APPENDIX ONLINE

Design of Study to Address Questions about M Turk.

The studies were designed to address concerns about workers on M Turk in the following ways. First, Turk workers have been viewed as possibly demotivated due to low pay (Kelley 2010; Paolacci and Chandler 2014); however, the respondents for this set of surveys were paid $1.70 for a 10 minute survey as compared to a typical survey paying $0.10 for 3 minutes (Paolacci, Chandler, and Ipeirotis 2010).

Second, two types of gaming are possible when deploying surveys using Mechanical Turk. The first centers on respondents possibly lying in order to be included in a survey. This concern focuses on research needing a target population, for example, a survey looking for smokers over 50 years old (Sharpe Wessling, Huber, and Netzer 2017). However, these studies were designed to be theoretically generalizable rather than statistically generalizable to a defined population (e.g., 'smokers over 50') rendering the first type of gaming not applicable: my only criteria for someone completing the survey was being from the United States and having over 95% acceptance rate for HITS (both verified by Amazon Mechanical Turk). The second type of gaming is due to inattentiveness (Cheung et al. 2017); this issue is further exacerbated as some MTurk workers regularly take surveys and can become inured to typical attention checks (Chandler, Mueller, and Paolacci 2014; Chandler et al. 2015). The design of the factorial vignette survey does not rely upon attention checks since the structure of the data – each individual rating 40 vignettes – allows the researcher to identify problematic respondents who just click through data analysis. A full explanation is in Appendix B as to how I calculated inattentiveness as well as respondent fatigue. I found 2-4% of respondents 'clicked through' across the surveys, which is consistent with previous analysis of MTurk as a crowdsourcing tool for research ("NSF Convergence Workshop on Crowdsourcing" 2018).

Finally, Amazon Mechanical Turk is found to have demographic differences from some targeted populations which can call into question the samples' representativeness (Cheung et al. 2017; Kelley 2010). A number of studies identify issues with data generalizability with Turk samples, which require matching of sample statistics to the target population. However, Turk has been used for theoretical generalizability quite successfully, as in the examination of the relationship between concepts or ideas (Kang, Brown, Dabbish, & Kiesler, 2014; Martin &

Nissenbaum, 2017a; Redmiles, Kross, Pradhan, & Mazurek, 2017). For example, Coppock successfully reproduced 15 experiments on Turk where the treatment effect was replicated (Coppock 2018). Specifically for an examination of online marketing practices, Turk has been used for consumer perceptions in marketing (Goldstein et al. 2014; Yang and Lynn 2014); MTurk captures consumers most likely to be online (Tucker 2014) and is found to be a reliable source of respondents (Daly and Nataraajan 2015).

In the work critiquing the generalizability of Turk samples, the Turk results are compared to phone surveys (Kang et al. 2014) as well as online nationally representative samples. The critiques of Turk samples center on statistical generalizability (Kang et al. 2014; Sharpe Wessling, Huber, and Netzer 2017). This study, on the other hand, is a theoretical examination– therefore the findings will support or not support the hypothesized relationships between vignette factors. Such research seeks the generalizability of ideas rather than the generalizability of data patterns within a specific population (Lynch Jr 1982).

To remain focused on theoretical generalizability, i.e., whether cause-effect relationships hold (Lynch Jr 1982), I examine the relative importance of the vignette factors in the regression analysis rather than the average vignette rating or the average of the control variables. Second, the control variables were standardized into quartiles and respondents are designated as 'high trust,' 'low trust,' etc relative to the other respondents rather than based on a static measure. Finally, the findings are centered on validating the conceptual definition of privacy and confirming (or not confirming) the privacy paradox as a concept. These steps ensure that the focus is on theoretical generalizability rather than statistical generalizability.

Quality Check of Data: Testing for Click Through Respondents

Attention checks within the factorial vignette surveys were not used both due to design issues and because research has suggested that MTurk respondents are experienced and can become inured to regular attention checks. Instead, the data was analyzed for 'click through' responses. Due to the structure of the multi-level data, where each respondent rated 40 vignettes, I could measure the number of times each respondent (a) never moved the slider (a rating of 0), (b) clicked through at the right end point of the slider (+95 to +100), or (c) clicked through at the left end point of the slider (-95 to -100).

The total counts are in Table B1 below. Across all survey respondents (all 4 surveys), seven (7) respondents clicked through with over 20 (out of 40) 0s, 10 respondents clicked through with over 20 responses at the right hand side, and 31 respondents clicked through with over 30 at the left hand side. The criteria for the left hand side was higher because many respondents quite legitimately rated the vignettes untrustworthy. In general, 2-4% of the respondents were then rated inattentive in analyzing the vignette as shown in Table B1. This is in keeping with previous analysis of MTurk as a sample and much lower than the same analysis of Knowledge Networks – a nationally representative online sample – where over 15% of the sample were found to 'click through' and not really take the survey.

Table B1: Click Through Respondents Per Survey

	Survey 1	Survey 2	Survey 3	Survey 4
N	393	381	400	399
V	15720	15240	16000	15960
ClickThru #	12	8	13	15
ClickThru %	3%	2%	3%	4%

Quality Check of Data: Testing for Respondent Fatigue

Additional analysis for respondent fatigue is included below. The sequence number for each vignette was recorded at the time of the factorial vignette surveys. Respondent fatigue was analyzed three ways: (a) comparing the mean and standard deviation for early and later blocks of vignettes, (b) including a dummy variable for early and later blocks of vignettes in the regression models, and (c) splitting the sample into two (early and later vignettes).

First, the average rating task was compared for early and later vignettes._In Table B2, the First 5 vignettes were rated higher on average (less negative). In other words, the last 20 vignettes (Last20Qs) were more negative and respondents were more critical of the second 20 vignettes compared to the first 20 vignettes – even though vignette factors were generated and assigned randomly with replacement. For example, in Survey 4 (with both the security and privacy violations), the average for the first 20 vignettes was -23.92 and the average of the last 20 vignettes was -26.54. However the standard deviation was the same (55.05 versus 54.24 later). Second, a dummy variable was added for first 5 and last 5 (First5Qs and Last5Qs) as well as the first 20 and last 20 vignettes for each survey and the significance was tested in the multi-

level regression analysis. For the same survey, the regression analysis was rerun including the dummy variables, and the coefficient for the dummy First20Qs was 3.16 ($p<0.005$). With the standard deviation being approximately the same for the first and second half, it is not clear which is 'correct' – the first 20 vignettes or the second 20 vignettes, since respondents go through a learning curve with these types of surveys.

Table B2: Differences and Similarities in beginning and ending vignettes.

		All Qs	First5Qs	Last5Qs	First20Qs	Last20Qs
Survey 1	Mean	-8.47	-2.84	-11.77	-5.65	-11.3
	SD	59.94	55.07	52.69	54.58	53.14
	Coef	n/a	6.51 ($p<0.005$)	-3.86 ($p < 0.005$)	6.40 ($p < 0.005$)	n/a
Survey 2	Mean	-16.97	-14.85	-18.9	-15.95	-18
	SD	56.63	55.41	57.67	56.41	56.83
	Coef	n/a	1.83 ($p = 0.13$)	-1.91 ($p = 0.12$)	2.08 ($p = 0.01$)	n/a
Survey 3	Mean	-18.43	-18.15	-20.71	-17.41	-19.45
	SD	54.43	56.15	53.23	55.15	53.68
	Coef	n/a	0.44 ($p = 0.71$)	-1.92 ($p = 0.104$)	1.96 ($p = 0.01$)	n/a
Survey 4	Mean	-25.23	-21.32	-28.9	-23.92	-26.54
	SD	54.66	55.64	53.88	55.05	54.24
	Coef	n/a	4.67 ($p < 0.005$)	-4.08 ($p < 0.005$)	3.16 ($p < 0.005$)	n/a

Finally, the survey samples were split between the first 20 vignettes and second 20 vignettes to see if the relative importance of vignette factors differed in a regression. The coefficients for each regression analysis are in Figure B2. The relative importance (the coefficients) are consistent across the samples. The differences across subsamples are insignificant and the theoretical findings as to the relative importance of vignette factors remain the same.

Figure B2: Comparison of Coefficients of Each Subsample Run in Table B1

There are a few possible reasons why fatigue would not be a problem in this design. First, the number of sentences read is actually akin to most surveys. For example, including standard controls, 1-2 survey instruments which include 10-20+ questions each, additional closing survey questions, and attention checks, a standard research survey could require 40+ *different* sentences to carefully read answer and some are reverse coded. Here, the vignettes are kept fairly simple by design in that the vignettes are standard in their format and the type of rating task asked. This is another reason why the methodology is insistent on one rating task for all the vignettes – answering different types of questions is tiring as they have different error terms. Finally, all vignettes are pooled in a multi-level analysis. Since the vignette factors are randomly assigned with replacement, the small differences in the first and second half of the vignettes as to the mean are averaged out in measuring the relative importance of the coefficients.

REFERENCES FOR APPENDIX ONLINE

Chandler, Jesse, Pam Mueller, and Gabriele Paolacci. 2014. "Nonnaïveté among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers." *Behavior Research Methods* 46 (1): 112–30.

Chandler, Jesse, Gabriele Paolacci, Eyal Peer, Pam Mueller, and Kate A Ratliff. 2015. "Using Nonnaive Participants Can Reduce Effect Sizes." *Psychological Science* 26 (7): 1131–39.

Cheung, Janelle H, Deanna K Burns, Robert R Sinclair, and Michael Sliter. 2017. "Amazon Mechanical Turk in Organizational Psychology: An Evaluation and Practical Recommendations." *Journal of Business and Psychology* 32 (4): 347–61.

Coppock, Alexander. 2018. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods*, 1–16. https://doi.org/10.1017/psrm.2018.10.

Daly, Timothy M, and Rajan Nataraajan. 2015. "Swapping Bricks for Clicks: Crowdsourcing Longitudinal Data on Amazon Turk." *Journal of Business Research* 68 (12): 2603–2609.

Goldstein, Daniel G, Siddharth Suri, R Preston McAfee, Matthew Ekstrand-Abueg, and Fernando Diaz. 2014. "The Economic and Cognitive Costs of Annoying Display Advertisements." *Journal of Marketing Research* 51 (6): 742–52.

Kang, Ruogu, Stephanie Brown, Laura Dabbish, and Sara B Kiesler. 2014. "Privacy Attitudes of Mechanical Turk Workers and the US Public." In *SOUPS*, 37–49.

Kelley, Patrick Gage. 2010. "Conducting Usable Privacy & Security Studies with Amazon's Mechanical Turk." In *Symposium on Usable Privacy and Security (SOUPS)(Redmond, WA*. 2010).

Lynch Jr, John G. 1982. "On the External Validity of Experiments in Consumer Research." *Journal of Consumer Research* 9 (3): 225–39.

Martin, Kirsten, and Helen Nissenbaum. 2017. "Measuring Privacy: Using Context to Expose Confounding Variables." *Columbia Science and Technology Law Review* 18: 176–218.

"NSF Convergence Workshop on Crowdsourcing." 2018. 2018. http://convergence2018.info/.

Paolacci, Gabriele, and Jesse Chandler. 2014. "Inside the Turk: Understanding Mechanical Turk as a Participant Pool." *Current Directions in Psychological Science* 23 (3): 184–88.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making* (5): 411–19.

Redmiles, Elissa M, Sean Kross, Alisha Pradhan, and Michelle L Mazurek. 2017. "How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk and Web Panels to the US."

Sharpe Wessling, Kathryn, Joel Huber, and Oded Netzer. 2017. "MTurk Character Misrepresentation: Assessment and Solutions." *Journal of Consumer Research* 44 (1): 211–30.

Tucker, Catherine E. 2014. "The Reach and Persuasiveness of Viral Video Ads." *Marketing Science* 34 (2): 281–96.

Yang, Sybil, and Michael Lynn. 2014. "More Evidence Challenging the Robustness and Usefulness of the Attraction Effect." *Journal of Marketing Research* 51 (4): 508–13.