Kjelsrud, Anders; Kotsadam, Andreas; Røgeberg, Ole; Brodeur, Abel

Working Paper

A Comment on "Raising Health Awareness in Rural Communities: A Randomized Experiment in Bangladesh and India" by Siddique et al. (2024)

IZA Discussion Papers, No. 17783

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Kjelsrud, Anders; Kotsadam, Andreas; Røgeberg, Ole; Brodeur, Abel (2025) : A Comment on "Raising Health Awareness in Rural Communities: A Randomized Experiment in Bangladesh and India" by Siddique et al. (2024), IZA Discussion Papers, No. 17783, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/316738

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17783

A Comment on "Raising Health Awareness in Rural Communities: A Randomized Experiment in Bangladesh and India" by Siddique et al. (2024)

Anders Kjelsrud
Andreas Kotsadam
Ole Rogeberg
Abel Brodeur

MARCH 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17783

A Comment on "Raising Health Awareness in Rural Communities: A Randomized Experiment in Bangladesh and India" by Siddique et al. (2024)

Anders Kjelsrud
Oslo Metropolitan University

Andreas Kotsadam
The Ragnar Frisch Centre for Economic Research

Ole Rogeberg
The Ragnar Frisch Centre for Economic Research

Abel Brodeur
University of Ottawa, Institute for Replication and IZA

MARCH 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

A Comment on "Raising Health Awareness in Rural Communities: A Randomized Experiment in Bangladesh and India" by Siddique et al. (2024)*

Siddique et al. (2024a) report massive effects of a mobile phone-based health awareness campaign in a randomized field experiment conducted in rural Bangladesh and India during the COVID-19 pandemic. Both awareness and compliance with preventive COVID-19 measures were higher when the information was received by voice call rather than text, and even higher for those receiving both. Reproducing the analyses we identify many severe issues, including that the study did not in fact randomize treatment assignment. We further find implausible response patterns in the data, undisclosed sampling criteria that negate the study motivation, and an (unreported) re-treatment where some of the respondents were also included in a separate study that provided additional COVID-19 information immediately before the last data collection.

Keywords: replication, health awareness, COVID-19

Corresponding author:
Abel Brodeur
University of Ottawa
75 Laurier Ave E
Ottawa, ON K1N 6N5
Canada

E-mail: abrodeur@uottawa.ca

* We are grateful to colleagues in Bangladesh who want to remain anonymous. All errors are our own.

1 Introduction

Siddique et al. (2024a) – henceforth INFO24 – describes a randomized field experiment that compares the effectiveness of text and voice based interventions to raise awareness of and compliance with COVID-19 prevention guidelines in rural villages in Bangladesh and India. Three modes of cell phone based delivery are compared: (i) text messages only, (ii) direct phone calls from a local NGO only, or (iii) both text messages and phone calls. The study uses the "text message only" group as the baseline against which the other treatment arms are compared. The expectation is that voice-based information will be more effective, as *"[...] many people struggle to understand text messages due to high illiteracy among adults in rural areas (Saleh 2020), which often makes text messages an ineffective method of communication among rural people"* (INFO24, p. 640).

Consistent with their expectations, the paper finds that both awareness and compliance are higher for those receiving phone calls instead of text messages alone, and even higher for individuals receiving both. The reported effects are very large – with the treatment raising compliance by 1-1.5 standard deviations in Bangladesh and 2.2-2.7 standard deviations in India.

In this comment prepared for the Institute for Replication (Brodeur et al. 2024), we reproduce and assess the analyses using the publicly available replication package from the original authors (Siddique et al. 2024b). Our work uncovered several issues related to the quality of the data and experimental design used in the paper. All our analyses were successfully reproduced by multiple coauthors. While our findings relate mainly to the samples from Bangladesh, which also provide the bulk of the data and analyses in the original paper, we also find problematic patterns in the data from India. We identify the following main problems:

First, contrary to the claims by the authors, the study intervention was in fact not randomized. Participants were drawn from two earlier studies and the treatment assignments of these earlier studies were simply reused. The treatment effects

estimated may therefore be contaminated by residual effects of earlier treatments or interactions between the different treatments, undermining the internal validity of the study.

Second, while the study motivates the need for voice calls by describing the sample as illiterate, 2/3 of the sample in Bangladesh come from a study where participants were *selected on being literate* (Ahmed et al. 2024). The other part of the sample comes from a study that explicitly recruited households with young children (Guo et al. 2024). The earlier sampling criteria, which partly undermined the rationale behind their research design, were not disclosed in the paper.

Third, we uncover unlikely data patterns across the two endline surveys whereby there are extreme discontinuities in the observation counts precisely at the thresholds used to discretize the outcome variables.

Fourth, a subset of the Bangladesh sample was also enrolled in another randomized field experiment that a subset of the authors conducted simultaneously. That study tested the effect of phone-based mental health counseling and COVID-19 information on mothers (Vlassopoulos et al. 2024). They gathered baseline data after the INFO24 intervention, and exposed a portion of the INFO24 sample to one or two 30 minute calls focused on COVID-19 information just prior to the second INFO24 endline. Merging data from the data repositories of both studies, we also show that a) some variables reportedly gathered at *different* times by the two studies are in fact identical, b) some joint distributions show implausible patterns and excessive regularities, c) other joint distributions that *should* be very similar are not (e.g., occupation of household head).

Fifth, for the Indian sample we document several stark anomalies in the data. This includes a response pattern that is statistically certain to occur never occurs.

In the remainder of this comment we document these findings in more detail.

2 Re-use of study participants from Ahmed et al. (2024) and Guo et al. (2024)

INFO24 states that they have randomly sampled households from a database of individuals previously surveyed by the two participating NGOs. Baseline data are reported to come from these earlier surveys – collected in 2019.[1] However, no further information is disclosed about these previous data collections. Using available data repositories we successfully matched *all* 5,840 observations in the main specification for the Bangladesh sample of INFO24 to the data samples used in Ahmed et al. (2024) (henceforth TURNOUT24, 4,066 matches) and Guo et al. (2024) (hereafter CHILDHOOD24, 1,774 matches).[2] All three studies were conducted in collaboration with the same implementing partner (Global Development Research Initiative – GDRI), and matches had identical RECORD_ID values, as well as identical income and age in the different samples. Given these findings, we are certain that INFO24 reused respondents from TURNOUT24 and CHILDHOOD24.

CHILDHOOD24 analyzes a field experiment in rural Bangladesh to evaluate the effects of two early childhood programs carried out between 2017 and 2019 on child development and parental networks. The programs targeted families with children aged 3 to 5, living in villages with no existing formal pre-school centers. The programs were found to improve children's cognitive and non-cognitive development and parenting practices in the household. TURNOUT24 analyzes a field experiment in rural Bangladesh conducted in 2018 that examined the differential effect of two information campaign treatments when these were implemented in government strongholds relative to opposition strongholds. The interventions were found to increase voter turnout in government strongholds and decrease them in opposition strongholds, shifting the partisan composition of voters towards the incumbent party without affecting overall turnout.

[1] *"Since conducting an extensive baseline was not possible during the pandemic, we matched respondents to data that was collected in 2019 by the same local organizations" (p. 642).*

[2] Guo et al. (2024) is an unpublished working paper but the data repository can be found here: 10.5255/UKDA-SN-855543.

2.1 No randomization of assigned treatments

As described in INFO24, the study used the existing database of surveyed individuals and *"[...] randomly selected about 8,000 phone numbers (where phone numbers belong to either female/male household-heads or their spouses and each number represents a household) in Bangladesh and 1,870 in India" (p.641).* The intervention for the Bangladesh sample was at the village level, and the authors write that they *"randomized 420 villages in Bangladesh to three different treatment arms" (p.641).*

This description is unequivocally false. Comparing the treatment assignment in INFO24 to the treatment assignments for *the same individuals* in CHILDHOOD24 and TURNOUT24 reveals that the treatment assignment was simply inherited from the earlier studies (see Table 1).

This undisclosed fact negates the main purpose of randomized treatment assignment, which is to ensure that no observed or unobserved characteristics differ systematically between the treatment arms in expectation. This is no longer the case when the treatment assignments are re-used: since the earlier treatments had large and persistent effects on multiple outcomes, the treatment groups in INFO24 will necessarily differ on those characteristics prior to the here studied intervention. And since the treatment assignment in INFO24 perfectly coincides with these earlier treatment assignments, the estimated effects of the INFO24 treatments will necessarily be bundled with any treatment effects these earlier treatments potentially had on COVID-19 awareness and compliance, and with any treatment interaction effects that may exist between the different interventions. For instance, we may suspect that the political interventions of TURNOUT24 could alter how participants react to public health information, given that INFO24 themselves note that *"... studies show that political polarization (Allcott et al. 2020) ..., sense of civic duty (Al-Dmour et al. 2020), ... can influence peoples compliance to health directives during the COVID-19 pandemic" (INFO24, p. 640).* We might also expect participants to be more receptive to information from an NGO that had already provided them with preschools that markedly improved their children's school performance

as well as the internal household dynamic through improved parenting practices.

In sum, the treatment estimates in INFO24 rest entirely on assumptions of no carryover effects, no differential attrition, and no baseline imbalances – none of which are discussed or even disclosed in the paper.

Table 1: Treatment groups and earlier treatment

	Treatment status in INFO24		
	SMS only	Call only	SMS+Call
Treatment status in CHILDHOOD24:			
Home-visit only	0	0	648
Pre-school only	627	0	0
Pre-school+home visit	0	499	0
Treatment status in TURNOUT24:			
Control	1,564	0	0
Legit	0	0	1,239
Policy	0	1,263	0

Note: The table illustrates the number of observations for each treatment arm in INFO24 and how they relate to the prior randomizations in CHILDHOOD24 and TURNOUT24. The upper panel shows the number of observations in the main estimation sample of INFO24 that are taken from the sample used for CHILDHOOD24, and illustrates that the treatment allocation of these 1,774 observations are completely based on an alleged randomization in CHILDHOOD24. The bottom panel shows that the treatment status of the 4,066 observations taken from TURNOUT24 are completely based on an alleged randomization in TURNOUT24.

2.2 Sampling of study participants

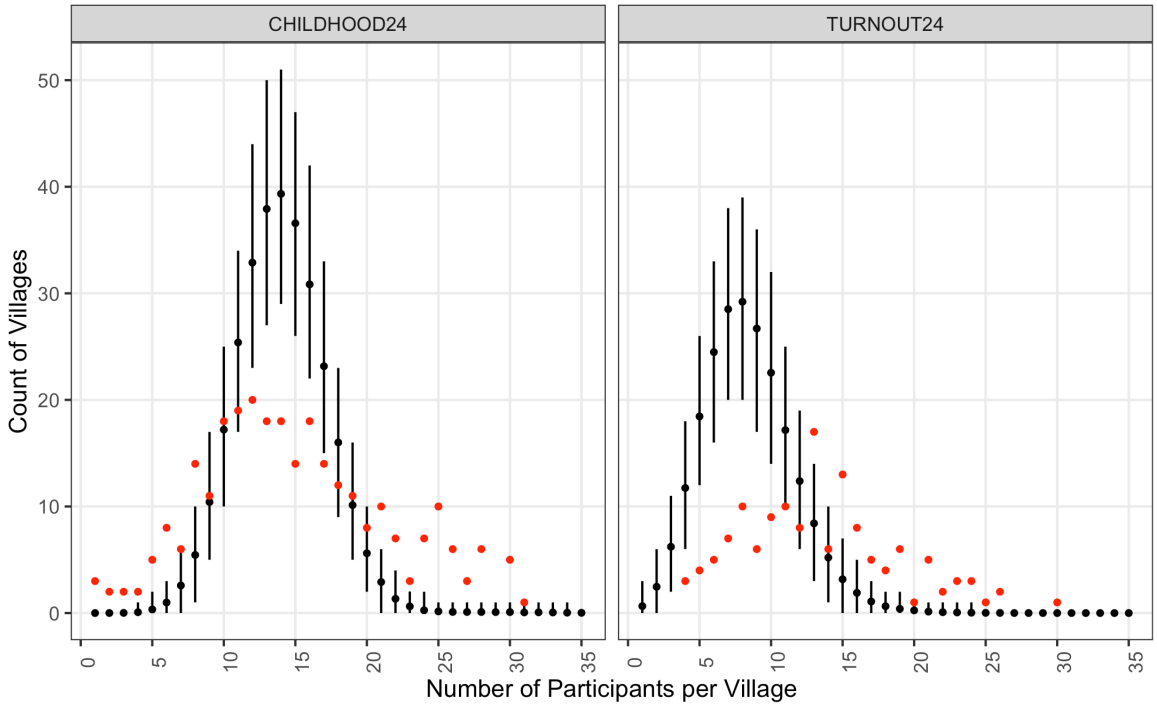
INFO24 also claims that all participants in their study were *randomly* sampled from the earlier participant samples.[3] To assess this claim, we drew repeated random samples of the same size as in INFO24 from the source samples and calculated bootstrap confidence intervals for the number of villages with exactly K respondents in the sample, for all possible values of K.

The comparison in Figure 1 reveals that the observed number of villages with different counts of participants in INFO24 frequently falls outside of the confidence intervals. The CHILDHOOD24 sample drawn in INFO24 has too few small villages and too many large relative to the reference distribution for random samples of this size. The TURNOUT24 sample has an excess of both small and large villages

[3] *"To randomly select households for the campaign, we obtained a list of households with mobile phone numbers that were previously surveyed by the two local organizations, GDRI and DPRN, in Bangladesh and India, respectively" (p. 641).*

and too few in the mid-range. This implies that the randomization procedure implemented in INFO24 was flawed or involved some form of conditioning or sample restriction not documented in the paper or supporting materials.

Figure 1: Distribution of village participant counts for reused samples



Note: The red dots are the observed sample and black dots and ranges are bootstrapped means and 95 percent confidence intervals.

2.3 Undisclosed sampling criteria negate study motivation

The central hypothesis examined by INFO24 is that text-based health information is insufficient in resource-poor rural communities in Bangladesh and India due to widespread illiteracy, and this hypothesis underpins the research design: there are no untreated controls, and the effectiveness of voice calls is only measured *relative to* the text-only group. The rationale for this hypothesis was that:

> *"... many people struggle to understand text messages due to high illiteracy among adults in rural areas (Saleh, 2020), which often makes text messages an ineffective method of communication among rural people. Therefore, to disseminate accurate, reliable information to these people on how to stay healthy and keep safe during the pandemic [...] we*

carried out two over-the-phone campaigns in Bangladesh and India in between early April and mid-May, 2020" (p. 640).

The illiteracy of the sample is also highlighted in the concluding section:

> *"A key lesson from our findings is the importance of targeted health communications during health crises in developing countries. [...] text and video messages [...] might not be as effective in hard-to-reach rural communities in developing countries. The reason being that illiteracy [etc.] can be strong barriers in communicating important health information to improve health literacy and choices of the poor" (p. 653).*

Testing this hypothesis in a way that has external validity for rural populations in these countries requires a sample that is representative of the general rural population regarding literacy rates. This is not the case: in fact, the majority of the participants were drawn from an earlier study sample that was *explicitly selected on* being literate:

> *"... we [...] focused on literate and married individuals between the ages of 20 and 55 years. We focused on literate individuals to ensure that they could read and understand our treatment messages (e.g., the leaflets)" (TURNOUT24, p. 9).*

Background variables lend further evidence to this difference in selection criteria for the two samples: average years of schooling is about 2.5 years higher for respondents from TURNOUT24 as compared to respondents from CHILDHOOD24 (9.1 years versus 6.6 years). Also, just 1.4% of those from TURNOUT24 have less than six years of schooling, as compared to 35.6% in CHILDHOOD24 sample.[4]

This undisclosed sampling restriction on the TURNOUT24 sample allows for a direct test of the hypothesis that text messages are ineffective *due to* widespread

[4]Approximately half of the respondents taken from TURNOUT24 were respondents also in the TURNOUT24 study, while the other half consist of their spouses. The difference in educational attainment compared to the CHILDHOOD24 sample is however quite similar across these two sub-groups. On average, the *respondents* in TURNOUT24 have 9.3 years of schooling, while their *spouses* have an average of 9.0 years.

illiteracy: since the TURNOUT24 sample are selected on literacy, they should be markedly better informed and show higher compliance with preventive measures under the "SMS only" treatment relative to the participants sampled from the CHILDHOOD24 study that does not select on literacy. This is not the case: the TURNOUT24 sample participants in the "SMS only" treatment group are significantly *less* informed ($p < 0.01$) and have identical compliance rates to the CHILDHOOD24 participants in the same treatment group. Compared to the non-literate CHILDHOOD24 participants, the literate TURNOUT24 sample also shows *significantly larger treatment effects* from receiving the phone calls (all interaction effects positive with $p<0.01$). This is summarized in Table 2.

Table 2: Differential treatment effects on COVID-19 awareness and compliance

	Awareness	Compliance
	(1)	(2)
TURNOUT24 sample	-0.315***	-0.036
	(0.081)	(0.062)
Call Only	1.114***	0.807***
	(0.094)	(0.065)
Call+SMS	1.881***	1.367***
	(0.086)	(0.059)
(Call Only)×(TURNOUT24 sample)	0.666***	0.294***
	(0.111)	(0.083)
(Call+SMS)×(TURNOUT24 sample)	0.314***	0.252***
	(0.100)	(0.073)
Observations	5840	5840
R^2	0.453	0.429

Note: The outcomes in the regressions are the awareness index (Column 1) and the compliance index (Column 2), both extracted from the first endline survey of INFO24. The variable "TURNOUT24 sample" denotes whether the observation was present in the TURNOUT24 sample. Both regressions include the controls used in INFO24. Robust standard errors clustered at the village level are in parentheses.
* p<0.10, ** p<0.05, *** p<0.01

2.4 Other issues related to the reuse of data

A few other issues related to the reuse of the TURNOUT24 and the CHILDHOOD24 samples are worth mentioning. Firstly, comparing the information provided in the

different studies, we found that CHILDHOOD24 and TURNOUT24 gathered their survey data in 2017[5] and 2018 (TURNOUT24, p.11), and not in 2019 as claimed in INFO24.[6]

Further, a closer comparison of the different data samples revealed errors in the data reused from TURNOUT24: all female respondents are incorrectly assigned the age of their husband in INFO24.

Finally, according to Table B4 in INFO24, only 7 of the 11 control variables used in the main specification were derived from the previously collected data. Instead, the controls for occupation, worries about household health and finances, and food insecurity were constructed from the new survey collected after the information campaign. This raises concerns because these variables may have been influenced by the treatment, making them "bad controls".[7]

3 Unreported re-treatment of participants and relation to Vlassopoulos et al. (2024)

The reuse of the TURNOUT24 and CHILDHOOD24 samples was not restricted to INFO24. Over the course of our work on INFO24 we discovered that a subsample of the same participants – about one quarter of the respondents (1,583) in the Bangladesh sample – also appears in the data of Vlassopoulos et al. (2024), hereafter referred to as COUNSELING24.[8] In this section we focus solely on the participants that appear in both studies.

COUNSELING24 was an analysis of a randomized field experiment of phone-based counseling and COVID-19 information targeting Bangladeshi women in rural

[5]Note that it is unclear from the study timeline in CHILDHOOD24 (p.11) whether the data was collected in 2016 or 2017. The relevant folder in the data repository, however, is labeled 2017.

[6]"*As participating individuals were also surveyed by the two NGOs in 2019, we consider this 2019 survey data our baseline and use it to check our sample characteristics and balance between treatment arms.*" (INFO24, p. 639).

[7]It also seems likely that it would have been possible to construct conceptually similar variables from the data of CHILDHOOD24 and TURNOUT24. Using occupation as reported in CHILDHOOD24 and TURNOUT24, we note imbalances across treatment categories. For instance, respondents in the "Call+SMS" group are 6.3% (p=0.016) less likely to be farmers as compared to respondents in the "SMS only" group.

[8]As above, we match respondents across datasets using RECORD_ID and compare their baseline characteristics to confirm that the IDs refer to the same individuals.

villages during the pandemic. As INFO24, this study also drew participants exclusively from the samples of TURNOUT24 and CHILDHOOD24. The intervention consisted of four counseling sessions spread across a three-month period and totaling about two hours of sessions per participant. The intervention was reported to have large and statistically significant effects across a wide range of indicators, including stress, depression, and compliance with COVID-19 health guidelines.

INFO24 and COUNSELING24 used overlapping study samples in two concurrently running and closely related field experiments. This should have been disclosed. We can be sure that this was known to the authors of INFO24: Four researchers are listed as authors on both papers and GDRI was the implementing partner for both studies.

The joint timeline of the two projects, shown in Figure 2, also reveals that the interventions and the data collection were tightly interwoven. This should have been disclosed in both papers. It should also have been noted in the handbook chapter on field experiments one of the authors later wrote, where the *lack of* sample overlap between INFO24 and a third study is highlighted as essential to validly estimate clean treatment effects in two concurrently running studies with shared implementation infrastructure.[9]

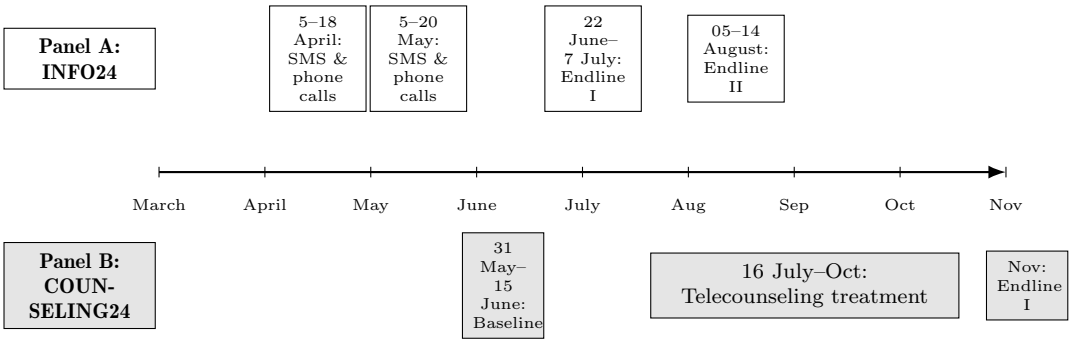
The key points to note are the following: the INFO24 intervention had ended *before* the baseline data of COUNSELING24 were collected, and the COUNSELING24 intervention had been ongoing for about a month *before* INFO24 collected their second endline data.

The two samples are also related in other ways: the probability that a participant from TURNOUT24 or CHILDHOOD24 is present in the INFO24 sample increases from 25% to 71% and from 35% to 67%, respectively, for participants also sampled and present in COUNSELING24. If the two studies were randomized independently

[9]"For example, during the Covid-19 pandemic, we provided a health information campaign among people to make them aware of basic health protocol on Covid-19 (Siddique et al. 2022), while also offering children remote learning opportunities (Wang et al. 2023), with the support of an NGO partner. Since both these interventions were run in the same geographic area but among different households, using the same NGO partner, this approach avoided contamination biases, and reduced the overall cost of running surveys and interventions" (Islam 2024, p.15).

in the way claimed, the probability of being in INFO24 should be the same regardless of whether a participant was also drawn to the COUNSELING24 sample.[10] Also, when we look at the participants included in the second endline survey of INFO24, described in the paper as a *"a second survey on roughly 1,600 randomly selected women participants from Bangladesh"* (p. 639), we find that this sample perfectly coincides with the sample of respondents that appear in both papers.

Figure 2: Joint timeline of INFO24 and COUNSELING24



Note: The upper part of the illustration reproduces the timeline presented in INFO24 (Figure 1, p. 641). The bottom part presents the most precise dates we were able to find for COUNSELING24. The exact dates for the baseline survey and for the start of the COUNSELING24 intervention are taken from the pre-registration report (https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=380128). These dates are consistent with the timeline presented in COUNSELING24 (Figure 1, p.430), although the timeline does not provide exact dates.

Table 3: Treatment groups and re-treatment in COUNSELING24

	Treatment status in INFO24		
Treatment status in COUNSELING24	SMS only	Call only	SMS & Call
Control	246	206	269
Therapy treatment	332	256	274

Note: The table illustrates the number of observations in the second endline of INFO24, by treatment arms, and how they relate to the therapy treatment in COUNSELING24.

3.1 Exposure to a new treatment

The COUNSELING24 treatment sessions were spaced out to occur every *"two to three weeks"* (COUNSELING24, p. 430). While we do not know exactly how many

[10]Benign explanations are possible, but would raise additional issues. E.g., imagine that 65% of the TURNOUT24 sample had changed their phone numbers and could not be reached. The INFO24 sample would now cover about 71% of those remaining, consistent with the probability of seeing TURNOUT24 participants in INFO24 conditional on inclusion in CHILDHOOD24. Having the same phone number as before, however, would now be unusual (the majority - 65% - would not), and this would then raise additional questions about selective attrition in the participant pool and the generalizability of the estimates.

of the four counseling sessions participants received before the second INFO24 endline, we can conservatively assume 1-2 sessions. Even though the treatment assignment in COUNSELING24 was unrelated to that in INFO24 (see Table 3), this is still important because the counseling sessions were extensive and covered similar COVID-19 information as the INFO24 interventions. But whereas the information voice calls assessed in INFO24 were said to last *"10-15 minutes"* (INFO24, p. 641), the COUNSELING24 sessions lasted *"roughly 30 minutes"* (COUNSELING24, p. 431). Also, the first session was centered almost entirely on tips for avoiding a COVID-19 infection. The scripts in COUNSELING24's Online Appendix D show talking points such as:

- *One way to stay safe from coronavirus is to stay at home. For emergencies, we all go outside the home, but the less you go outside the home is better.*
- *We should avoid going to places where a lot of people gathers such as social ceremonies, general meeting with people.*
- *If we go outside the home, we should keep a distance of a minimum of 1.5 meters or three times your arm's length.*
- *We should wash our hands with soap and water for at least 20 seconds after coming home.*
- *We should cover our mouth with a handkerchief or with the fold of our elbow while coughing or sneezing.*

This information is essentially the same as that covered by the compliance items in the second endline of INFO24, which means that the INFO24 participants in the COUNSELING24 treament group would get a comprehensive information session directly tailored to the compliance questions they would be asked a short time later. As a result, roughly half of the INFO24 participants in the "SMS only" group actually received an extensive voice-based information intervention prior to answering their second endline, while roughly half of the INFO24 participans in the

"Call only" and "Call+SMS' groups had an extra dose of information. None of this was disclosed or discussed in the paper.

Surprisingly, while the brief INFO24 calls had very large and substantive effects on compliance with COVID-19 prevention guidelines, the longer and more comprehensive COVID-information sessions in COUNSELING24 had *no impact* on reported compliance with COVID-19 guidelines as measured in the INFO24 August endline. This is shown in the first column of Table 4, which presents treatment effects of the different interventions. The reference group in the regression consists of those assigned to "SMS only" in INFO24 and the control group (no counseling sessions) in COUNSELING24. To ease interpretation, scores are standardized using the standard deviation in this reference group, so the coefficients are in standard deviation units. As can be seen, the COUNSELING24 treatment had no effect on compliance for any of the treatment groups from INFO24.

The second column in Table 4 is the same analysis, but using the COVID-19 compliance items from the COUNSELING24 November endline, gathered three months after the INFO24 August endline.[11] In this data, the COUNSELING24 treatment has a large effect on compliance, raising scores by 1.357 SD ($p<0.001$), while the INFO24 treatment effects have largely disappeared. The effect of receiving a INFO24 call (but no COUNSELING24 treatment) has fallen from 1.781 SD ($p<0.001$) to 0.254 SD ($p<0.05$), while the initially stronger "Call+SMS" effect has been erased – going from 2.808 SD ($p<0.001$) to 0.086 ($p=0.421$).

In sum, although both studies have treatments emphasizing the same COVID-19 guidelines and use outcome measures that are closely related, the effects of the INFO24 treatment are not discernible in the COUNSELING24 November endline while the effects of the COUNSELING24 treatment are not discernible in the INFO24 August endline.

[11] The compliance index in COUNSELING24 is measured on a different scale, but we standardize the index to the same reference group to facilitate the comparison across specifications.

Table 4: Treatment effects on compliance to COVID-19 guidelines

	INFO24 (Aug-20) (1)	COUNSELING24 (Nov-20) (2)
Call Only	1.781***	0.254**
	(0.113)	(0.123)
Call+SMS	2.808***	0.086
	(0.091)	(0.107)
Therapy Treatment	0.114	1.357***
	(0.082)	(0.106)
(Call Only)×(Counseling Treatment)	0.182	-0.360**
	(0.143)	(0.151)
(Call+SMS)×(Counseling Treatment)	-0.122	-0.135
	(0.125)	(0.147)
Observations	1582	1463
R^2	0.576	0.330

Note: The table displays treatment effects on compliance to COVID-19 health guidelines. The outcome in Column 1 is the compliance index of INFO24, as measured in the second endline collected in August 2020. Column 2 uses as outcome the compliance index of COUNSELING24, as measured in their first endline collected in November 2020. Both outcome variables are re-scaled to the group of respondents in the "Call only" group from INFO24 and the control group from COUNSELING24. Both regressions include the same controls as in Table 4, Column 4 of INFO24. Robust standard errors clustered at the village level are in parentheses.
* p<0.10, ** p<0.05, *** p<0.01

3.2 Comparing COVID-19 compliance measures across studies

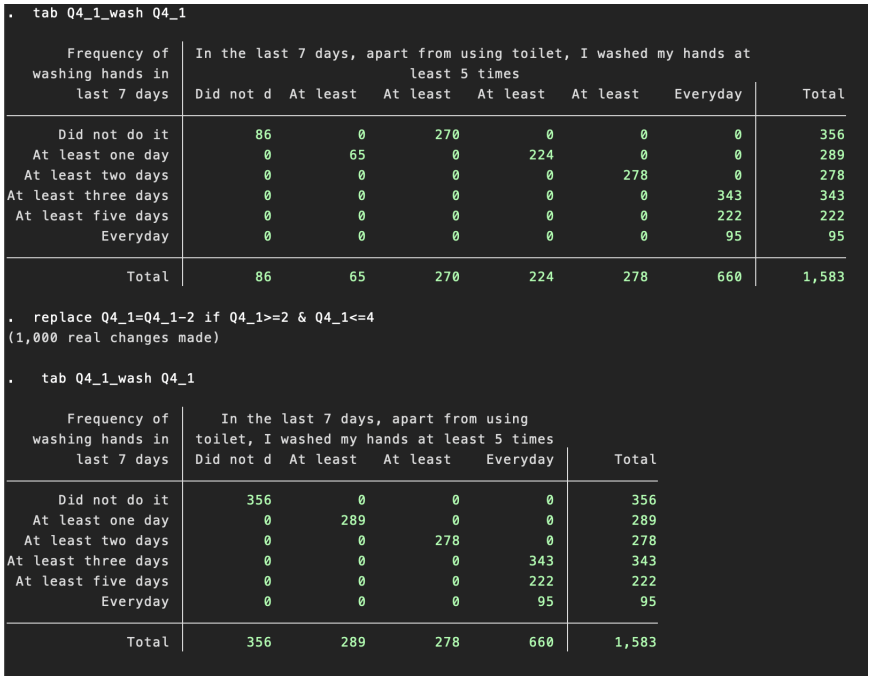
The baseline data of COUNSELING24 and the first endline of INFO24 contain essentially the same six items to gauge compliance with COVID-19 guidelines, all measured on a six point scale.[12] The responses to these questions should reasonably be similar but not identical, as they were collected about a month apart: May 31-June 15 versus June 22-July 7.

Comparing responses on the COUNSELING24 items and the analogous INFO24 items for the overlapping sample we found a systematic and statistically implausible joint distribution where responses in one data source map neatly onto the responses in the other. Visually, the impression is that all the observations with middle range values in COUNSELING24 shifted two levels up on the response scale relative to their response in INFO24. This pattern is present for every single one of these items. For instance, Figure 3 shows this for the hand washing item: *every* respondent answering *"At least 2 days"* in COUNSELING24 answered *"Didn't do it"* in INFO24; *every* respondent that answered *"At least 3 days"* in COUNSELING24 answered *"At least 1 day"* in INFO24; and so forth. As is shown in the bottom part of the figure, a simple recoding aligns all response values except for one. Analogous figures for the remaining items are in the appendix (Figures A1 to A5).

This pattern between items reputedly from two separate surveys, collected one month apart, is systematic and consistent to an extent that cannot be rationalized as plausibly correct data.

[12] COUNSELING24 has an additional item not mirrored in INFO24 that asks respondents how often they attended work last week. It is worth noting that this is the only compliance variable where we find a *negative* treatment effect of the INFO24 treatment (coefficient of -0.059 for the "Call+SMS" group, p=0.060).

Figure 3: Frequency of hand washing in INFO24 and COUNSELING24.



```
. tab Q4_1_wash Q4_1

      Frequency of |  In the last 7 days, apart from using toilet, I washed my hands at
     washing hands in |                          least 5 times
       last 7 days | Did not d  At least   At least   At least   At least    Everyday |     Total
-------------------+-----------------------------------------------------------------+----------
      Did not do it |        86         0       270         0         0         0 |       356
    At least one day |         0        65         0       224         0         0 |       289
   At least two days |         0         0         0         0       278         0 |       278
 At least three days |         0         0         0         0         0       343 |       343
  At least five days |         0         0         0         0         0       222 |       222
            Everyday |         0         0         0         0         0        95 |        95
-------------------+-----------------------------------------------------------------+----------
               Total |        86        65       270       224       278       660 |     1,583

. replace Q4_1=Q4_1-2 if Q4_1>=2 & Q4_1<=4
(1,000 real changes made)

. tab Q4_1_wash Q4_1

      Frequency of |    In the last 7 days, apart from using
     washing hands in |  toilet, I washed my hands at least 5 times
       last 7 days | Did not d  At least   At least    Everyday |     Total
-------------------+---------------------------------------------+----------
      Did not do it |       356         0         0         0 |       356
    At least one day |         0       289         0         0 |       289
   At least two days |         0         0       278         0 |       278
 At least three days |         0         0         0       343 |       343
  At least five days |         0         0         0       222 |       222
            Everyday |         0         0         0        95 |        95
-------------------+---------------------------------------------+----------
               Total |       356       289       278       660 |     1,583
```

Note: The screenshot displays the relationship between the "washing hands" variables from the first endline of INFO24 (Q4_1_wash) and the baseline of COUNSELING24 (Q4_1). The bottom part of the screenshot shows that a simple recoding aligns most response values with the exception of the highest possible response in COUNSELING24.

3.3 No short-term treatment effects on COVID-19 perceptions

The timing of the two studies, where the baseline of COUNSELING24 was collected immediately after the INFO24 intervention, also gives us the opportunity to assess the very short-term effects of the information campaign.

In Table 5 we display estimated treatment effects for the sample of respondents that appear in both samples. To facilitate the comparison across outcomes, we re-scale all outcomes using the standard deviation of "SMS only" group, such that the coefficients are presented in standard deviation units.

We start with compliance to COVID-19 guidelines. The outcome for each sample is the sum of the discretized variables using the INFO24 thresholds. As we already have shown that the compliance items in the two surveys are systematically related (see Section 3.2), it is not surprising that we find strong treatment effects in both samples. Still, the estimated effects in the COUNSELING24 data are about half the size of those in the INFO24 data.

We next sum the four binary COVID-19 awareness items that appear in both samples and estimate treatment effects on this new composite measure. As can be seen from Columns 3 and 4, we find that the immediate impact of the information campaign was larger than the impact one month later: 2.60 SD versus 1.44 SD for the "Call+SMS" group, and 1.65 SD versus 0.94 SD for the "Call only" group. These differences cannot be explained by a general erosion of knowledge as we move further away in time from the intervention. As we show in the appendix (Table A1), the level of awareness is systematically higher in INFO24 than in COUNSELING24. For instance, only 12% of the respondents in the "SMS only" group knew about the importance of hand washing in the COUNSELING24 baseline (just after the intervention of INFO24), while this had increased to 75% in the INFO24 endline.

Finally, the baseline data from COUNSELING24 includes 16 items capturing COVID-19 "perceptions". These are conceptually related to COVID-19 "awareness", allowing us to assess treatment effects on a new set of outcomes. In COUNSELING24, these items are summed to yield an overall measure of perceptions, with each individual item having binary outcomes with 1 indicating the "correct" response (see Figure 4 for a full list of items). Column 5 of Table 5 displays treatment effects on this measure. In stark difference to the estimated effects on COVID-19 awareness, we find very little impact on COVID-19 perceptions: the effect is effectively zero for the "Call only" group (coefficient of -0.019, p=0.833) and just 6% of the effect on awareness for the "Call+SMS" (coefficent of 0.153, p=0.048).

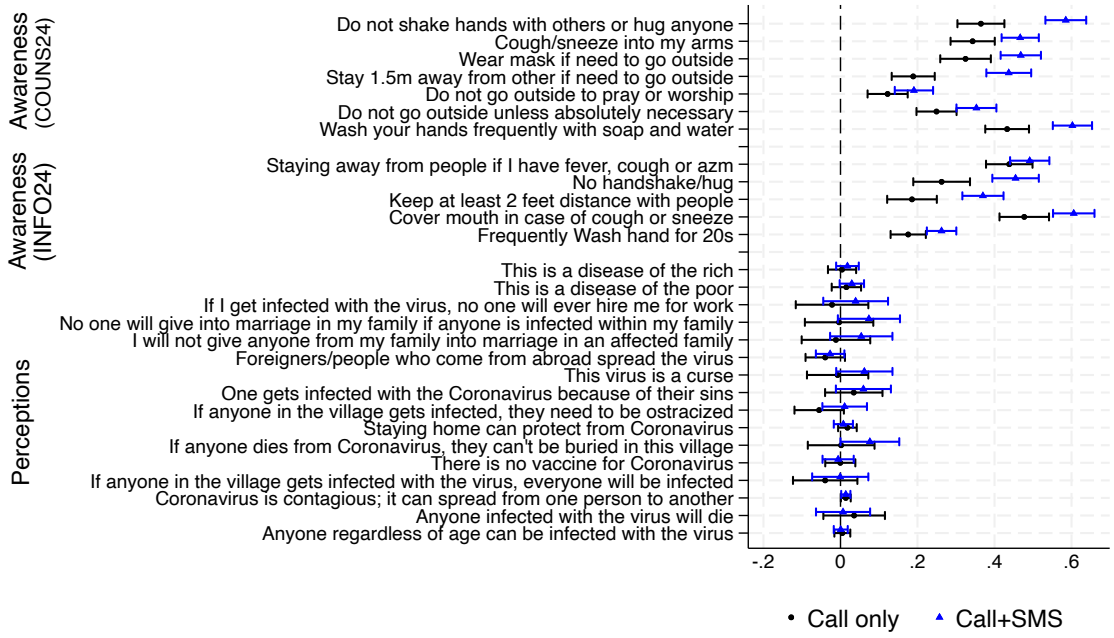
To examine this discrepancy more closely, we estimated the treatment effects on each awareness and perception item using separate regressions (Figure 4). Treatment effects are large and statistically significant at the 1% level for every single COVID-19 awareness item, whereas only 1 of 16 perception items is statistically significant at the 5% level.

Table 5: Treatment effects on COVID-19 awareness and perceptions

	Compliance		**Awareness**		**Perceptions**
	COUNS24	INFO24	COUNS24	INFO24	COUNS24
Dates :	(31.5-15.6)	(22.6-7.7)	(31.5-15.6)	(22.6-7.7)	(31.5-15.6)
	(1)	(2)	(3)	(4)	(5)
Call Only	0.691***	1.244***	1.652***	0.936***	-0.019
	(0.058)	(0.082)	(0.081)	(0.066)	(0.091)
Call+SMS	0.891***	1.884***	2.598***	1.438***	0.153**
	(0.051)	(0.064)	(0.075)	(0.059)	(0.077)
Observations	1421	1421	1421	1421	1421
R^2	0.292	0.457	0.545	0.429	0.141

Note: The outcome in Column 1 is the COVID-19 perception index of COUNSELING24, as measured in their baseline in May-June 2020. Column 2 uses as outcome the COVID-19 awareness index of INFO24, as measured in June-July 2020. To facilitate the comparison between the two outcomes, we have re-scaled both outcomes, such that the coefficients represent standard deviations of the "SMS only" group. Both regressions include the controls used in INFO24. Robust standard errors clustered at the village level are in parentheses.
* p<0.10, ** p<0.05, *** p<0.01

Figure 4: Treatment effects on individual COVID-19 awareness and perception indicators



Note: The figure plots estimated treatment effects, relative to the "SMS only" group, on each of the individual COVID-19 awareness and perception indicators. The lines display 95% confidence intervals, based on standard errors clustered at the level of villages. All regressions include the controls used in INFO24. The upper panel shows the awareness indicators from INFO24, while the bottom panel shows the perception indicators from COUNSELING24.

3.4 Comparing other items measured in both studies

The compliance items are not the only ones where we find implausible patterns across the samples of INFO24 and COUNSELING24. The same systematic "two shifts down" pattern also appears for a variable capturing financial worries: item-responses differ across the two data sources, but the differences look consistent and systematic (see Figure A6).[13] Moreover, for an item capturing health worries, we found that *all* 1583 respondents gave identical responses on a 3-point scale in the two surveys (see Figure A7). This is also true for an item capturing household chores from the second endline of INFO24 (collected in August) and the baseline of COUNSELING24 (see Figure A8).

For other items, on the other hand, the joint correlations are unexpectedly low. For instance, the occupation of the household head switches markedly in just one month: 44% of household heads working as farmers in May-June no longer do so in June-July, while only 20% of household heads working as laborers in May-June still do so in June-July.

4 Unlikely data patterns across the two endline surveys

INFO24 uses data from two endline surveys – one collected from everyone in June/July and one collected from a subset of the female participants in August. We use this as an opportunity to examine within-individual correlations in three items with close analogues across the two surveys. These all relate to compliance with COVID-19 precautionary measures: frequency of washing hands, avoiding physical contact, and going out for religious purposes.

All items are answered on a 6-point scale in the raw data file, and transformed to binary variables before analysis – taking the value 1 for the three highest response levels and 0 otherwise. The binary compliance variables are strongly correlated within individuals across time, with a correlation coefficient exceeding 0.5 for all

[13] Note also that the thresholds used to discretize these variables in the two studies differ in a way that perfectly negates this shift, meaning that the binary variables used in the analyses have identical values for every respondent in the two files.

three outcomes. Unusually, however, the correlation for the more information dense 6-leveled variables is substantially lower – though still exceeding 0.3 for all outcomes.

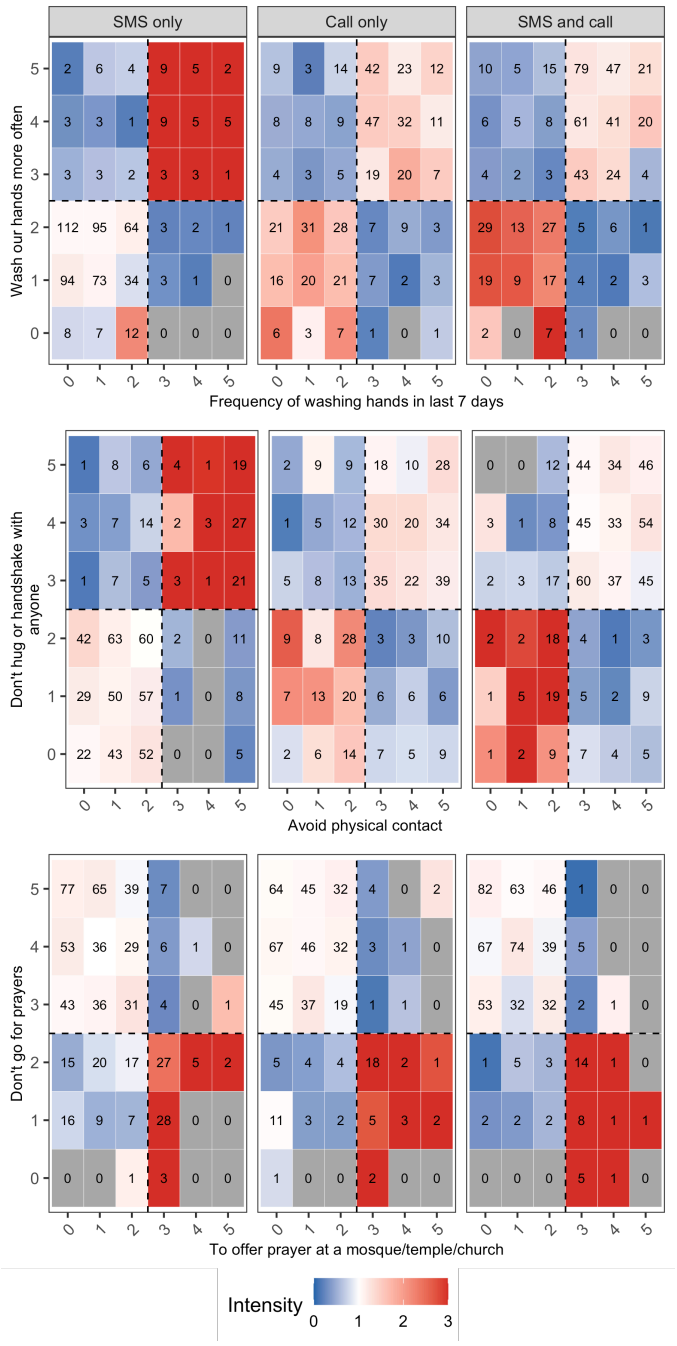
This motivated us to take a closer look at the joint distribution across the two endlines for each variable – separately by treatment group (Figure 5). The color gradients within each plot display to what extent cell counts are over-represented (red) or under-represented (blue) relative to expected counts when observed marginal distributions are taken as given and we assume independence between the two measurements. The dotted black lines demarcate the thresholds used in INFO24 to make both variables binary, and separates the top three from the bottom three levels of each variable. Observed counts are displayed within each cell.

The figures explain why we see higher correlations for the less granular binary version of the variables than we do for the six-level variables in their raw form: the within-individual correlation is primarily at the level of the four quadrants defined by the thresholds – while there is essentially no systematic patterns to speak of within the 3x3 grid of cells within any of these quadrants. As a result, we also see extreme discontinuities in the counts around the thresholds where high-count and low-count quadrants meet. As an example, consider the "SMS only" group in the first panel row comparing "hand washing" from the first and second endline survey. In the first three columns the counts around the horizontal threshold jump from 112 to 3, from 95 to 3 and from 64 to 2. Comparing the bottom three rows on both sides of the vertical threshold we jump from 64 to 3, from 34 to 3 and from 12 to 0.

Another way of showing the same pattern is to pool the treatment groups and plot the average response in the second endline by the responses in the first endline. Figure 6 clearly shows how the average score jumps as we pass the threshold. Learning what side of the threshold you were at in the first endline tells us a lot about your score on the same item in the second endline – learning whether you were close to or far from the threshold tells us more or less nothing. Figure A9 shows a similar pattern in average responses in the first endline, conditioning on the response value in the second endline.

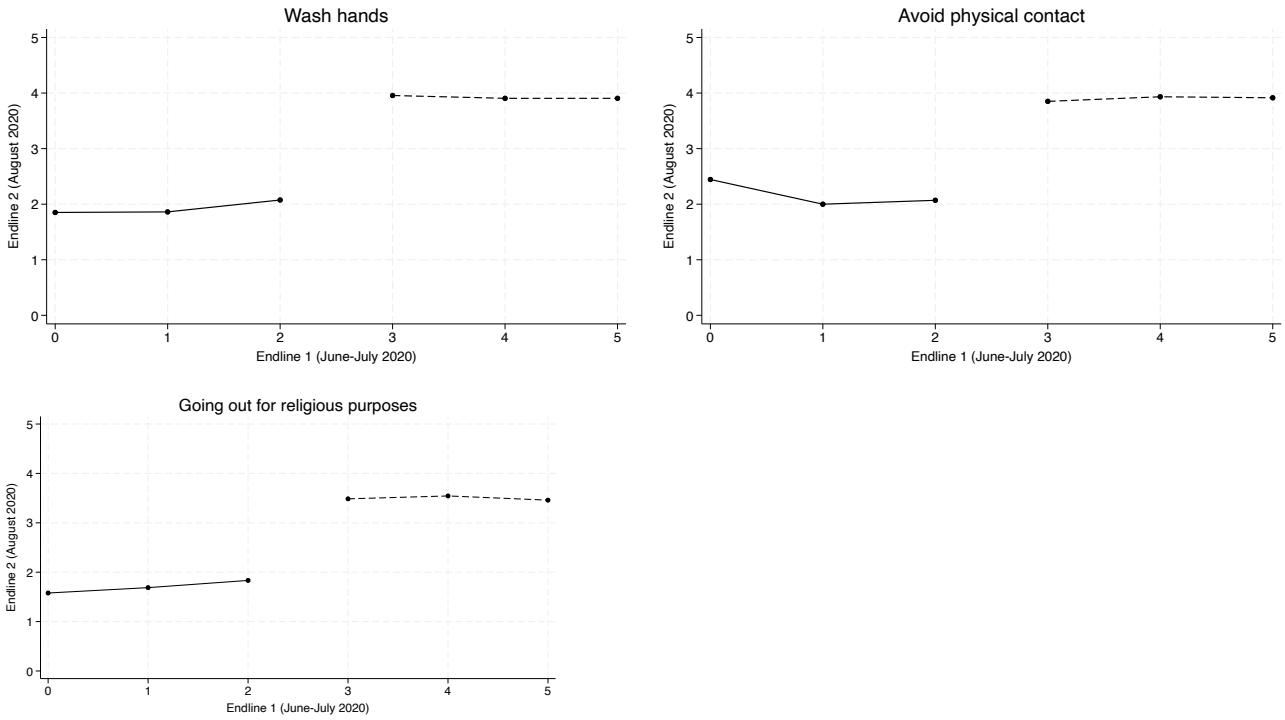
These discontinuities are why within-individual correlations increase when the variables are discretized. The systematic nature of this pattern across multiple survey items is concerning and hard to reconcile with typical survey response behavior.

Figure 5: Comparison of COVID-19 compliances outcomes across the two endline surveys



Note: The figures show the joint distribution of three COVID19 compliance items for the two endline surveys in Bangladesh. The horizontal axis displays values from the first endline survey, while the vertical axis displays values from the second endline survey. The shading of the cells shows the ratio of the observed count to the expected count under fixed marginal distributions and an assumption of item independence. Empty cells are grey.

Figure 6: Discontinuous responses across the two endline surveys



Note: The figures display average responses to three questions related to compliance to COVID-19 guidelines in the second endline survey, calculated separately for each response value in the first endline survey. For example, the first dot in the upper left figure displays the average response to the hand washing question in the second endline for the respondents that answered *0="Didn't do it"* in the first endline survey. The break in the figures correspond to the cutoff values used to discretize the outcome variables in the analysis of INFO24. See Figure A9 for a similar plot when conditioning on the responses in the second endline survey.

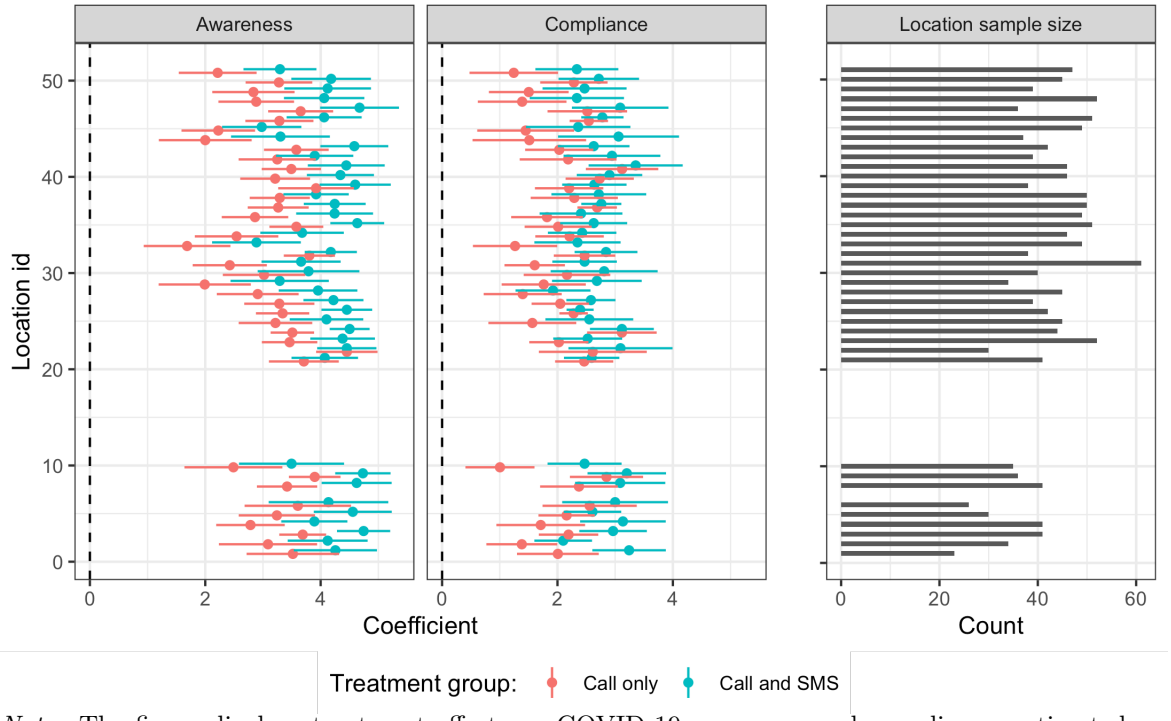
5 Statistically improbable data patterns in the India data

Opportunities for assessing the data from the Indian sample are more sparse. The study uses the same awareness and compliance items as the Bangladeshi study, but has a smaller sample size and fewer variables available in the data sets. Although the India sample, like the Bangladesh one, was recruited from previously surveyed individuals, we have not been able to find available data repositories from other projects with overlapping samples.

The Indian sample is about a quarter of the sample size from Bangladesh (n=1680 vs. n=6722), and estimated treatment effects are all statistically significant ($p<0.001$) with magnitudes substantially exceeding those seen in Bangladesh. Where the "Call+SMS" treatment increases awareness in Bangladesh by an esti-

mated 2.1 (p<0.001), for instance, the same treatment in India raises awareness by 3.4 (p<0.001). The effects are also remarkably consistent: If we estimate the treatment effects independently within each location, operating with samples sizes in the range of 23 to 61, we nonetheless get statistically significant treatment effects with p-values below 0.01 in *every single village* (Figure 7).

Figure 7: Village-wise treatment effects in India



Note: The figure displays treatment effects on COVID-19 awareness and compliance, estimated separately for each of the 40 Indian villages in the INFO24 sample. Outcome variables are standardized to the "SMS only" group. The lines display 95% confidence intervals for each village. The rightmost panel displays the number of observations in each village.

5.1 Irregular data patterns

To illustrate the magnitude of the above effect sizes, we tabulate the number of correct awareness responses for each individual and compare the distribution across treatment groups in the two countries (Table 6). This shows that *no one* in the Indian "SMS only" group got all items correct, while 84.4% in the "Call+SMS" treatment got everything right. This is an increase of 84.4 percentage points (!) and is twice the size of the increase seen for Bangladesh.

These numbers indicate a very low within-group heterogeneity in the outcomes. Another way of illustrating this is to compare the R^2 from regressions of the main

outcomes on either the treatment indicators only or the background controls only. When including all control variables at the same time (excluding the region fixed effects), they explain less than 13% of the variation in both the awareness and the compliance indices. In contrast, the treatment indicators alone have a very large explanatory power, explaining 61% of the variation in compliance and 79% of the variation in awareness.[14]

Table 6 also reveals a puzzling anomaly. While the Indian "SMS only" group does very poorly, with 73.8% of the respondents managing answering only a single correct response, we do not see *any observations* that fail all five items. This does not reflect the inclusion of an excessively simple item: each individual item is failed by the majority of respondents within the same treatment group when the items are assessed separately. Assuming that the items are statistically independent, we can take the product of the individual item failure rates to calculate the probability that a respondent in the Indian "SMS only" group would fail all the items as $0.544 * 0.829 * 0.548 * 0.781 * 0.848 = 0.163$. With 561 participants, this gives a probability of $(1 - 0.163)^{561}$ of failing all items, which is effectively zero, indicating that it is almost certain that at least one participant will fail all the items.

[14]We find a similar but less extreme pattern in the Bangladesh sample: the control variables jointly explain less than 3% of the variation in awareness and less than 17% of the variation in compliance, whereas the treatment indicators alone explain 41% and 26% respectively.

Table 6: Frequency of observations with different number of correct COVID-19 awareness responses

Number of correct responses	**Treatment status** SMS only	Call only	SMS+Call	**Sum**
Panel A: India				
0	0% (0)	0% (0)	0% (0)	0
1	73.8% (414)	0% (0)	0% (0)	414
2	10.3% (58)	4.8% (29)	0.2% (1)	88
3	12.8% (72)	20.1% (121)	1.7% (9)	202
4	3.0% (17)	42.1% (253)	13.7% (71)	341
5	0% (0)	32.9% (198)	84.4% (437)	635
Total	100% (561)	100% (601)	100% (518)	1,680
Panel B: Bangladesh				
0	14.5% (342)	0.3% (7)	0.1% (2)	351
1	16.5% (389)	3.2% (65)	1.0% (21)	475
2	28.0% (660)	11.6% (236)	5.0% (104)	1,000
3	30.8% (727)	19.0% (385)	18.4% (385)	1,497
4	6.4% (151)	28.7% (583)	27.0% (565)	1,299
5	3.9% (92)	37.2% (755)	48.5% (1,016)	1,863
Total	100% (2,361)	100% (2,031)	100% (2,093)	6,485

Note: The table displays the percentage of observations (with counts in parentheses) with different number of correct COVID-19 awareness responses, separately by treatment arms. Percentages are calculated within each treatment group.

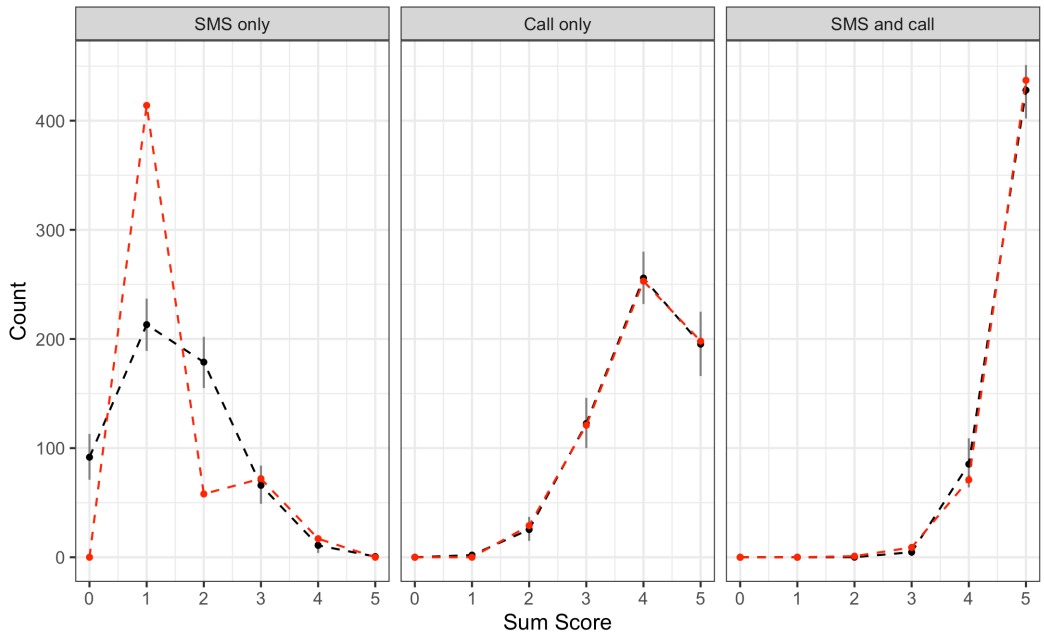
The above patterns prompted us to examine how well the Indian data fits a simple statistical model that assumes that responses are uncorrelated across items. Using this model, we can estimate the success probability of each item from the observed share with correct responses, and then draw multiple synthetic data samples using these estimated probabilities to obtain a reference distribution for the number of correct responses.

The results for the Indian data are shown in Figure 8, with the red dashed line showing the observed distribution while the black lines show the expected distribution with 95% credibility intervals. The model is estimated separately for each treatment group.[15] The first thing to note is that the model fits the data almost

[15]The model was estimated in Stan, a programming language for probabilistic Bayesian models. The estimation takes weak prior distributions for the parameters and updates these in light of the data, returning a set of representative draws from the updated (posterior) distribution. For each of the 4000 posterior draws we generate a synthetic data sample using the posterior draw for the model parameters and counting the sum score distribution in the resulting data set. In this way, the credibility intervals will reflect two sources of uncertainty: the uncertainty with regards to the model parameters, and the uncertainty resulting from sampling variation.

perfectly for the "Call only" and "Call+SMS" groups (implying that the awareness items are indeed uncorrelated). The second thing to note is the much worse fit for the "SMS only" group. We again see the anomaly we identified manually: the implausible lack of any "all failure" observations in the "SMS only" group.[16] In addition, the model finds an excessive number of "1 correct response" observations and too few "2 correct response observations".[17]

Figure 8: COVID-19 awareness in India: Expected vs. observed number of correct responses



Note: The figure shows the observed distribution of the number of correct COVID-19 awareness responses (Sum Score) in red against the expected distribution under item independence in black. The model treats each item response as a bernoulli trial with a success probability estimated from the data. The model is estimated in the Stan language for probabilistic Bayesian models, with the reference distributions generated by drawing synthetic response data from each posterior draw of parameter values.

[16]The manual and model-based analyses reach the same conclusion: The independent item model expects to see 91.2 (95% CI: 72-112) observations with no items correct in the "SMS only" group, which is essentially the same as 561 x 0.163 = 91.4 (treatment group size times the manually calculated probability).

[17]Intuitively, we would expect responses to be correlated across related survey items, but replacing the independence assumption and using a multivariate probit model to allow item correlations finds essentially identical results (see Figure A11). This is different for the Bangladesh data, where only the correlated items model is consistent with the observed distribution (Figure A12). Comparing the estimated item-level correlations across each treatment group x country combination finds low and unsystematic correlations in every Indian treatment group and persistently high across every Bangladesh group (Figure A13).

5.2 Discontinuities in COVID-19 compliance measures

The above patterns motivated us to take a closer look at the underlying item-level data from India. For two of the compliance variables, this uncovered discontinuities in the response counts at the thresholds used to discretize the variables before analysis in INFO24. This is most evident when we compare the "SMS only" group to the "Call+SMS" group.

Figure 9 plots the share of participants with different responses to the "hand washing" and "avoiding physical contact" items. As can be seen, the most frequent response for the "Call+SMS" group is the value just above the threshold, while the most frequent response for the "SMS only" group is the value just below.

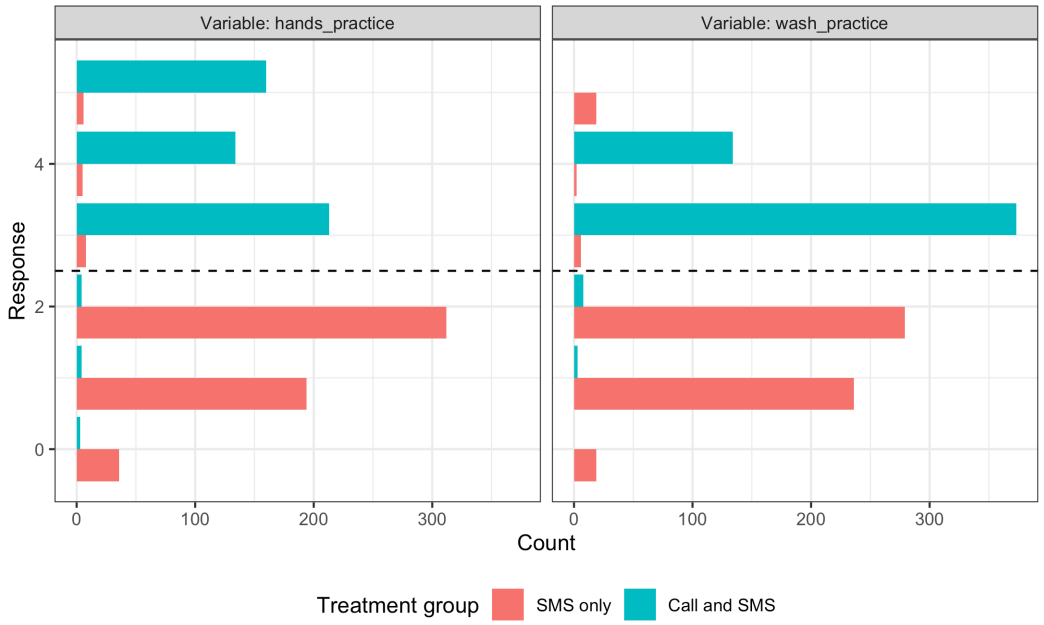
Next, looking at the joint distribution of responses across these two items with the discretizing thresholds in dashed lines, reveals unusually systematic patterns (Figure 10). Examining the top right subquadrant in all panels, there are only three cells that are populated with counts. These three cells are the same in every treatment group, and those responding 3 on the hand washing item are always responding either 3 or 5 *but never* 4 on the physical contact item.[18] This is particularly striking in the "Call+SMS" group where 373 individuals answer 3 on hand washing and split themselves neatly into either response 3 (n=213) or 5 (n=160) on the physical contact item. The empty response category between these, response level 4, is not impossible, as *every single individual* who responded 4 on the hand washing item also responded 4 on the physical contact item. This was so in all groups: "SMS only" group (where the subgroup in question has n=3), "Call only" (n=98) and"Call+SMS" (n=134).

Assessing treatment effects using each item separately finds that the two items discussed above drive the bulk of the compliance effect (Figure 11). This is even more salient running the estimation within each village (Figure A10): for the

[18]It is easily seen that a simple recoding will make responses across the two items identical for all respondents with in the treatment group (subtracting 1 to those with response 1 to the hand washing item, and adding 1 and subtracting 2 to the responses 1 and 5 to the physical contact item respectively).

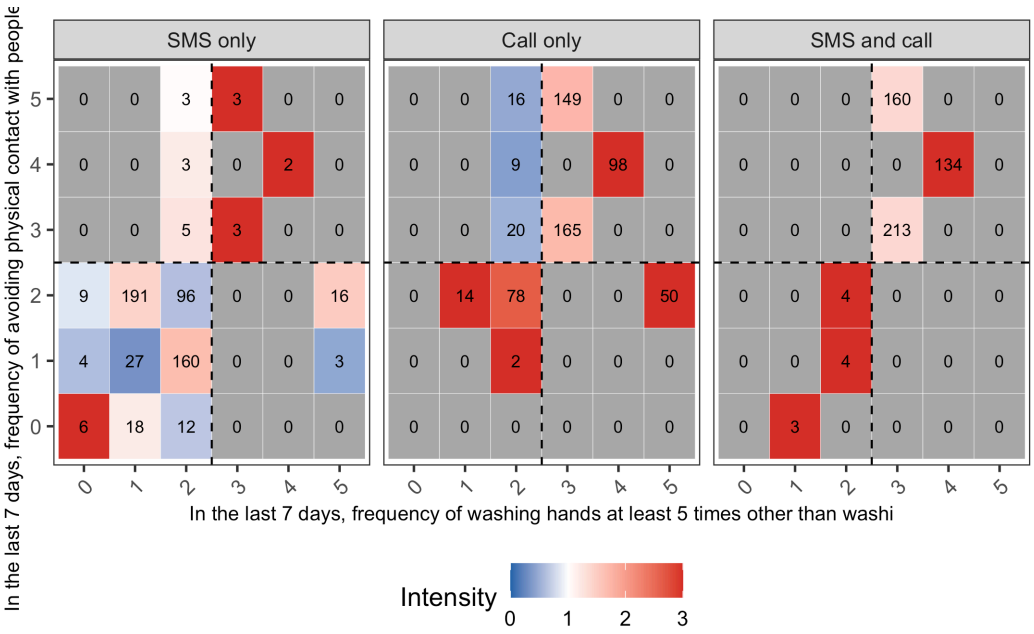
"Call+SMS" group, we find a positive treatment effect for both items, statistically significant at the 5% level, for *all* 40 villages, while we do not find a single positive and significant treatment effect for any of the other items

Figure 9: Discontinuities in compliance measures



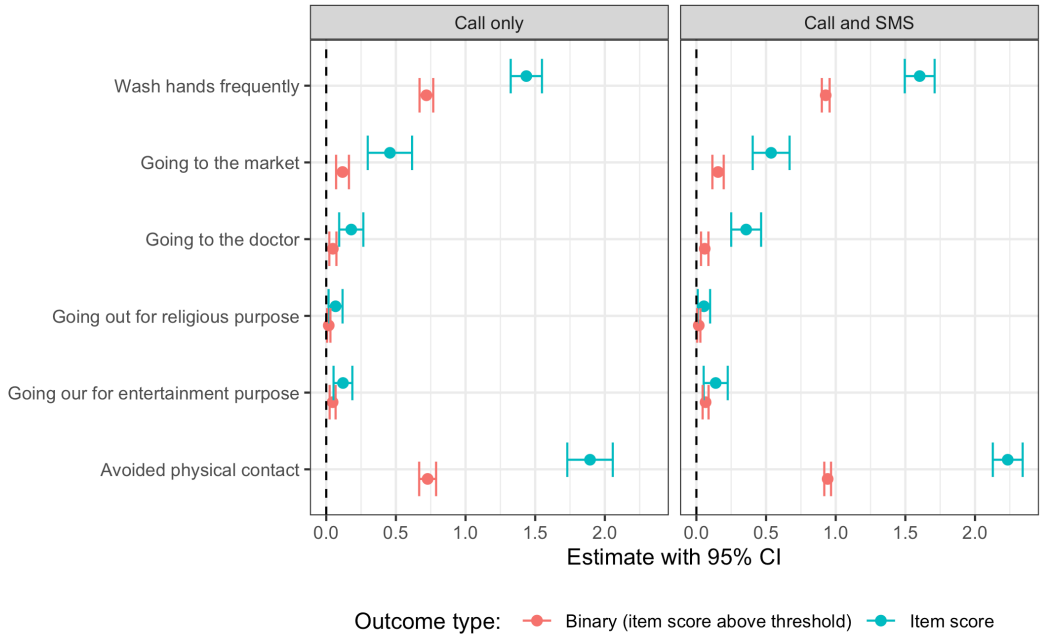
Note: The figure shows the count of different response options for the "SMS only" and "SMS and call" group for two items in the Indian sample. The dashed line shows the threshold used in INFO24 to discretize the variables before analysis.

Figure 10: Joint distribution of "hand washing" and "avoiding physical contact" items



Note: The figure shows the joint distribution of two COVID19 compliance items used in the Indian arm of the INFO24 field experiment. The shading of the cells shows the ratio of the observed count to the expected count under fixed marginal distributions and an assumption of item independence. Empty cells are grey.

Figure 11: Treatment effects for individual compliance measures in India



Note: The plot shows the coefficient and 95% confidence intervals when treatment effects are estimated using each item as an outcome in separate OLS regressions. Each variable is used as the outcome in two ways: in its raw form (taking values from 0 to 5) and discretized.

5.3 Other issues in the India sample

We end by briefly mentioning four additional issues in the sample from India.

First, the data show stark differences across the treatment groups in terms of being "worried about health". Whereas 99.6% of the "SMS only" group and 94.0% of the "SMS+Call" group responded that they were worried about their health, only 22.6% of the respondents in the "Call only" group stated the same. If we take this data as correct, it implies that the text messages caused an increase in health worries of more than 70 percentage points.

Second, there is a non-credible negative association between employment and income. Regressing log of income on a binary variable denoting whether the respondent has employment yields a coefficient of -0.203 ($p<0.01$). The negative relationship is present for male and female respondents alike.[19]

Third, 92% of the female respondents are reportedly working full-time, which does not match other statistics from India.

Fourth, although the income variable in the data reportedly is continuous, it only takes on five values (1000, 3500, 7500, 15000 and 35000).

6 Conclusion

In this comment we have documented a wide ranging set of issues with Siddique et al. (2024a). The paper hypothesizes that text messages are ineffective communication tools in rural Asian populations due to widespread illiteracy, but draws the bulk of their participants from a sample explicitly selected for literacy. The paper acknowledges that it reuses previously surveyed individuals, but does not disclose that all participants in Bangladesh come from two earlier highly selected populations. More importantly, the authors fail to disclose that treatment assignments were inherited from these earlier studies and reused. Simply put: *this was not a randomized study as claimed.* Since the earlier experiments had large effects on

[19] In principle, the negative relationship could be due to the fact that income is measured at the level of household, while employment is measured for the respondent. However, as the relationship holds for both males and females, this explanation is not plausible.

a broad range of outcomes (Ahmed et al. 2024, Guo et al. 2024), the treatment groups were unbalanced at baseline and treatment effects are bundled with earlier treatment effects.

Interpretation is further affected by the undisclosed fact that the experiment was one of two running concurrently with substantial sample overlap and closely related interventions. Linking data across these two experiments at the individual level, we find that individuals present in both showed highly implausible systematic patterns between conceptually related survey items said to be gathered at different points in time. Based on a reconstructed joint timeline of the two projects, both sets of treatments should have discernible effects in both sets of data – but most treatment effects are only large and consistent in the outcome measures from "their own" studies. Comparing participant responses on similar items across the two endlines in Siddique et al. (2024a), we also find highly implausible discontinuities in response counts around the thresholds used to discretize the variables before analysis. Finally, in a smaller data sample from India, we find statistically impossible response patterns – most strikingly a lack of any observations answering all items incorrectly in the least informed treatment group where this would be expected for at least 16% of the sample given the large share of participants answering all but one of the items correctly.

These issues gain additional weight given the magnitude and consistency of the treatment effects reported. The relatively small additional interventions raise compliance by 1-1.5 standard deviations in Bangladesh and 2.2-2.7 standard deviations in India. This is extreme in itself, but made more so by the consistency of the effects: In the Indian arm of the study, where randomization is at the household level, we find statistically significant effects of both treatments on both awareness and compliance outcomes in *every single location*, even though these locations only have sample sizes in the range of 23-61.

To place the effect sizes in context, interventions considered highly effective in psychology will often have effect sizes a third of the size seen in Siddique et al.

(2024a) (Cuijpers et al. 2016). A large educational intervention with effect sizes a tenth of those seen in Siddique et al. (2024a) would be viewed as highly successful (Evans and Yuan 2022). The average t-statistic for treatment effects in Siddique et al. (2024a) is 32.9, and *every single estimate* has a t-value > 12. The results are clear outliers when compared to RCTs across multiple fields,[20] and remain so when we compare them directly to RCTs previously published in top 25 economics journals in the period 2015-2018: the average z-statistics in the other economics studies was 3.97 – with only 1% of papers reporting average z-statistics above 20 (Brodeur et al. 2020).

We were only able to uncover and document these issues following a very detailed analysis of project data, available as a result of recent data sharing requirements in several journals. Even this would have been insufficient, however, had we not discovered that the participant samples overlapped and could be linked to publicly available data from other projects. This revealed additional data anomalies and failures to disclose important facts.

In our judgment, the issues identified in the paper jointly rise to a very problematic level. This judgment reflects more than the simple sum of the noted issues – it also reflects the need for trust in science: we need to trust that the authors gave an accurate description of what was done, we need to trust that they were honest and forthcoming about any weaknesses or limitations in the experimental design, implementation or analysis, and we need to trust that they carefully maintained data integrity from measurement all the way to the final published estimates. The findings of this report break this trust.

[20] Only 0.35% of the 186,822 estimates in the Cochrane Database of Systematic Reviews (Schwab 2020) have a t-statistic above 12, and only 4 of around 35,000 RCT studies in the database have an average t-statistic above 34.

References

Ahmed, F., Hodler, R. and Islam, A.: 2024, Partisan effects of information campaigns in competitive authoritarian elections: Evidence from Bangladesh, *The Economic Journal* **134**(660), 1303–1330.

Al-Dmour, H., Salman, A., Abuhashesh, M., Al-Dmour, R. et al.: 2020, Influence of social media platforms on public health protection against the COVID-19 pandemic via the mediating effects of public health awareness and behavioral changes: Integrated model, *Journal of Medical Internet Research* **22**(8), e19996.

Allcott, H., Boxell, L., Conway, J., Gentzkow, M., Thaler, M. and Yang, D.: 2020, Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic, *Journal of Public Economics* **191**, 104254.

Brodeur, A., Cook, N. and Heyes, A.: 2020, Methods matter: P-hacking and publication bias in causal analysis in economics, *American Economic Review* **110**(11), 3634–3660.

Brodeur, A., Mikola, D., Cook, N. et al.: 2024, Mass reproducibility and replicability: A new hope. I4R Discussion Paper 107 (Preprint).

Cuijpers, P., Cristea, I. A., Karyotaki, E., Reijnders, M. and Huibers, M. J.: 2016, How effective are cognitive behavior therapies for major depression and anxiety disorders? a meta-analytic update of the evidence, *World Psychiatry* **15**(3), 245–258.

Evans, D. K. and Yuan, F.: 2022, How big are effect sizes in international education studies?, *Educational Evaluation and Policy Analysis* **44**(3), 532–540.

Guo, K., Islam, A., List, J., Vlassopoulos, M. and Zenou, Y.: 2024, Early childhood education, parental social networks, and child development, *CDES Working Paper No.* **15/24**.

Islam, A.: 2024, Dos and don'ts when implementing randomized controlled trials in developing countries, *CDES Working Paper No.* **02/24**.

Saleh, A.: 2020, In Bangladesh, Covid-19 threatens to cause a humanitarian crisis. Weblink. Online; Accessed April 10, 2020.

Schwab, S.: 2020, Re-estimating 400,000 treatment effects from intervention studies in the cochrane database of systematic reviews [data set], *Open Science Framework. doi. org/10.17605/OSF. IO/XJV9G* .

Siddique, A., Rahman, T., Pakrashi, D., Islam, A. and Ahmed, F.: 2024a, Raising health awareness in rural communities: A randomized experiment in Bangladesh and India, *Review of Economics and Statistics* **106**(3), 638–654.

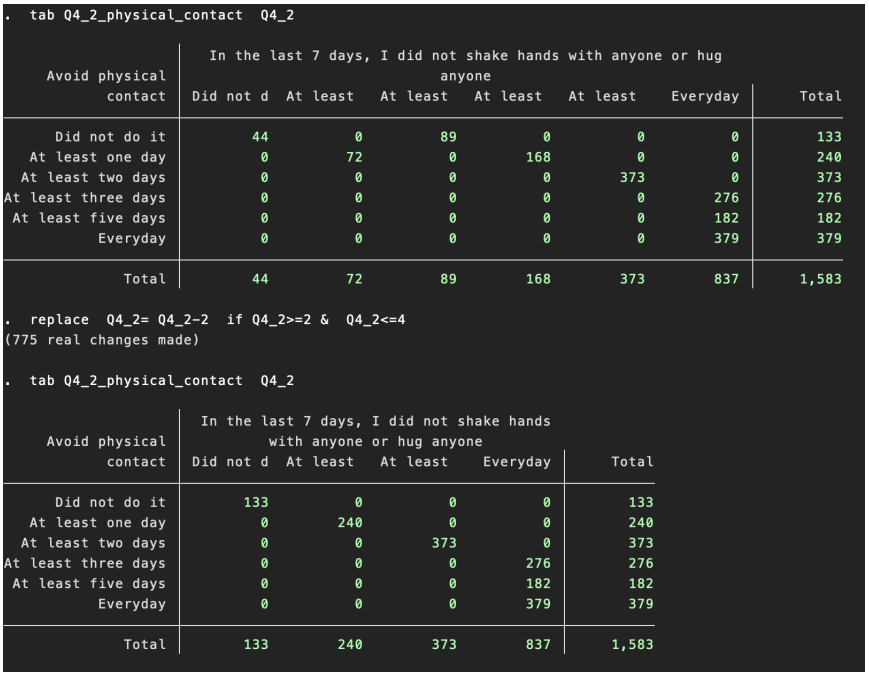
Siddique, A., Rahman, T., Pakrashi, D., Islam, A. and Ahmed, F.: 2024b, Replication Data for: Raising Health Awareness in Rural Communities: A Randomized Experiment in Bangladesh and India. https://doi.org/10.7910/DVN/VAOEHQ, Harvard Dataverse, V1.

Vlassopoulos, M., Siddique, A., Rahman, T., Pakrashi, D., Islam, A. and Ahmed, F.: 2024, Improving womens mental health during a pandemic, *American Economic Journal: Applied Economics* **16**(2), 422–455.

Online appendix

A Appendix Figures

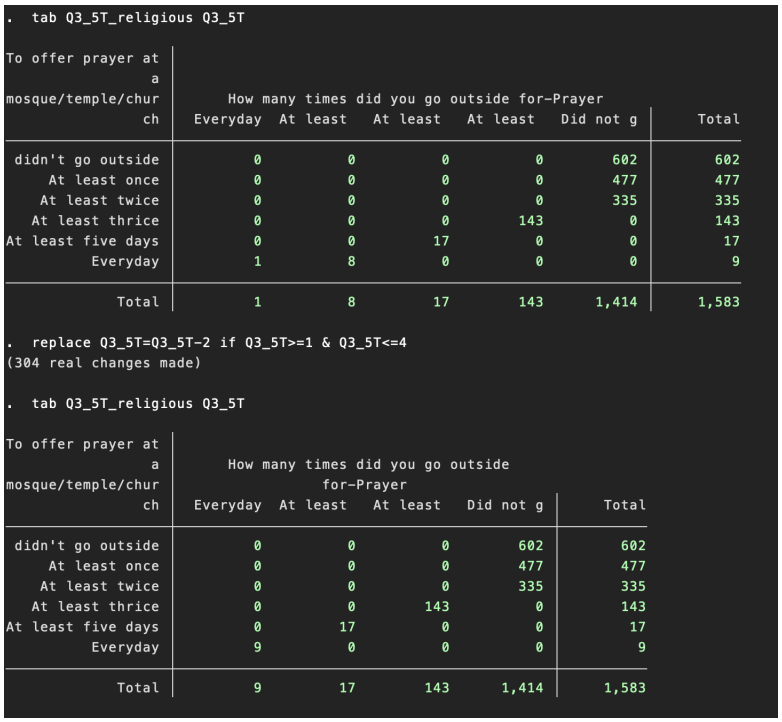
Figure A1: COVID-19 compliance in INFO24 and COUNSELING24, "avoiding physical contact"



```
. tab Q4_2_physical_contact Q4_2

                  |  In the last 7 days, I did not shake hands with anyone or hug
                  |                                anyone
Avoid physical    |
        contact   | Did not d  At least   At least   At least   At least    Everyday |     Total
------------------+------------------------------------------------------------------+----------
     Did not do it |        44         0        89         0         0         0 |       133
   At least one day |         0        72         0       168         0         0 |       240
  At least two days |         0         0         0         0       373         0 |       373
At least three days |         0         0         0         0         0       276 |       276
 At least five days |         0         0         0         0         0       182 |       182
          Everyday |         0         0         0         0         0       379 |       379
------------------+------------------------------------------------------------------+----------
             Total |        44        72        89       168       373       837 |     1,583

. replace Q4_2= Q4_2-2 if Q4_2>=2 & Q4_2<=4
(775 real changes made)

. tab Q4_2_physical_contact Q4_2

                  |  In the last 7 days, I did not shake hands
                  |         with anyone or hug anyone
Avoid physical    |
        contact   | Did not d  At least   At least    Everyday |     Total
------------------+---------------------------------------------+----------
     Did not do it |       133         0         0         0 |       133
   At least one day |         0       240         0         0 |       240
  At least two days |         0         0       373         0 |       373
At least three days |         0         0         0       276 |       276
 At least five days |         0         0         0       182 |       182
          Everyday |         0         0         0       379 |       379
------------------+---------------------------------------------+----------
             Total |       133       240       373       837 |     1,583
```

Note: The screenshot displays the relationship between the "avoiding physical contact" variables from the first endline of INFO24 (Q2_1_physical_contact) and the baseline of COUNSELING24 (Q4_2). The bottom part of the screenshot shows that a simple recoding aligns most response values with the exception of the highest possible response in COUNSELING24.

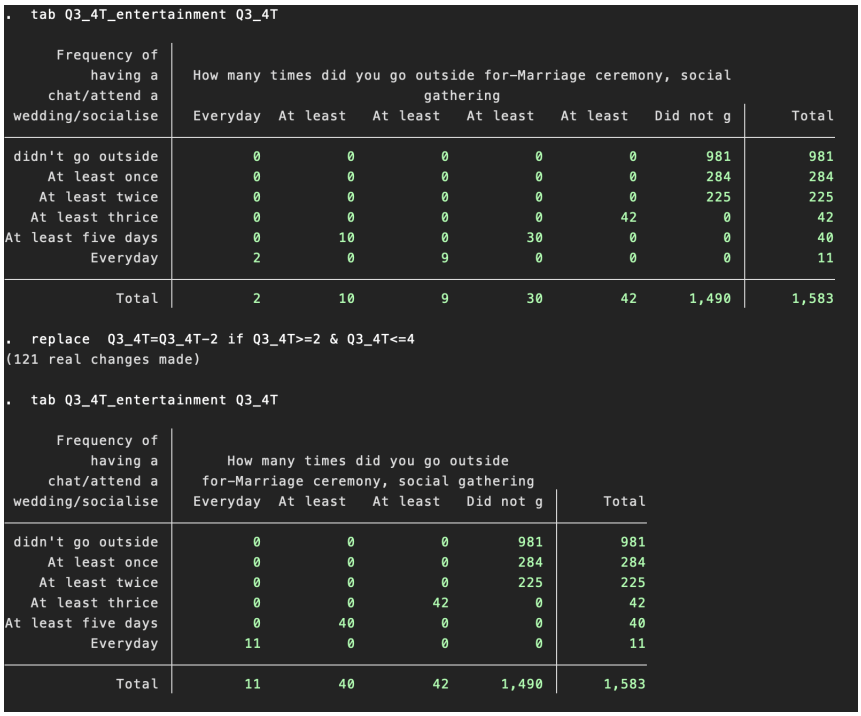
Figure A2: COVID-19 compliance in INFO24 and COUNSELING24, "going out for religious purposes"



```
. tab Q3_5T_religious Q3_5T

To offer prayer at a mosque/temple/church | How many times did you go outside for-Prayer
                                          | Everyday  At least   At least   At least   Did not g |     Total
------------------------------------------+-----------------------------------------------------+----------
                   didn't go outside      |        0          0          0          0        602 |       602
                       At least once      |        0          0          0          0        477 |       477
                      At least twice      |        0          0          0          0        335 |       335
                     At least thrice      |        0          0          0        143          0 |       143
                  At least five days      |        0          0         17          0          0 |        17
                            Everyday      |        1          8          0          0          0 |         9
------------------------------------------+-----------------------------------------------------+----------
                               Total      |        1          8         17        143      1,414 |     1,583

. replace Q3_5T=Q3_5T-2 if Q3_5T>=1 & Q3_5T<=4
(304 real changes made)

. tab Q3_5T_religious Q3_5T

To offer prayer at a mosque/temple/church | How many times did you go outside for-Prayer
                                          | Everyday  At least   At least   Did not g |     Total
------------------------------------------+-----------------------------------------+----------
                   didn't go outside      |        0          0          0        602 |       602
                       At least once      |        0          0          0        477 |       477
                      At least twice      |        0          0          0        335 |       335
                     At least thrice      |        0          0        143          0 |       143
                  At least five days      |        0         17          0          0 |        17
                            Everyday      |        9          0          0          0 |         9
------------------------------------------+-----------------------------------------+----------
                               Total      |        9         17        143      1,414 |     1,583
```

Note: The screenshot displays the relationship between the "going out for religious purposes" variables from the first endline of INFO24 (Q3_5T_religious) and the baseline of COUNSELING24 (Q3_5T). The bottom part of the screenshot shows that a simple recoding aligns most response values with the exception of the highest possible response in COUNSELING24.

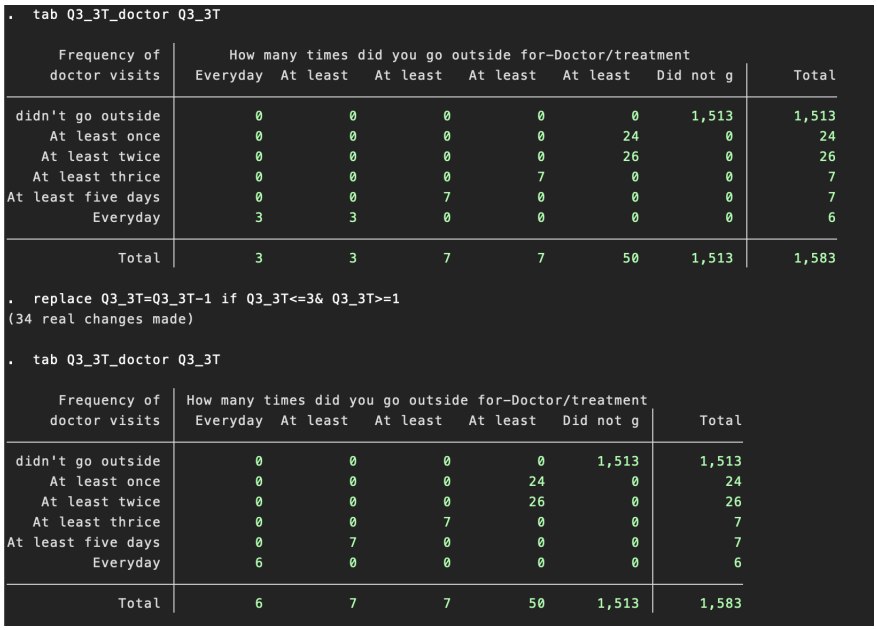
Figure A3: COVID-19 compliance in INFO24 and COUNSELING24, "going out for entertainment"



```
. tab Q3_4T_entertainment Q3_4T

      Frequency of |
          having a |  How many times did you go outside for-Marriage ceremony, social
      chat/attend a |                              gathering
 wedding/socialise | Everyday  At least   At least   At least   At least   Did not g |     Total
-------------------+-----------------------------------------------------------------+----------
 didn't go outside |        0          0          0          0          0        981 |       981
     At least once |        0          0          0          0          0        284 |       284
    At least twice |        0          0          0          0          0        225 |       225
   At least thrice |        0          0          0          0         42          0 |        42
At least five days |        0         10          0         30          0          0 |        40
         Everyday  |        2          0          9          0          0          0 |        11
-------------------+-----------------------------------------------------------------+----------
             Total |        2         10          9         30         42      1,490 |     1,583

. replace  Q3_4T=Q3_4T-2 if Q3_4T>=2 & Q3_4T<=4
(121 real changes made)

. tab Q3_4T_entertainment Q3_4T

      Frequency of |
          having a |      How many times did you go outside
      chat/attend a |  for-Marriage ceremony, social gathering
 wedding/socialise | Everyday  At least   At least   Did not g |     Total
-------------------+----------------------------------------+----------
 didn't go outside |        0          0          0        981 |       981
     At least once |        0          0          0        284 |       284
    At least twice |        0          0          0        225 |       225
   At least thrice |        0          0         42          0 |        42
At least five days |        0         40          0          0 |        40
         Everyday  |       11          0          0          0 |        11
-------------------+----------------------------------------+----------
             Total |       11         40         42      1,490 |     1,583
```

Note: The screenshot displays the relationship between the "going out for entertainment" variables from the first endline of INFO24 (Q3_4_entertainment) and the baseline of COUNSELING24 (Q3_4T). The bottom part of the screenshot shows that a simple recoding aligns most response values with the exception of the highest possible response in COUNSELING24.

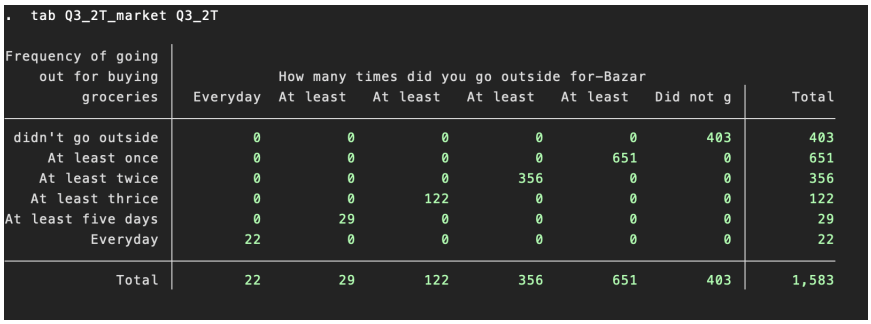
Figure A4: COVID-19 compliance in INFO24 and COUNSELING24, "going out for doctor"



```
. tab Q3_3T_doctor Q3_3T

     Frequency of |      How many times did you go outside for-Doctor/treatment
     doctor visits |  Everyday  At least   At least   At least   At least   Did not g |     Total
-------------------+------------------------------------------------------------------+----------
 didn't go outside |         0          0          0          0          0      1,513 |     1,513
     At least once |         0          0          0          0         24          0 |        24
    At least twice |         0          0          0          0         26          0 |        26
   At least thrice |         0          0          0          7          0          0 |         7
At least five days |         0          0          7          0          0          0 |         7
          Everyday |         3          3          0          0          0          0 |         6
-------------------+------------------------------------------------------------------+----------
             Total |         3          3          7          7         50      1,513 |     1,583

. replace Q3_3T=Q3_3T-1 if Q3_3T<=3& Q3_3T>=1
(34 real changes made)

. tab Q3_3T_doctor Q3_3T

     Frequency of | How many times did you go outside for-Doctor/treatment
     doctor visits |  Everyday  At least   At least   At least   Did not g |     Total
-------------------+----------------------------------------------------------+----------
 didn't go outside |         0          0          0          0      1,513 |     1,513
     At least once |         0          0          0         24          0 |        24
    At least twice |         0          0          0         26          0 |        26
   At least thrice |         0          0          7          0          0 |         7
At least five days |         0          7          0          0          0 |         7
          Everyday |         6          0          0          0          0 |         6
-------------------+----------------------------------------------------------+----------
             Total |         6          7          7         50      1,513 |     1,583
```

Note: The screenshot displays the relationship between the "going out for doctor" variables from the first endline of INFO24 (Q3_3T_doctor) and the baseline of COUNSELING24 (Q3_3T). The bottom part of the screenshot shows that a simple recoding aligns most response values with the exception of the second to highest response in COUNSELING24.

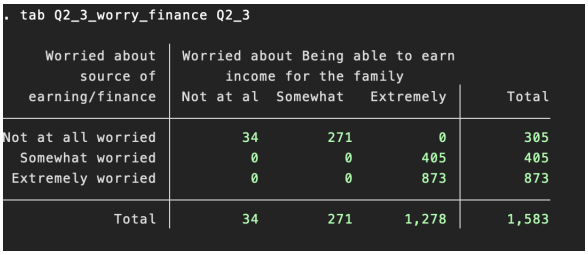
Figure A5: COVID-19 compliance in INFO24 and COUNSELING24, "going to market"



```
. tab Q3_2T_market Q3_2T

Frequency of going |
    out for buying |          How many times did you go outside for-Bazar
         groceries |  Everyday  At least   At least   At least   At least   Did not g |     Total
-------------------+------------------------------------------------------------------+----------
 didn't go outside |         0          0          0          0          0        403 |       403
     At least once |         0          0          0          0        651          0 |       651
    At least twice |         0          0          0        356          0          0 |       356
   At least thrice |         0          0        122          0          0          0 |       122
At least five days |         0         29          0          0          0          0 |        29
          Everyday |        22          0          0          0          0          0 |        22
-------------------+------------------------------------------------------------------+----------
             Total |        22         29        122        356        651        403 |     1,583
```

Note: The screenshot displays the relationship between the "going to market" variables from the first endline of INFO24 (Q3_2T_market) and the baseline of COUNSELING24 (Q3_2T). As can be seen, all participants responded exactly the same in the two surveys.

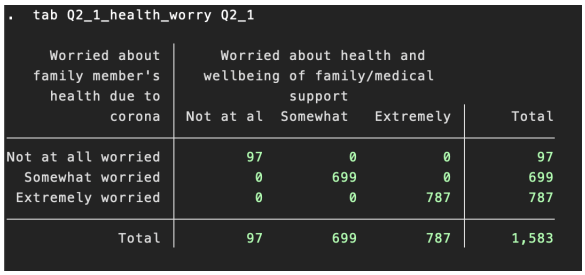
Figure A6: Financial worries in INFO24 and COUNSELING24



```
. tab Q2_3_worry_finance Q2_3

    Worried about |  Worried about Being able to earn
        source of |       income for the family
  earning/finance | Not at al  Somewhat   Extremely |     Total
------------------+---------------------------------+----------
Not at all worried |        34        271          0 |       305
 Somewhat worried |         0          0        405 |       405
Extremely worried |         0          0        873 |       873
------------------+---------------------------------+----------
            Total |        34        271      1,278 |     1,583
```

Note: The screenshot displays the relationship between the "financial worries" variables from the first endline of INFO24 (Q2_3_worry_finance) and the baseline of COUNSELING24 (Q2_3).

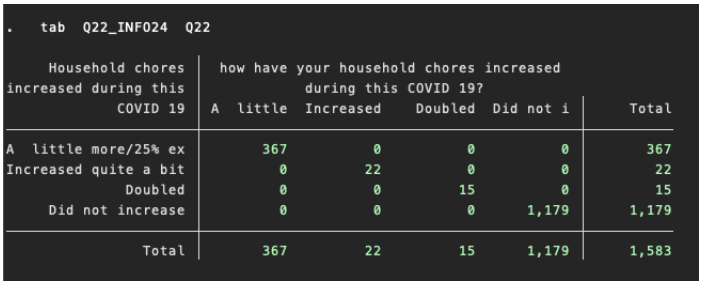
Figure A7: Health related worries in INFO24 and COUNSELING24



```
.  tab Q2_1_health_worry Q2_1

    Worried about |    Worried about health and
   family member's |   wellbeing of family/medical
    health due to |              support
            corona | Not at al  Somewhat   Extremely |     Total
-------------------+---------------------------------+----------
Not at all worried |        97          0          0 |        97
 Somewhat worried |         0        699          0 |       699
Extremely worried |         0          0        787 |       787
-------------------+---------------------------------+----------
            Total |        97        699        787 |     1,583
```

Note: The screenshot displays the relationship between the "health worries" variables from the first endline of INFO24 (Q2_1_health_worry) and the baseline of COUNSELING24 (Q2_1). As can be seen, all participants responded exactly the same in the two surveys.

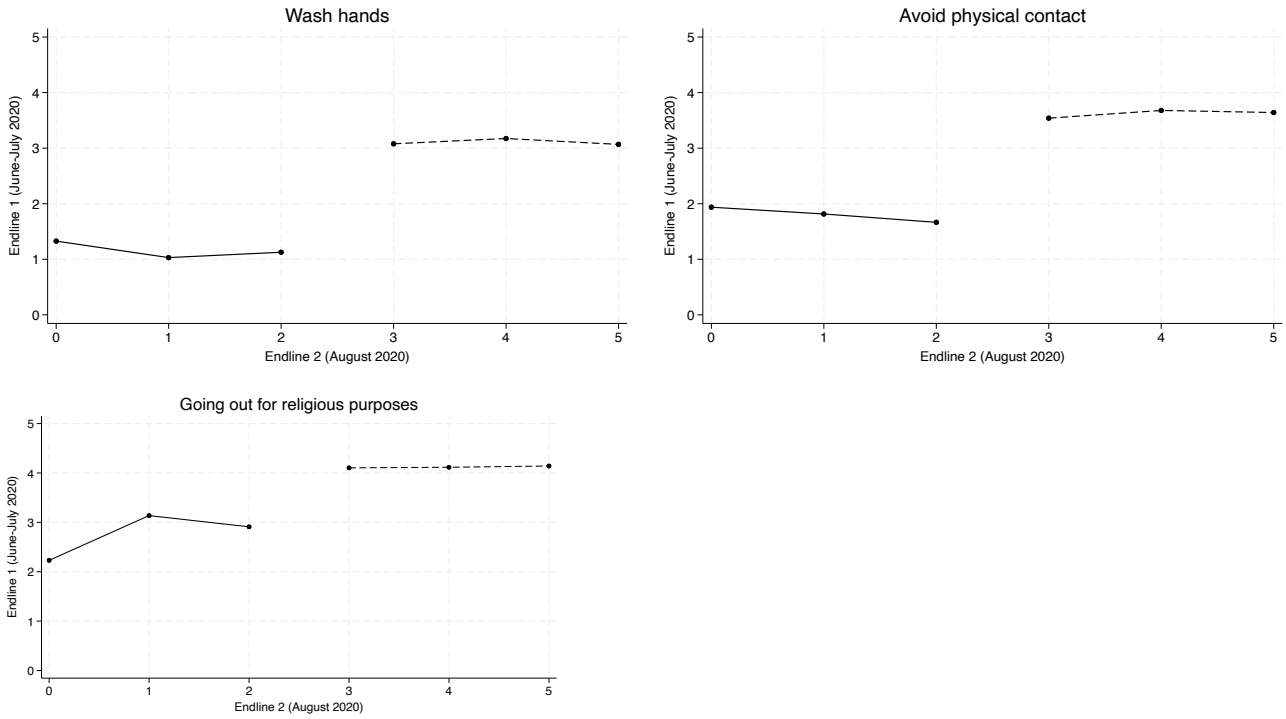
Figure A8: Increased household chores due to COVID-19 in INFO24 and COUNSELING24



```
.   tab  Q22_INFO24  Q22

    Household chores |  how have your household chores increased
increased during this |            during this COVID 19?
            COVID 19 | A  little  Increased    Doubled  Did not i |     Total
---------------------+-------------------------------------------+----------
A  little more/25% ex |       367          0          0          0 |       367
Increased quite a bit |         0         22          0          0 |        22
              Doubled |         0          0         15          0 |        15
     Did not increase |         0          0          0      1,179 |     1,179
---------------------+-------------------------------------------+----------
                Total |       367         22         15      1,179 |     1,583
```

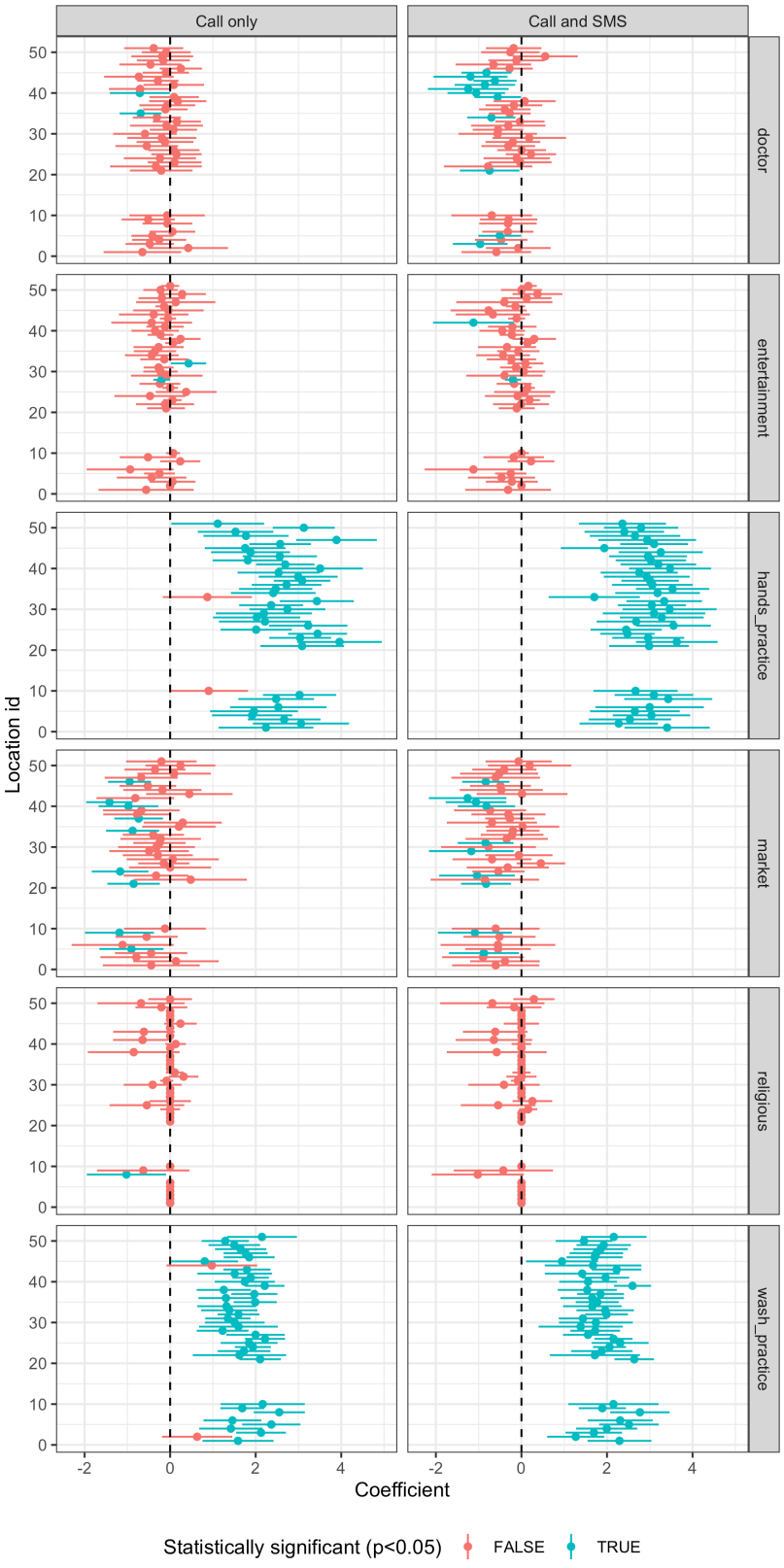
Note: The screenshot displays the relationship between the "household chores" variables from the second endline of INFO24 (Q22_INFO24) and the baseline of COUNSELING24 (Q22). As can be seen, all participants respondents exactly the same in the two surveys. Note that we renamed the variable Q22 in INFO24 to Q22_INFO24, to avoid the same variable name in the two surveys.

Figure A9: Discontinuous responses across the two endline surveys, conditioning on the second endline



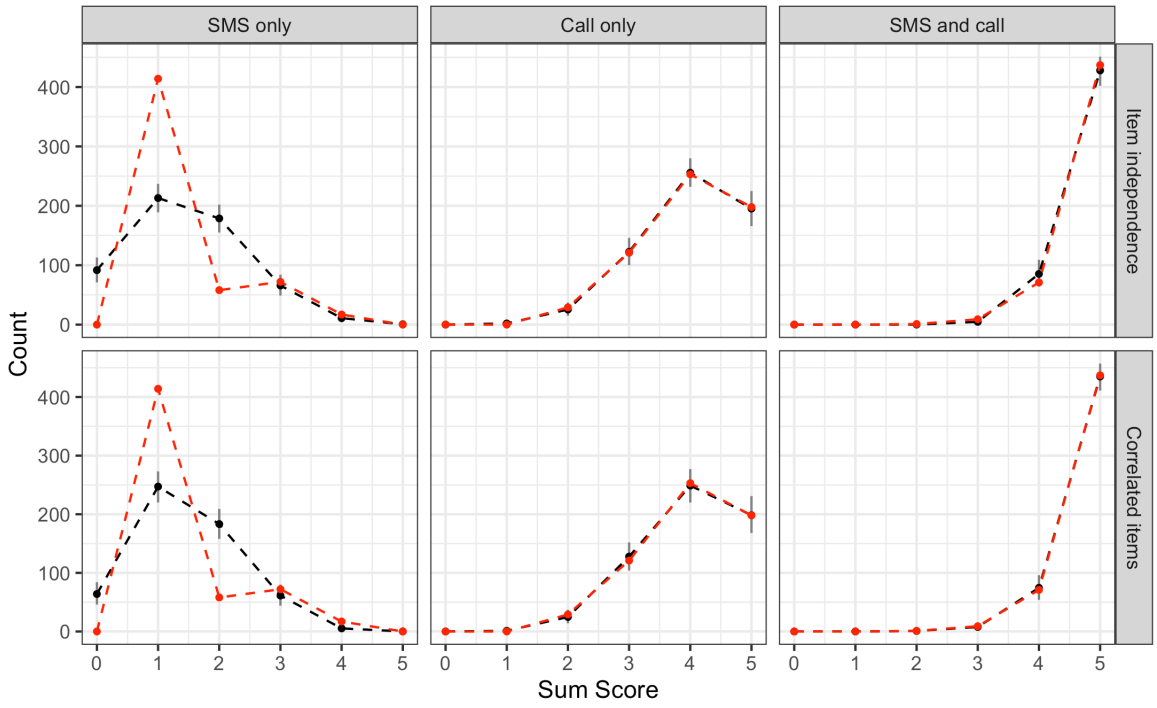
Note: The figures are similar to those in Figure 6, but with the axes flipped. In other words, these figures display average responses to three COVID-19 compliance question in the first endline survey, calculated separately for each response value in the second endline survey. For example, the first dot in the upper left figure displays the average response to the hand washing question in the first endline for the respondents that answered *0="Didn't do it"* in the second endline survey. The break in the figures correspond to the cutoff values used to discretize the outcome variables in the analysis of INFO24.

Figure A10: Village-wise treatment effects in India, items in the COVID-19 compliance measure



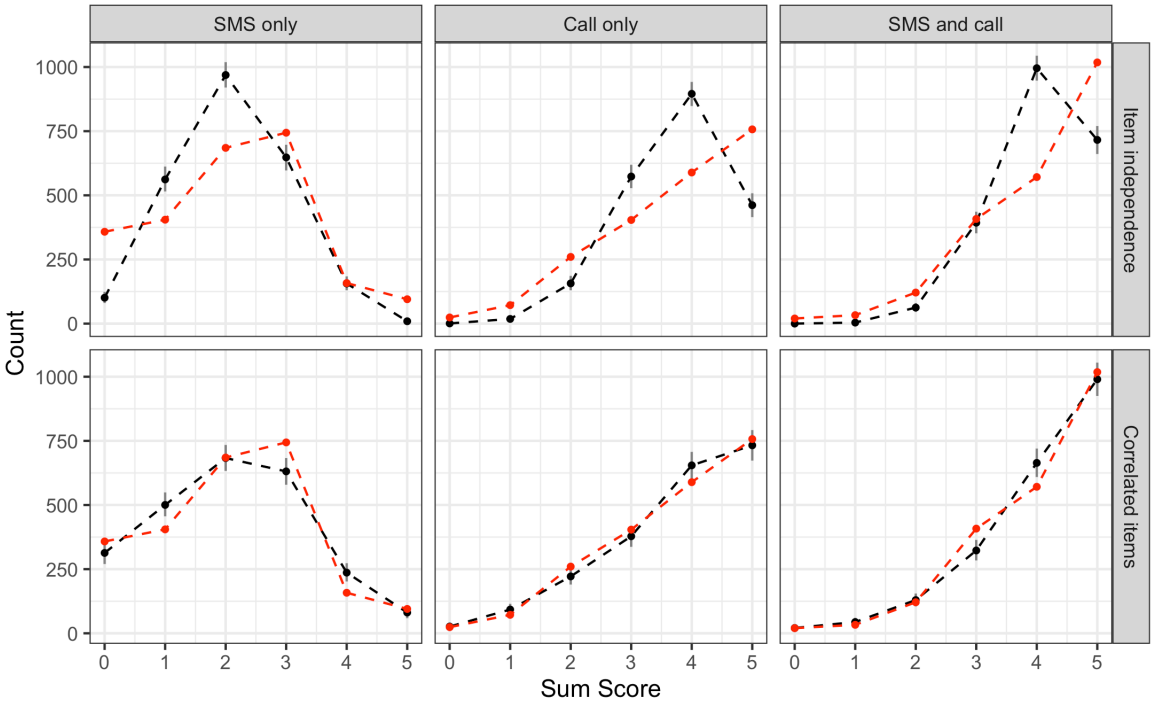
Note: The figure displays treatment effects on each of the COVID-19 compliance items, estimated separately for each of the 40 Indian villages in the INFO24 sample. Outcome variables are standardized to the "SMS only" group. The lines display 95% confidence intervals for each village. Green lines indicate estimates statistically different from zero at a 5% level.

Figure A11: COVID-19 awareness in India: Expected vs. observed sumscores



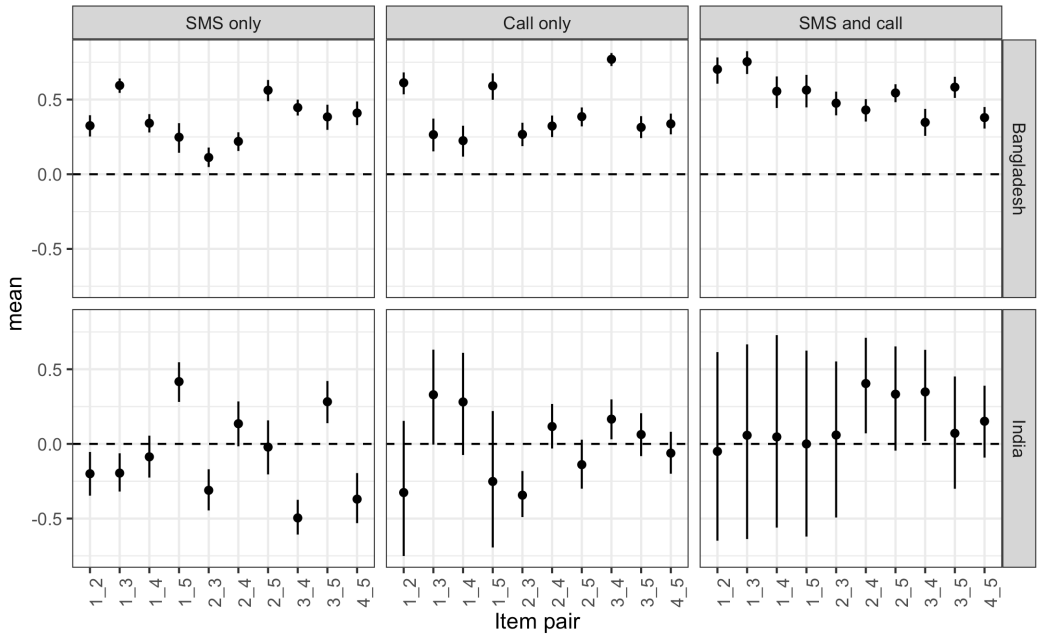
Note: The figure shows the observed distribution of the number of correct responses in red against the distribution expected under two statistical models, both estimated separately for each treatment group. The upper part of the figure is a reproduction of Figure 8. The "Item independence" model treats each item response as a bernoulli trial with a success probability estimated from the data, while the "Correlated items" model estimates a multivariate probit. Both models are estimated in the Stan language for probabilistic Bayesian models, with the reference distributions generated by drawing synthetic response data from each posterior draw of parameter values..

Figure A12: COVID-19 awareness in Bangladesh: Expected vs. observed number of correct responses



Note: The figure shows the observed distribution of the number of correct responses in red against the distribution expected under two statistical models, both estimated separately for each treatment group. The "Item independence" model treats each item response as a bernoulli trial with a success probability estimated from the data, while the "Correlated items" model estimates a multivariate probit. Both models are estimated in the Stan language for probabilistic Bayesian models, with the reference distributions generated by drawing synthetic response data from each posterior draw of parameter values.

Figure A13: COVID-19 awareness: Item-wise correlation



Note: The figure shows the estimated item-pair correlations from the correlation matrix estimated in the multivariate probit. The model was estimated separately for each treatment group x country combination.

B Appendix Tables

Table A1: Average compliance to COVID-19 guidelines, INFO24 and COUNSELING24

	SMS		Call		Call+SMS	
	COUNS24	INFO24	COUNS24	INFO24	COUNS24	INFO24
COVID-19 Compliance						
Washing hands	0.519	0.090	0.842	0.532	0.871	0.667
Avoid physical contact	0.715	0.187	0.933	0.630	0.983	0.807
Going out: religious	0.995	0.855	0.989	0.903	0.998	0.926
Going out: markets	0.751	0.751	0.959	0.959	0.982	0.982
Going out: doctor	0.981	0.976	0.998	0.994	0.998	0.994
Going out: entertainment	0.972	0.886	0.989	0.955	1.000	0.989
COVID-19 Awareness						
Washing hands	0.119	0.749	0.597	0.937	0.733	0.994
Avoid physical contact	0.171	0.573	0.396	0.844	0.628	0.930
Cover mouth	0.135	0.266	0.468	0.745	0.619	0.855
No handshake	0.230	0.445	0.649	0.794	0.818	0.877

Note: The table displays average compliance to COVID-19 health guidelines from the first endline of INFO24, collected in June-July 2020, and the baseline of COUNSELING24, collected in May-June 2020. The variables in the table are binary, where 1 indicates higher compliance. The binary variables are constructed using the cutoff values from INFO24.