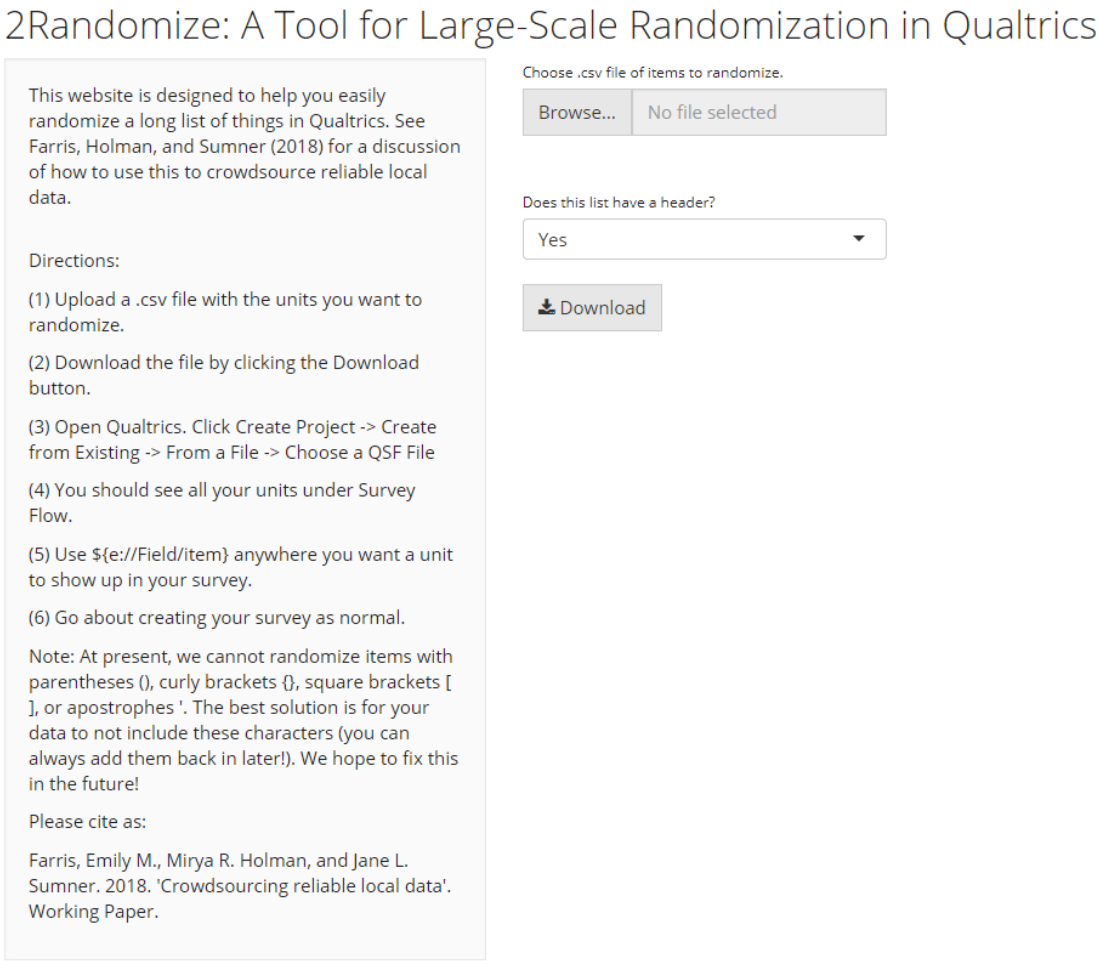
Appendix for Crowdsourcing reliable local data

Jane Lawrence Sumner, Emily M. Farris, and Mirya R. Holman

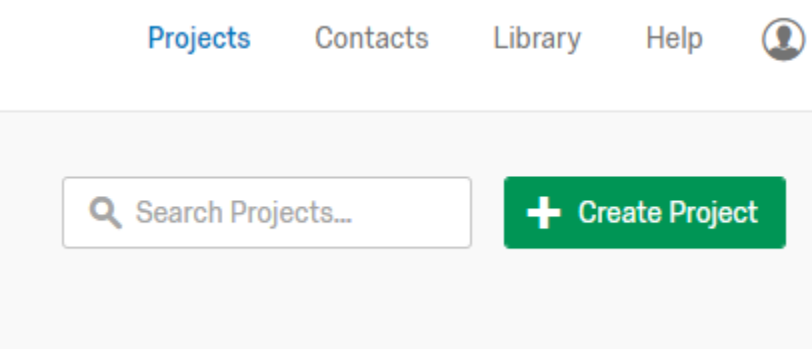
Appendix A: A How-To Guide for Crowdsourcing Data Collection

This data guide assumes that you have established three different accounts: an MTurk "Requester" account, a TurkPrime account (ideally an academic one to save you money), and a Qualtrics account.

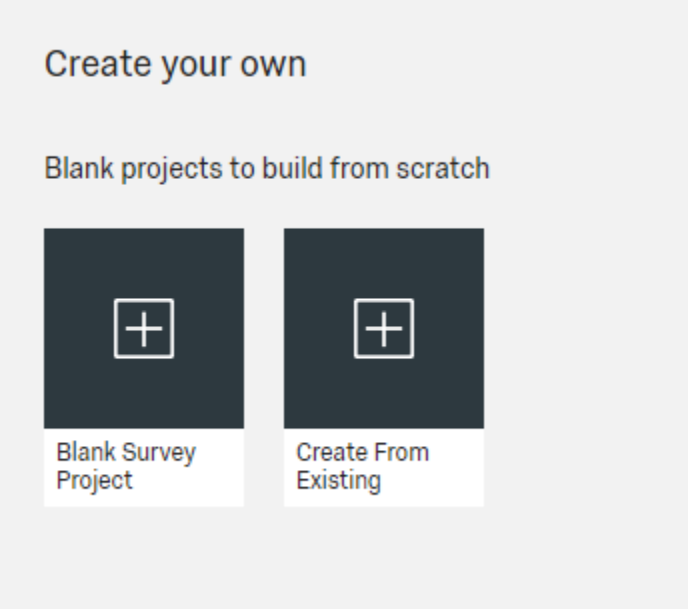
Step 1: Using 2Randomize (available here: https://jlsumner.shinyapps.io/2randomize/) upload whatever data you would like randomly assigned to the MTurk Workers in .csv form. You should be able to then download your data as a qsf file by clicking "Download." Save that file somewhere where you can find it! If you are prepared to program the randomization in Qualtrics, you can skip to step 5.



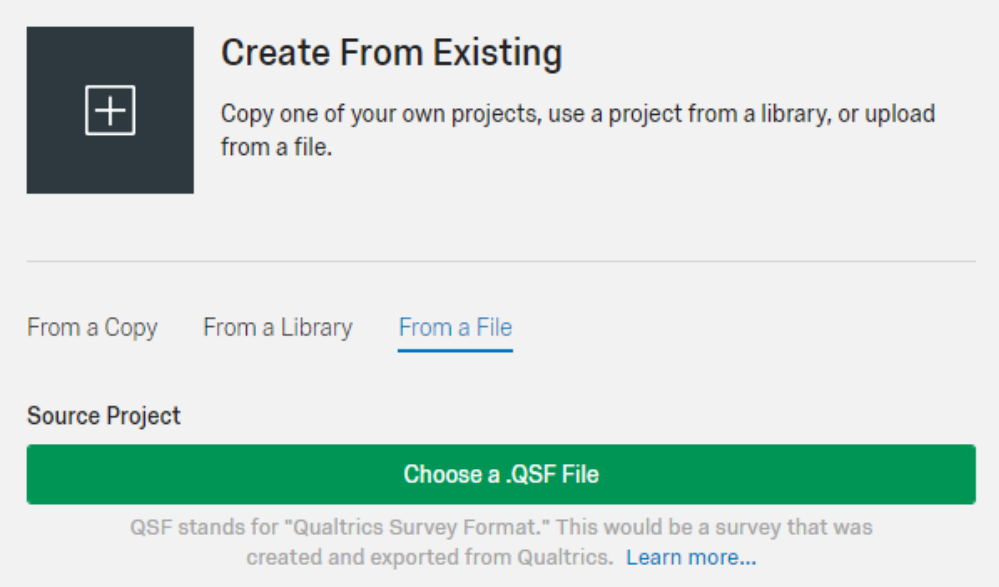
Step 2: Log into Qualtrics. From there, select "Create Project"



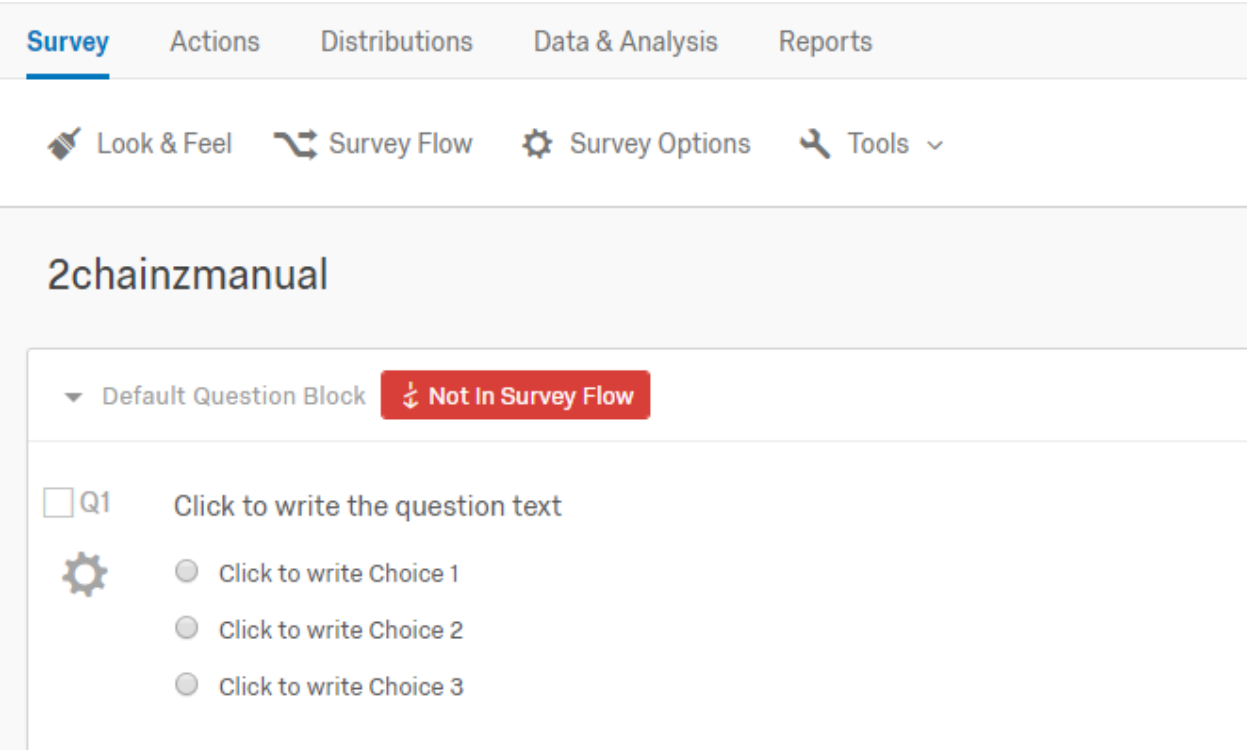
Step 3: Selection “Create from existing”



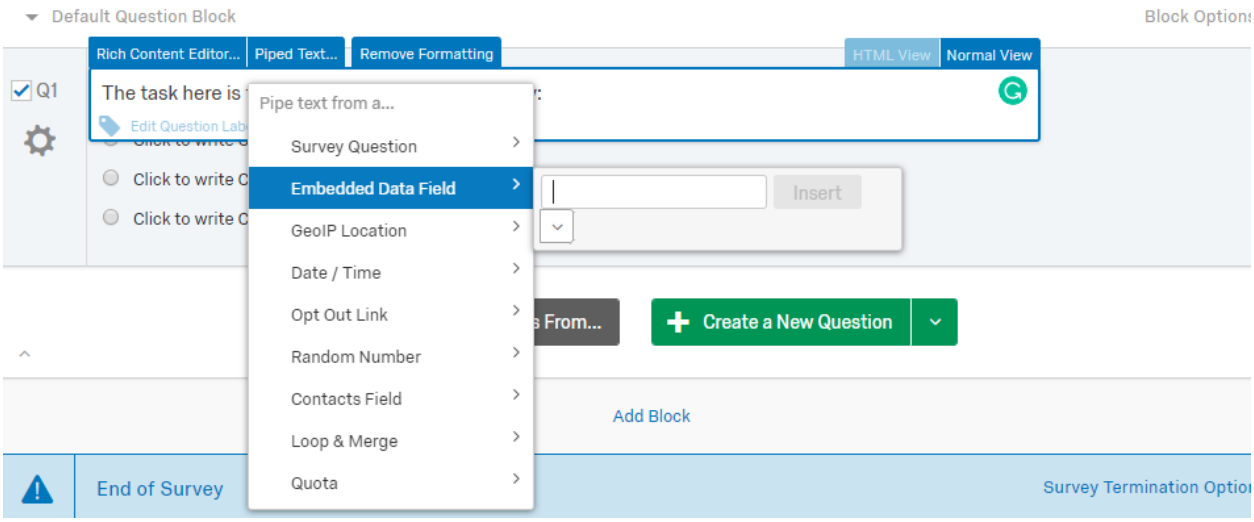
Step 4: Select “From a file” and select “Choose a .QSF file.” Navigate to where you saved your .qsf file from Step 1. Name your project. Click “Create Project”



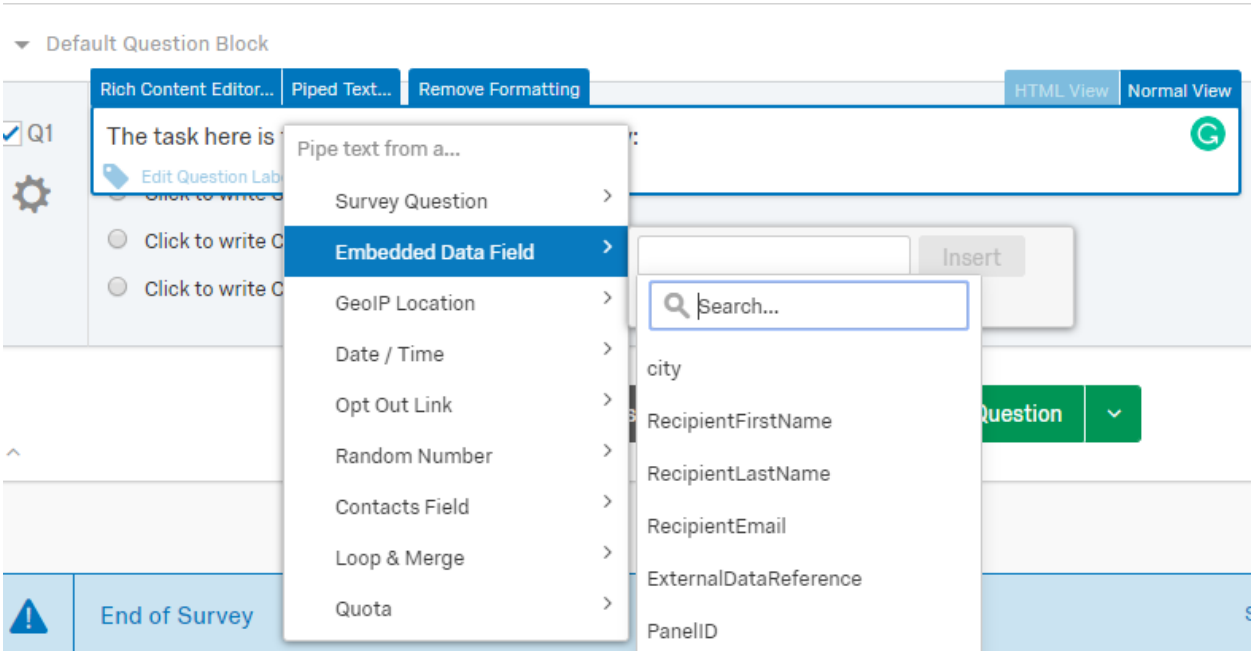
Step 5: You will see a blank survey. Here is where you will create questions that relate to your data coding task. You should write this exactly as you want your workers to see it!



Step 6: Are you ready to assign randomization? Write the question where you would like the randomized information inserted. Within the question text, select "Piped Text" → Embedded Data Field → click on the downward arrow



Step 7: Select the data field that you would like randomized (in our case, "City")



Step 8: Select "Preview Survey" at the top to make sure that your randomization is working as planned.

Step 9: Once your survey is complete, launch it and copy the link for it – you will need this in TurkPrime!

TurkPrime Steps:

We use TurkPrime as an add-on because it provides to us an ability to track how much we are paying our workers (to ensure an ethical wage), the opportunity to change the payment, and the ability to pause and restart our MTurk task.

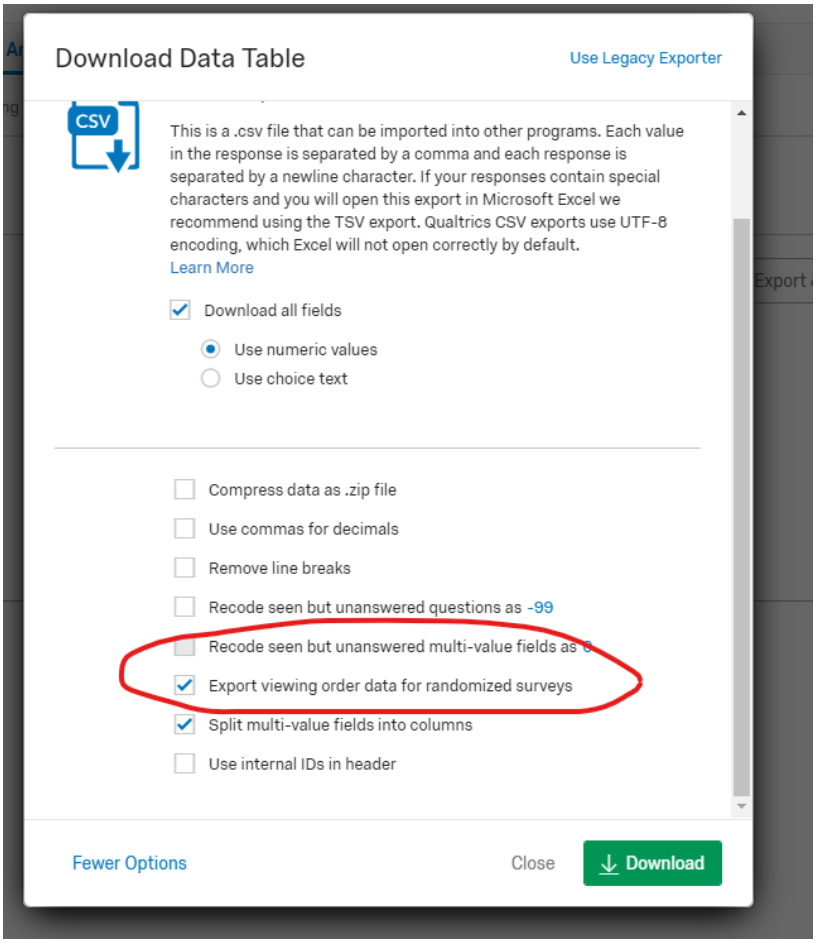
Step 10: Log into TurkPrime and click "Create a Study" at the top, selecting "MTurk Toolkit." Then select "Create a study with your own MTurk account."

Step 11: Calculate the number of participants. You will want 3 to 4 workers per task, so if you have 200 cities that you want to collect data on, estimate at least 600 "survey participants." You can choose demographic characteristics, but keep in mind that TurkPrime will **charge you for those choices.** Nothing in our research would suggest that any of these choices will improve the quality of your data.

Step 12: Make your way through the TurkPrime process. Set a price, select a title, description, instructions, and keywords, and make decisions about payment. Please pay MTurk workers at least federal minimum wage (ideally more than $10/hour)! Copy your url from the Qualtrics survey into TurkPrime. When given an opportunity, select "Microbatch", which will save you money in MTurk's system.

Step 13: Launch your survey! TurkPrime's dashboard lets you keep track of how long the task is taking each worker, the average hourly wage, and the number of completed tasks. One warning: TurkPrime's tasks will time out after a week. If your task it taking this long, it will need to be need to be copied to a new task and restarted with the number of incompletes as the number of workers in the second task.

Step 14: After your task has completed, go to Qualtrics and pause data collection. Make sure to close out all surveys that remain open. Navigate to the "Data and Analysis" tab and select "Export & Import" and then "Export Data". Select "More Options" and make sure the "Export viewing order data for randomized surveys" option is selected. This will ensure that the piece of data randomized to the worker appears in the dataset.


Download Data Table
Use Legacy Exporter
CSV
This is a .csv file that can be imported into other programs. Each value in the response is separated by a comma and each response is separated by a newline character. If your responses contain special characters and you will open this export in Microsoft Excel we recommend using the TSV export. Qualtrics CSV exports use UTF-8 encoding, which Excel will not open correctly by default.
Learn More
Download all fields
Use numeric values
Use choice text
Compress data as .zip file
Use commas for decimals
Remove line breaks
Recode seen but unanswered questions as -99
Recode seen but unanswered multi-value fields as
Export viewing order data for randomized surveys
Split multi-value fields into columns
Use internal IDs in header
Fewer Options
Close
Download

Appendix Table B1: Research start-up funds

	No research funds	<$1,000	$1,000-2,499	$2,500-4,999	$5,000-7,499	$7,500-10,000	>$10,000
PhD	6%	0%	2%	8%	21%	12%	40%
MA	43%	3%	3%	17%	13%	3%	3%
BA (standalone)	33%	4%	21%	21%	6%	2%	4%
BA (combined)	50%	7%	21%	7%	0%	0%	7%
Social Science	50%	0%	0%	0%	50%	0%	0%
Private	19%	7%	15%	20%	13%	4%	11%
Public	32%	0%	8%	10%	12%	7%	19%
Total	28%	3%	10%	14%	13%	6%	16%

Note: Data from the 2012 APSA Salary Survey. Percentages calculated by the authors.

Appendix C: What Makes a Good Coder?

For the mayoral analysis, we also set out to determine what, if anything, predicts (a) a coder being in the consensus, and (b) a coder being correct (if the consensus is incorrect), with the aim of providing best practices for targeting good coders. For ethical and practical reasons, we collected no personal data on our coders beyond what Qualtrics automatically collects. As such, we know the time a user started and ended the task, and therefore both the time of day they did it (Figure C1) and how long it took (Figure C2), and their state, so we can evaluate whether the coder is located in the same state as the city they coded (108 coders, or 5.6%). In addition, we asked researchers questions about the task, including the URLs they used to find both the mayor's email and the election date, and whether they wanted to volunteer to be a part of additional collection efforts (76.2% did). From those URLs, we were able to determine common sources of information (Table C1). Any or all of these may predict correctness, which we expected might create additional

guidelines and best practices (e.g., time constraints on the HIT, giving guidance about which sources to use)

.

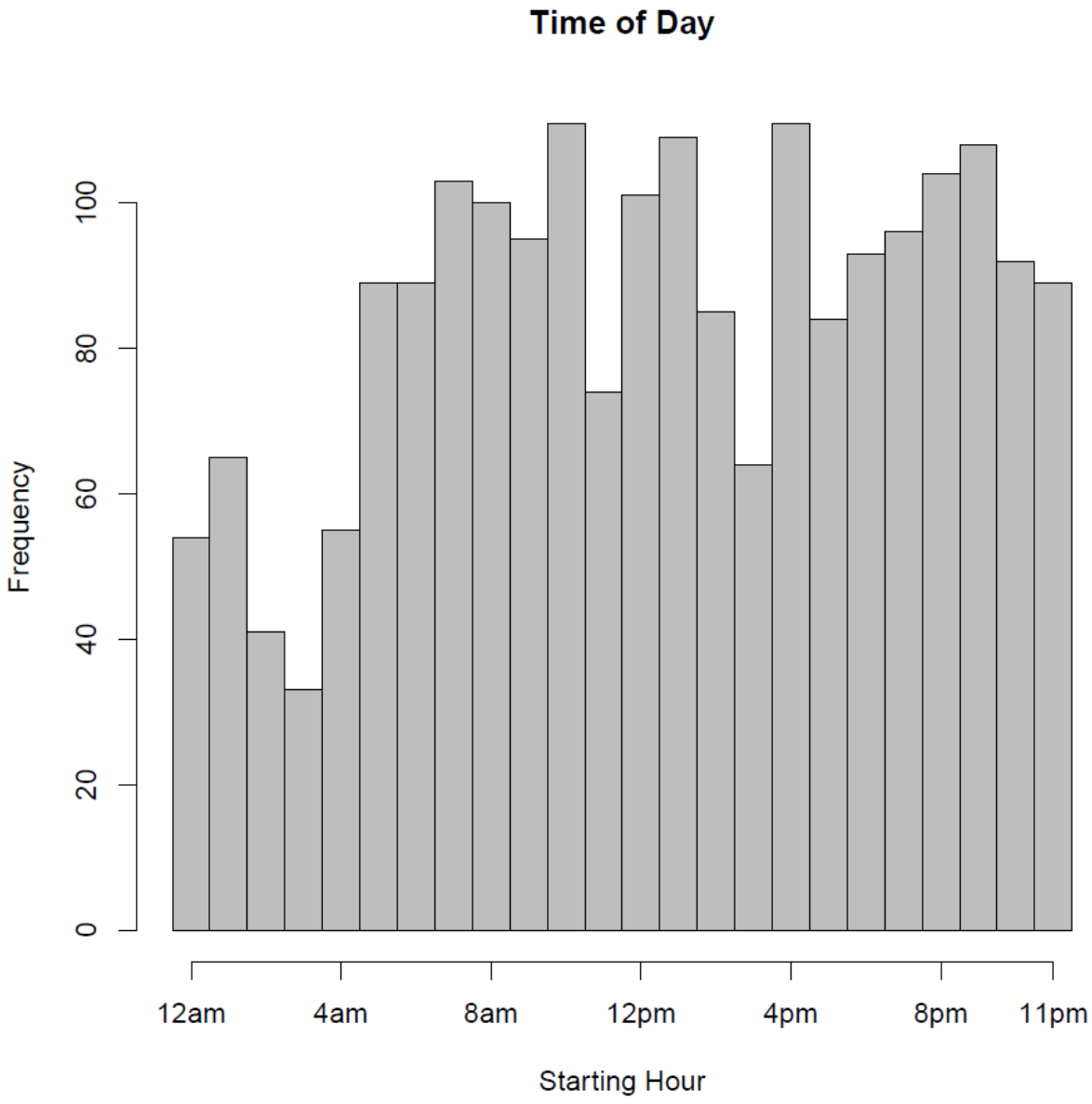


Figure C1: Distribution of starting hour for coders.

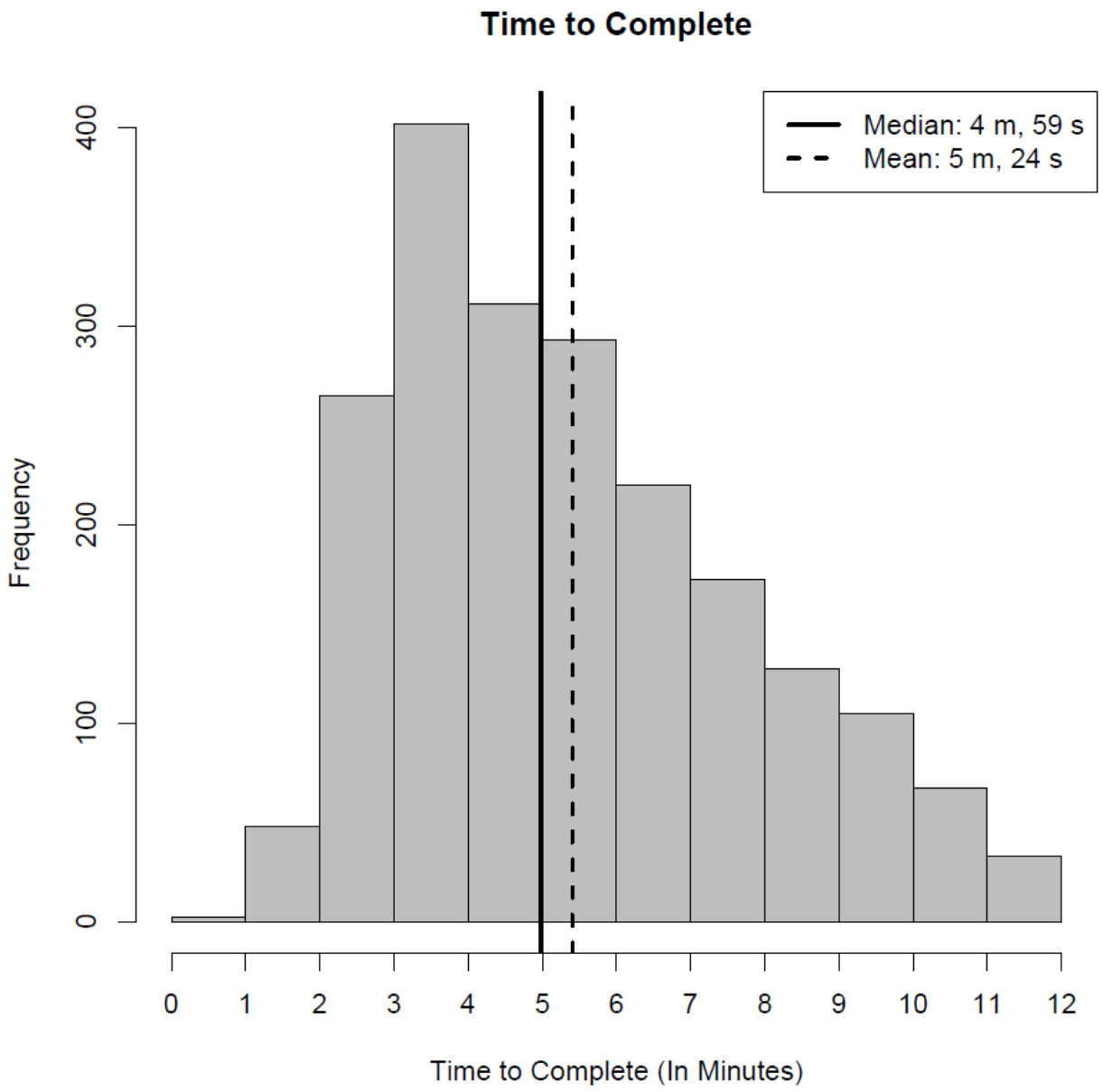


Figure C2: Distribution of time required to complete task, measured in minutes.

	Email	Election Date
Wikipedia	0.1%	13.9%
.gov Website	46%	18.4%
Ballotpedia	0%	8.6%
News website (contains 'news' or 'local' in URL)	0.8%	20.3%

Table C1: Percentage of users using common websites.

To test whether any of these factors are associated with correct information or a higher or lower chance of a worker ending up in the consensus (or both), we run a series of logit models (Table C2). We include as independent variables all of the data we expect might have an effect, and we include duration in quartiles, rather than in absolute time, for the sake of generalizability.

We find that a strong and consistent predictor of being in the consensus is the size of the city. The larger the city's population, the more likely its coders were to be in the consensus for both surname and election date. The models suggest that coders who used Wikipedia were less likely to be in the consensus than coders who used other websites (excluding the others in the model), even when we control for city size, and that being slightly faster than the median is predictive of being out of the consensus for date, but neither of those are especially telling. We conclude that consensus is driven by ease of access to information and the degree to which a worker seeks out up to date information; neither relates to the characteristics of the workers themselves.

Similarly, we find that the strongest two predictors of being 'correct' are city population and being in the consensus. There is slightly more robust evidence here that speed hurts, meaning that the fastest coders were less likely to be in the consensus or correct (even when controlling for city size, a potential proxy for the ease of finding the data) and that Wikipedia can be misleading.

	Dependent variable:					
	Coder is in Consensus			Coder is Correct		
	Surname	Email	Date	Surname	Email	Both
Same State	-0.459	-0.068	0.524	-0.325	-0.332	-0.219

	(0.335)	(0.311)	(0.358)	(0.268)	(0.218)	(0.205)
International	11.038	10.592	10.550	-14.452	-14.006	-12.852
	(535.411)	(324.744)	(324.744)	(324.744)	(535.411)	(324.744)
Second Quartile for Duration	-0.239	0.078	-0.444**	-0.441**	-0.142	-0.150
	(0.264)	(0.225)	(0.199)	(0.197)	(0.143)	(0.134)
Third Quartile for Duration	-0.052	-0.273	-0.308	-0.335*	-0.225	-0.327**
	(0.272)	(0.210)	(0.202)	(0.198)	(0.142)	(0.133)
Slowest Quartile for Duration	-0.177	-0.339	-0.272	-0.512***	-0.221	-0.399***
	(0.267)	(0.209)	(0.205)	(0.196)	(0.144)	(0.134)
Morning	0.171	0.022	-0.053	0.152	-0.010	0.018
	(0.260)	(0.204)	(0.187)	(0.184)	(0.137)	(0.128)
Evening	-0.004	0.054	0.039	0.004	0.024	-0.022
	(0.247)	(0.202)	(0.188)	(0.178)	(0.137)	(0.128)
Night	-0.093	-0.008	-0.021	0.155	0.293*	0.225
	(0.278)	(0.228)	(0.214)	(0.209)	(0.159)	(0.147)
Volunteer	0.207	0.141	0.016	-0.073	0.107	0.090
	(0.210)	(0.171)	(0.163)	(0.160)	(0.118)	(0.110)
.Gov Website source for email		0.645***			0.224**	
		(0.159)			(0.102)	
Wikipedia source for email		-1.732			-13.114	
		(1.435)			(329.496)	
Ballotpedia source for email		.			.	
News website source for email		-1.680***			-1.517*	
		(0.567)			(0.811)	
.Gov Website source for date			0.023			
			(0.186)			
Wikipedia source for date			-0.136			
			(0.242)			
Ballotpedia source for date			-0.214			
			(0.296)			
News website source for email			0.040			

			(0.180)			
In Consensus for last name				3.378***		
				(0.255)		
In Consensus for email					2.079***	
					(0.201)	
Wikipedia source for date or email						-0.479***
						(0.162)
Ballotpedia source for date or email						0.182
						(0.180)
News website source for email or email						0.302**
						(0.120)
.Gov Website source for date or email						0.274***
						(0.094)
City Population (logged)	0.544***	-0.011	0.257**	0.217**	0.071	0.191**
	(0.161)	(0.099)	(0.123)	(0.098)	(0.067)	(0.077)
Constant	-3.750**	1.953*	-0.893	-3.842***	-2.534***	-2.207**
	(1.887)	(1.187)	(1.423)	(1.177)	(0.832)	(0.896)
Observations	1,823	1,823	1,823	1,774	1,774	1,869
Log Likelihood	-453.788	-635.388	-718.281	-742.454	-1,125.981	-1,269.587
Akaike Inf. Crit.	929.576	1,298.776	1,466.562	1,508.908	2,281.962	2,569.175

*p<0.1; **p<0.05; ***p<0.01. Ballotpedia drops out of email models.

Table C2: Estimating consensus and correctness