# A Comment on "Improving Women's Mental Health During a Pandemic"

Abel Brodeur, Lenka Fiala, Jack Fitzgerald, Essi Kujansuu, David Valenta, Ole Rogeberg, and Gunther Bensch

February 24, 2025

**Abstract**

Vlassopoulos et al. (2024) find that after providing two hours of telephone counseling over three months, a sample of Bangladeshi women saw significant reductions in stress and depression after ten months. We find three anomalies. First, estimates are almost entirely driven by reverse-scored survey items, which are handled inconsistently both in the code and in the field. Second, participants in this experiment are reused from multiple prior experiments conducted by the paper's authors, and estimates are extremely sensitive to the experiment from which participants originate. Finally, inconsistencies and irregularities in raw survey files raise doubts about the data.

KEYWORDS: Reproduction, Replication, Mental health, COVID-19 pandemic

JEL CODES: B41, C12, I12, I18, J16, O12

Brodeur: University of Ottawa and Institute for Replication. Corresponding author's e-mail: abrodeur@uottawa.ca. Fiala: University of Bergen and Institute for Replication. Fitzgerald: Vrije Universiteit Amsterdam, Tinbergen Institute, and Institute for Replication. Kujansuu: University of Innsbruck, University of Turku, and Institute for Replication. Valenta: University of Ottawa and Institute for Replication. Rogeberg: The Ragnar Frisch Centre for Economic Research and Institute for Replication. Bensch: RWI - Leibniz Institute for Economic Research and Institute for Replication. We thank RushTranslate for providing official translations of the Bengali text in the raw data into English; all other errors are our own. We are grateful to colleagues in Bangladesh who wish to remain anonymous. This project has received approval from the Ethical Review Board of Vrije Universiteit Amsterdam's School of Business and Economics. Replication materials for this paper can be found at https://osf.io/pvkhy/.

## 1 Introduction

Vlassopoulos et al. (2024b) – henceforth VEA24 – report the results of a randomized field experiment offering telephone talk therapy to a sample of Bangladeshi women in rural villages during the COVID-19 pandemic. Those assigned to the treatment group received four counseling sessions that provided mental health support and information on COVID-19. The four counseling sessions lasted two hours in total, and were spread across a three-month period.

VEA24 report that data is collected in three waves. The first wave, collected in May 2020, serves as a baseline before counseling is administered between July and October 2020. The second wave, a one-month endline, was collected in November 2020, and the third wave, a ten-month endline, was collected in August 2021.

The intervention's effects are reported to be numerous and large. VEA24 report that the intervention yields significant improvements in stress, depression, happiness, life satisfaction, future aspirations, food security, time-intensive parental investment, attitudes about gender norms, attitudes about intimate partner violence, altruism, COVID-19 compliance, COVID-19 vaccination, and confidence in overcoming the COVID-19 pandemic, among other outcomes. Focusing on mental health outcomes, the intervention is reported to result in "reductions of 20 percent in the prevalence of moderate and severe stress and 33 percent in depression" after ten months (pg. 422). In standardized effect size terms, VEA24 report that after ten months, the intervention decreased both stress and depression scores by over 0.5 standard deviations (relative to the control group). VEA24 themselves note that these effect sizes are particularly large in this literature:

> "In fact, these estimated effects are large compared to the short-run impact of cognitive behavioral therapy interventions in Pakistan (Baranov et al. 2020) and Kenya (Bryant et al. 2017) and those found by telephone-delivered interventions (Mohr et al. 2008) and studies that use psychotherapy to improve individual psychological well-being (Cui-

> jpers et al. 2013, Cuijpers et al. 2010), considering that these earlier interventions were typically long, delivered by mental health professionals, and addressed people who suffered from depression. Notably, the impact of such interventions (including our own) is more sizable than the average effect size of economic transfers on mental health, which has been estimated to be 0.10 SD in low- and middle-income countries (McGuire et al. 2020). The effects of the current intervention are in the upper range of those reported in a recent meta-analysis of app-supported mental health interventions, in which effect sizes were found to range from 0.28 to 0.58 (Linardon et al. 2019)." (pg. 439).

In this comment prepared for the Institute for Replication (Brodeur et al. 2024), we examine the reproducibility, robustness, and credibility of VEA24. Our analysis relies on VEA24's publicly available replication repository (Vlassopoulos et al. 2024a). All our analyses were successfully reproduced by multiple coauthors.

Section 2 summarizes the intervention, and Section 3 shows that VEA24's conclusions about this intervention's effects on mental health are extremely sensitive to inconsistent handling of the main outcome variables. The primary mental health outcomes are Likert scales with several items that should be reverse-scored. These items are reverse-coded in VEA24's code at the one-month endline, but not at the ten-month endline. Consistently coding these scales across endlines reverses the results at the ten-month endline – the intervention is estimated to significantly *increase* participants' stress and depression after ten months. When we approached VEA24 about this issue, author Asad Islam claimed that this coding inconsistency was intentional, as enumerators were instructed in the field to systematically deviate from survey materials for reverse-scored questions in the ten-month endline survey. Sensitivity analyses show that these reverse-scored items drive over 70% of VEA24's treatment effect estimates at the ten-month endline, and that when these items are removed from the scales, estimated treatment effects on mental health are not statistically significantly different from zero after ten months. We also com-

pare treatment effect estimates on thematically identical questions within the same scales in the ten-month endline surveys, where some are reverse-scored and some are straight-scored. Estimated treatment effects on the reverse-scored items are up to 87 times larger than estimates on thematically identical straight-scored items.

In Section 4, we show that most observations in VEA24 can be traced back to data from two prior experiments run by the Global Development and Research Initiative (GDRI) – i.e., VEA24's implementing NGO – and several authors of VEA24 (Guo et al. 2024, Ahmed et al. 2024). We find that the extent of reverse-scoring inconsistency systematically differs between these data sources. Interaction effects between treatment and sample origination are also quite large, often exceeding 0.35 standard deviations in size. We additionally find another paper published by several authors of VEA24 that uses a virtually identical sample to VEA24 despite reporting a completely different sampling approach (Rahman, Hasnain and Islam 2021).

Section 5 uncovers numerous inconsistencies between VEA24's published paper, its documentation, and the replication repository. In both endline surveys, the raw survey forms show that the labels on the primary stress index do not scale monotonically in agreement with the question, contradicting the design of the scale in the baseline survey and the description of the scale in VEA24's published materials. We also obtain official translations of the raw survey form's Bengali text; these back-translations show many discrepancies between the question labels in the raw survey files and their description in VEA24's published materials. Additionally, there are several inconsistencies concerning VEA24's pre-registration, including undisclosed deviations in sample selection criteria and missing data in the replication repository.

Finally, Section 6 uncovers irregularities in the raw survey data and survey forms. Several multiple-choice questions take on values that are not part of the original survey options, and many variables take on missing values where this should not be possible. The survey is reported to be deployed in the Kobo software suite, which accommodates survey forms in XLS format. However, the raw XLS survey forms in VEA24's repository contain formatting issues that prevent survey deploy-

ment and imply that Bengali translations to English answer labels are not displayed in the form. Additionally, we find that many question labels in the raw survey forms are abnormally cut off mid-sentence or even mid-word. Many of these label cutoffs correspond to Stata's 80-character cutoff for variable labels. Finally, we show that the ordering of strings in the `sharedStrings` element of the Excel file for the raw baseline data is inconsistent with a programatically-created Excel spreadsheet.

## 2 The Intervention and Its Quality

The treatment described in VEA24 is a telephone counseling intervention delivered in four sessions, totaling roughly two hours of counseling over three months. VEA24's Figure 1 states that sessions were conducted once every two to three weeks from July to October 2020. VEA24's Online Appendix D provides detailed telecounseling scripts for each of these four sessions.

The brief sessions were split between genuine counseling and data collection. VEA24 varyingly describe the average session as being "about 25 minutes" or "roughly 30 minutes" long.[1] VEA24's Online Appendix D shows that each session was structured around a common script. Sessions began with introductions of names and the overall topic of the session.[2] Before starting the counseling, participants were asked to rate their well-being and anxiety on a 0–10 scale, both of which were to be recorded. Participants thereafter received concrete advice related to mental health or COVID-19. Participants were then asked whether they understood the counseling provided, and if they answered 'no', clarification may have been given. The counseling session ended with participants being asked whether they had learned something new that day, and if so, to detail what they had learned. Participants' responses were again to be recorded. The data collected in these sessions are not available in the replication repository.

Additionally, half of the counseling sessions were not focused on mental health.

[1] For the former time quotation, see pgs. 423 and 427; for the latter, see pg. 431. Data on session times is not available in the replication repository.

[2] It is not clear if the same person counseled participants throughout all sessions.

The session scripts in VEA24's Online Appendix D show that only the second and fourth sessions focus on common mental health topics.[3] The first and third sessions focus almost exclusively on providing COVID-19 prevention advice, with the third session providing additional details about COVID-19 prevention behaviors involving pregnancy, babies, and children.[4] Consequently, treated participants effectively only received roughly one hour of targeted mental health counseling.

The individuals chosen to provide this counseling were not experienced with providing mental health therapy at the start of the experiment. The counselors were 18 "recent graduates in either psychology, public health, or social sciences from public universities in Bangladesh without any significant prior real-world counseling experience" who were trained via video conferencing (see pgs. 427-428). The therapy described in this section, provided by these people, is estimated in VEA24 to reduce stress by one fifth, and depression by one third, after ten months.

## 3 Index Scoring

### 3.1 Mental Health Indices

VEA24 primarily highlight treatment effects on two mental health outcomes, the first of which is stress. VEA24 use the ten-item Perceived Stress Scale (PSS), which consists of ten survey items on a five-point Likert-scale (Cohen and Williamson 1988). The scale ranges from 0 to 40; higher values indicate higher stress. VEA24 argue that individuals with scores above 13 can be diagnosed as 'stressed'.[5]

The second main mental health variable is depression. In the endline surveys, depression is measured using the ten-item Center for Epidemiological Studies De-

[3]The second session focuses on fostering better mental states by not dwelling on the pandemic, by not placing personal blame for hardship, and by following healthy routines, whereas the fourth session focuses on maintaining personal relationships with family and neighbors.

[4]The first session script does include text stating that knowing more about COVID-19 can reduce one's fear of it, and the third session script includes advice to not scold children when instructing them to stay at home and quarantine.

[5]VEA24 support this cutoff with two citations, though neither makes any mention of a PSS score of 13 or 14 as a stress diagnosis cutoff (Cohen et al. 1983, 1997). However, evidence from the United States suggests that a PSS score of 13 corresponds to that of a representative American (Cohen and Williamson 1988, Harris et al. 2023).

pression Scale (CES-D-10), a scale of ten four-point Likert items (Andresen et al. 1994). This scale takes values from 0 to 30, increasing with depressive symptoms. Prior literature validates that CES-D-10 scores of 10 or higher can reliably diagnose people with depression (Andresen et al. 1994).

VEA24 dichotomize the PSS and CES-D-10 scales into binary indicators that diagnose participants with stress and/or depression, but code these indicators inconsistently across survey waves. VEA24 use the aforementioned PSS score cutoff of 13 to diagnose whether participants are stressed. However, VEA24 varyingly treat this cutoff as either strict or weak at different waves. In the ten-month endline, VEA24 treat the inequality as weak by coding individuals as 'stressed' if PSS $\geq 13$ on line 428 of `Generate_variables.do`. At the baseline and one-month endline, VEA24 effectively treat the inequality as strict by coding participants as 'not stressed' if PSS $\leq 13$ on lines 160 and 290 of `Generate_variables.do` (respectively). Likewise, VEA24 also use the aforementioned CES-D-10 score cutoff of 10 to diagnose participants as 'stressed', but similarly code participants as 'not depressed' if CES-D-10 $\leq 10$ in the one-month endline and 'depressed' if CES-D-10 $\geq 10$ in the ten-month-endline (see lines 296 and 434 of repository file `Generate_variables.do`, respectively). Again, this effectively treats the diagnosis cutoff as strict in the one-month endline and weak in the ten-month endline. Throughout this paper, we maintain this inconsistent coding of diagnosis indicators to prevent our main checks on VEA24's findings from being confounded by other coding adjustments.

Depression is measured differently between baseline and the endlines, as the CES-D-10 is not part of the baseline survey. Instead, the baseline survey asks respondents to separately rate their feelings of anxiety, loneliness, hopelessness, and worthlessness in four separate four-point Likert scales (ranging from one to four), with lower values corresponding to higher intensities of these feelings. Participants are classified with binary indicators as feeling anxious, lonely, hopeless, and worthless if they answer each respective question with a value of one or two. These binary indicators are then summed to create a 0-4 depression scale for the baseline. Par-

ticipants are classified as 'depressed' if they are above-median on this summed scale (i.e., if the summed scale takes values three or four).[6] This difference between baseline and endline depression diagnosis is relevant for VEA24's Figure 4, where this variable is used to represent the proportion of depressed participants at baseline.

### 3.2 Reverse-Scoring

There are four items in the PSS questionnaire and two items in the CES-D-10 questionnaire that should be reverse-coded. These include the fourth, fifth, seventh, and eighth items in the PSS scale, along with the fifth and eighth items in the CES-D-10 scale.[7] Repository files `Endline1_kobo form.xls` and `Endline2_kobo form.xls` show that the questions eliciting these items share roughly the same text across both the one-month and ten-month endline surveys, that higher values on these items correspond to less stress or depression, and that responses are scored the same way for all questions in both questionnaires across both survey waves.

VEA24's code reverse-scores all items that should be reverse-scored in the one-month endline survey, but reverse-scores none of these items in the ten-month endline. Repository file `Generate_variables.do` generates the PSS and CES-D-10 scales using the code captured in the screenshots displayed in Figure 1. The reverse-scoring code for the one-month endline can be found on lines 210-213 and 220-221 of Panel A. Panel B shows that the respective code for generating mental health scales in the ten-month endline is missing the reverse-scoring script entirely.

If the reverse-coding is done consistently across survey waves – as it should be based on the questions in the raw survey files – then VEA24's results sign-flip at the ten-month endline. I.e., with consistent reverse-coding, the treatment group is estimated to be significantly *more* stressed and depressed than the control group after ten months. Figure 2 demonstrates the effect of this consistency check

[6] This account of the 'depressed' indicator's construction at baseline is based on the note in VEA24's Figure 4 and the repository files `Baseline_kobo form.xls` and `Generate_variables.do`.

[7] For confirmation, see pre-analysis plan file `PAP_covid_mental_health_aearegistry.pdf`. See Section 5.2 for more details.

Panel A: One-Month Endline

```
204  ****************************************
205  ** Generating 1-month endline outcomes **
206  ****************************************
207
208  **SCORES**
209  //PSS score
210  replace end_Q5_4=4-end_Q5_4 //reverse scoring
211  replace end_Q5_5=4-end_Q5_5
212  replace end_Q5_7=4-end_Q5_7
213  replace end_Q5_8=4-end_Q5_8
214
215  egen end_pss_score = rowtotal(end_Q5_1 end_Q5_2 end_Q5_3 end_Q5_4 end_Q5_5
     end_Q5_6 end_Q5_7 end_Q5_8 end_Q5_9 end_Q5_10)
216  replace end_pss_score=. if attrition==1
217  label variable end_pss_score "Perceived Stress Scale score (between 0 and 40)"
218
219  //Depression score
220  replace end_Q6_5=3-end_Q6_5 //reverse scoring
221  replace end_Q6_8=3-end_Q6_8
222
223  egen end_ces_score = rowtotal(end_Q6_1 end_Q6_2 end_Q6_3 end_Q6_4 end_Q6_5
     end_Q6_6 end_Q6_7 end_Q6_8 end_Q6_9 end_Q6_10)
224  replace end_ces_score=. if attrition==1
225  label variable end_ces_score "CES-D score (between 0 and 30)"
```

Panel B: Ten-Month Endline

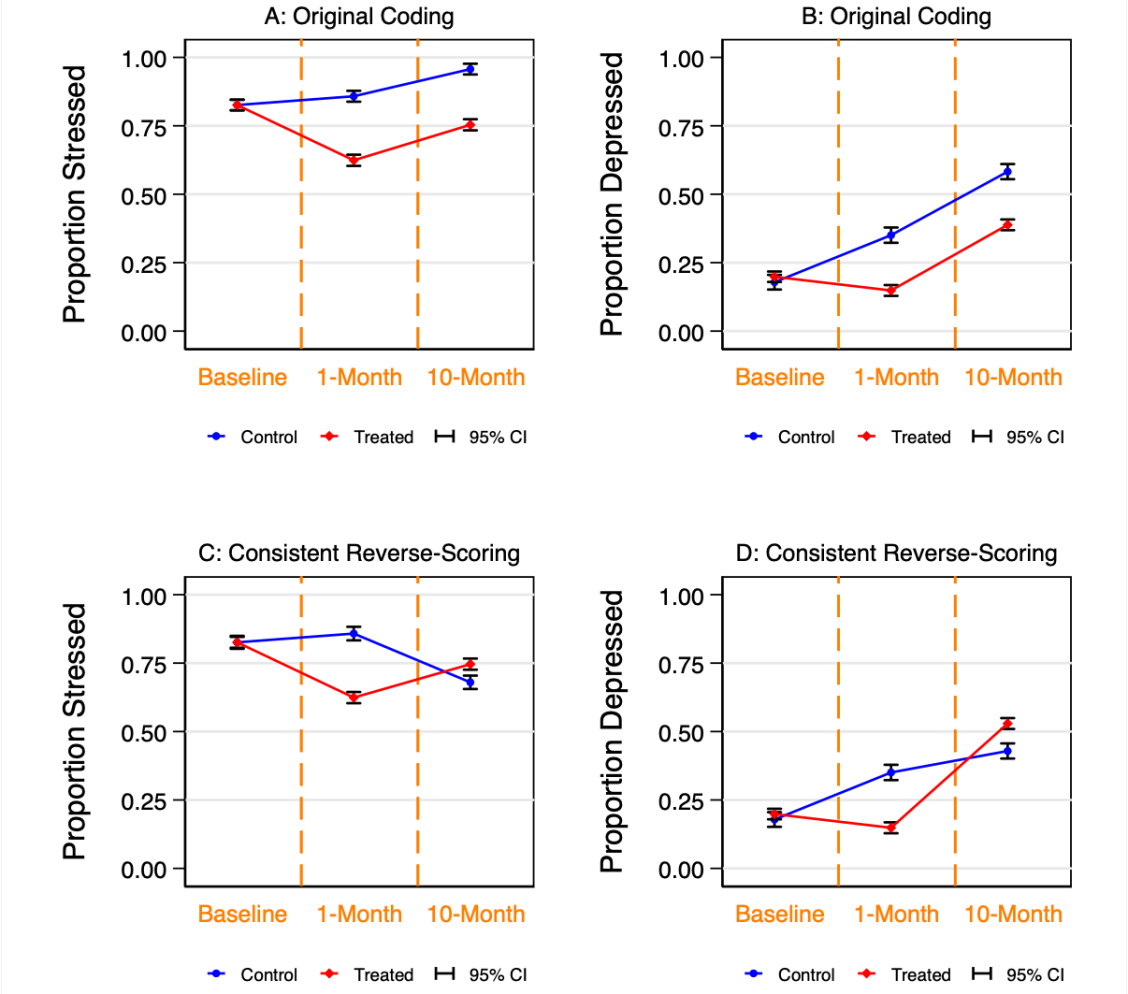
```
397  ************************************************
398  ** Generating 10-month endline outcome scores **
399  ************************************************
400
401  **SCORES**
402  //Stress
403  egen end2_pss_score = rowtotal(MH_1_1 MH_1_2 MH_1_3 MH_1_4 MH_1_5 MH_1_6 MH_1_7
     MH_1_8 MH_1_9 MH_1_10) if attrit2==0
404  label variable end2_pss_score "Perceived stress score (0-40) at 2nd endline"
405
406  //Depresison
407  egen end2_ces_score = rowtotal(Depr_1 Depr_2 Depr_3 Depr_4 Depr_5 Depr_6 Depr_7
     Depr_8 Depr_9 Depr_10) if attrit2==0
408  label variable end2_ces_score "Depression score (0-30) at 2nd endline"
```

*Note:* Code originates from VEA24's repository file `Generate_variables.do`.

Figure 1: VEA24's Code for Generating Mental Health Scales

graphically. Panels A and B in Figure 2 respectively capture the proportions of stressed and depressed participants by treatment group and survey wave under VEA24's original coding, reproducing Figure 4 in VEA24. Panels C and D in Figure 2 depict the results when reverse-coding is done consistently across survey waves. Consistently reverse-scoring the PSS and CES-D-10 scales across both endlines flips the signs of estimated treatment effects on stress and depression after ten months.

Table 1 shows the impact of this consistency adjustment numerically, displaying estimated treatment effects on standardized PSS scores, standardized CES-D-10 scores, and indicators for stress and depression, both at one-month and ten-month endlines. Panel A uses VEA24's original coding, which replicates the results in




*Note:* Participants are classified as 'stressed' and 'depressed' based on VEA24's original diagnosis cutoff scheme, as discussed in Section 3.1. Means by baseline treatment status and survey wave are displayed along with 95% confidence intervals.

Figure 2: Mental Health of Treatment Groups With and Without Consistent Reverse-Scoring

VEA24's Table 2 down to rounding error at three decimal places. Panel B applies consistent reverse-scoring at the ten-month endline. Accordingly, results in Panels A and B in Table 1 are identical at the one-month endline. However, estimated treatment effects in Panel B flip signs at the ten-month endline: treated participants exhibit significantly higher stress and depression than untreated participants. Remarkably, the sign-flipped treatment effect estimates on index scores retain most of the magnitude of the original treatment effects obtained using VEA24's coding scheme. In Table 1, the sign-flipped treatment effects in Columns 5 and 6 of

Panel B retain between 87-103% of the effect sizes of those estimates in Panel A. However, sign-flipped treatment effect estimates on diagnosis indicators are substantially smaller than the original estimates obtained using VEA24's repository code. In Table 1, sign-flipped treatment effect estimates in Columns 7 and 8 of Panel B retain between 33-54% of the magnitude of those estimates in Panel A.

### 3.3 Response by Asad Islam

We approached the authors of VEA24 with an earlier draft of this report to offer them an opportunity to respond. One of the authors, Asad Islam, promised a full response in due time, which we have not yet received after more than five weeks. In a partial response, Islam claimed that items which should ordinarily be reverse-scored in the PSS and CES-D-10 scales should actually not be reverse-scored in VEA24's code for the second endline, as enumerators were instructed in the field to deviate from survey materials during the ten-month endline survey. From Islam's email response on 16 January 2025:

> "Regarding the main point about the purported "coding error" concerning the PSS and CESD measures, we can confirm that the relevant PSS and CESD items that are meant to be reverse-coded, were collected in reverse order in the second endline and therefore did not require additional reversing. The decision to reverse these items during the second endline data collection was based on feedback from enumerators during the piloting phase of the 10-month endline questionnaire. Enumerators noted that respondents found it difficult to alternate between positively and negatively framed mental health questions during phone interviews conducted under lockdown conditions. To address this, the field team instructed enumerators to administer the questions in a way that eliminated the need for subsequent reversing of the answers. However, we recognize that this procedural detail was inadvertently omitted in Appendix B of the paper (Section B.2) and the replication readme file."

| | One-Month Endline Outcomes | | | | Ten-Month Endline Outcomes | | | |
|---|---|---|---|---|---|---|---|---|
| | Standardized PSS Score (1) | Standardized CES-D-10 Score (2) | Stress Diagnosis (3) | Depression Diagnosis (4) | Standardized PSS Score (5) | Standardized CES-D-10 Score (6) | Stress Diagnosis (7) | Depression Diagnosis (8) |
| **Panel A: Original Coding** | | | | | | | | |
| Treatment Effect, Without Covariates | -0.712 | -0.638 | -0.229 | -0.2 | -0.576 | -0.525 | -0.202 | -0.193 |
| | (0.06) | (0.052) | (0.023) | (0.026) | (0.077) | (0.065) | (0.018) | (0.03) |
| | [0] | [0] | [0] | [0] | [0] | [0] | [0] | [0] |
| | {0} | {0} | {0} | {0} | {0} | {0} | {0} | {0} |
| Treatment Effect, With Covariates | -0.696 | -0.652 | -0.22 | -0.207 | -0.551 | -0.513 | -0.551 | -0.191 |
| | (0.059) | (0.05) | (0.022) | (0.025) | (0.075) | (0.063) | (0.018) | (0.029) |
| | [0] | [0] | [0] | [0] | [0] | [0] | [0] | [0] |
| | {0} | {0} | {0} | {0} | {0} | {0} | {0} | {0} |
| **Panel B: Consistent Reverse-Scoring** | | | | | | | | |
| Treatment Effect, Without Covariates | -0.712 | -0.638 | -0.229 | -0.2 | 0.559 | 0.462 | 0.067 | 0.099 |
| | (0.06) | (0.052) | (0.023) | (0.026) | (0.063) | (0.059) | (0.023) | (0.028) |
| | [0] | [0] | [0] | [0] | [0] | [0] | [0.004] | [0.001] |
| | {0} | {0} | {0} | {0} | {0} | {0} | {0.004} | {0.001} |
| Treatment Effect, With Covariates | -0.696 | -0.652 | -0.22 | -0.207 | 0.571 | 0.47 | 0.07 | 0.104 |
| | (0.059) | (0.05) | (0.022) | (0.025) | (0.065) | (0.058) | (0.024) | (0.028) |
| | [0] | [0] | [0] | [0] | [0] | [0] | [0.004] | [0] |
| | {0} | {0} | {0} | {0} | {0} | {0} | {0.006} | {0} |
| Observations | 2220 | 2220 | 2220 | 2220 | 2254 | 2254 | 2254 | 2254 |

*Note*: Treatment effect estimates on the outcome displayed in the column title are shown along with standard errors clustered at the village level in parentheses, raw $p$-values in straight brackets, and Westfall-Young family-wise error rate-controlled $p$-values in curled brackets (Westfall and Young 1993). Standardized scores hold a mean of zero and a standard deviation of one in the control group. Treatment effects are based on baseline treatment status. For each panel and each treatment effect type (i.e., with covariates and without covariates), Westfall-Young $p$-values are calculated as in VEA24's Table 2 across all eight outcomes in the columns of this table, happiness, life satisfaction, future aspirations, and COVID-19 compliance. All models control for union fixed effects, and all models on PSS scores and diagnosis indicators control for baseline PSS scores and diagnosis indicators (respectively). Additional covariates include respondents' age, education, and occupation, the occupation of respondents' husbands, a dummy indicating whether the respondent reports that their household lost income during the COVID-19 pandemic, a dummy indicating whether the respondent reports that their household chores increased after the COVID-19 lockdowns, a dummy indicating whether the respondent is the head of their household, the number of people in the respondent's household, and the number of children in the respondent's household.

Table 1: Treatment Effect Estimates With and Without Consistent Reverse-Scoring

This email effectively claims that for the second endline survey, enumerators were instructed in the field to reverse certain questions in the PSS and CES-D-10 scales to participants and record people's agreement with the newly-reversed questions. This is not reflected in the repository's survey files, and could only occur if enumerators were explicitly instructed to go off-script for these specific questions. E.g., one can imagine that statement eight on the CES-D-10 scale was framed to participants as "I was sad" rather than the original text of "I was happy", which would eliminate the need for reverse-scoring in the ten-month endline data.

There is evidence corroborating Islam's account. Online Appendix Table A1 shows cross-wave correlations between reverse-scored survey items. Items that should be reverse-coded in the one-month endline positively correlate with one another, but negatively correlate with the same items in the ten-month endline.

If this is indeed how reverse-scored items were collected for mental health outcomes at the ten-month endline, then the paper fails to disclose a serious source of bias risk. Endline survey forms `Endline1_kobo form.xls` and `Endline2_kobo form.xls` require enumerators to enter treatment status variables `treat2` and `treat3`. This implies that not only could enumerators know participants' treatment status, but they were additionally being instructed by the field team to systematically deviate from written survey protocols in ways that remain undisclosed in the paper and its published materials. This is not standard practice in development trials, and Section 3.4 shows that these reverse-scored items in the ten-month endline are abnormally sensitive to treatment. Additionally, Section 4 shows that many participants in this experiment are reused from other prior experiments, and in Section 4.4, we demonstrate that the sensitivity of reverse-scored items to treatment significantly differs depending on the experiment from which the sample originates.

### 3.4 Sensitivity Analysis

Several sensitivity analyses of the reverse-scoring issue show that VEA24's findings are highly sensitive to the reverse-scored items. First, we create versions of the PSS

and CES-D-10 scores based only on the straight-scored items in each index which do not require reverse-scoring. We classify participants as 'stressed' and/or 'depressed' analogously to VEA24, but adjust the diagnosis cutoffs by the proportion of items that are straight-scored. For example, the PSS stress diagnosis cutoff is 13, and six of the ten items in the PSS index are straight-scored – for this sensitivity check, we thus classify participants as stressed based on a cutoff of $13 \times 6/10$.[8] Second, we create similar versions of the PSS and CES-D-10 scores based only on the items in each index that should be reverse-scored, adjusting diagnosis cutoffs similarly.

Table 2 shows that despite comprising a minority of the stress and depression indices, reverse-scored items drive over 70% of VEA24's estimates at the ten-month endline. At this second endline, the standardized treatment effect estimates from indices constructed only using straight-scored items attenuate by 72-101% compared to the respective standardized estimates in Table 1's Panel A. Panels A and B in Figure 3 visualize mental health outcomes by treatment group and survey wave when indices are built using only straight-scored items. Treated participants still exhibit statistically significant reductions in stress and depression compared to control participants at the one-month endline. However, these differences are much smaller than those observed in the indices that include reverse-scored items, and are no longer statistically significantly different from zero by the ten-month endline.

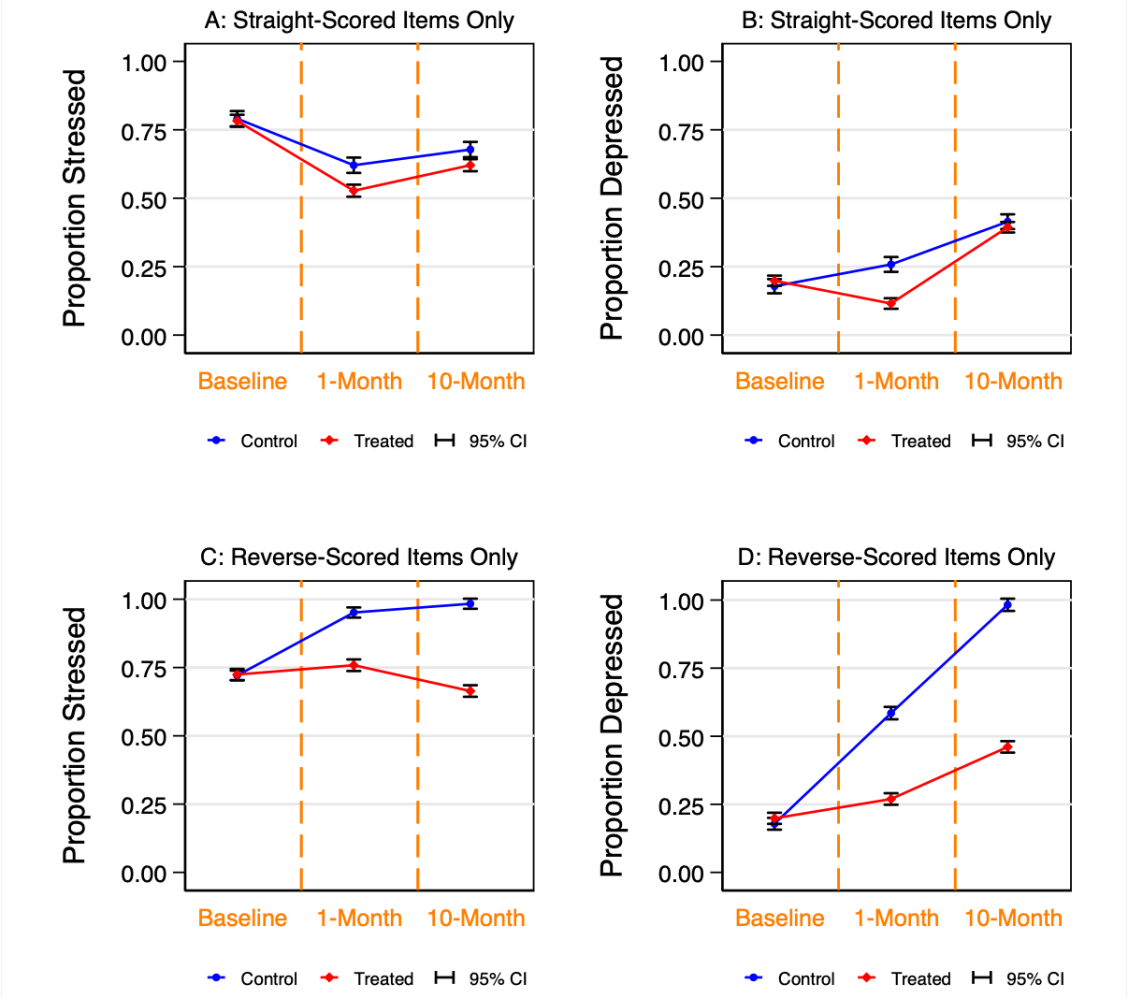
The vast majority of the magnitude of VEA24's treatment effect estimates is found only in the reverse-scored items in each scale. Panels C and D in Figure 3 show estimated mental health trajectories by treatment group when indices are built only using items that should be reverse-scored. Across both endlines, the gaps in stress and depression between participants in the treatment and control groups become far larger for indicators based only on reverse-scored scale items. In particular, when diagnosis indicators are built only using reverse-scored items, the control group appears to experience significant increases in stress and depression over time that are not seen in the treatment group.

[8] Here we retain VEA24's inconsistent indicator coding, as discussed in Section 3.1.

| | One-Month Endline Outcomes | | | | Ten-Month Endline Outcomes | | | |
|---|---|---|---|---|---|---|---|---|
| | Standardized PSS Score | Standardized CES-D-10 Score | Stress Diagnosis | Depression Diagnosis | Standardized PSS Score | Standardized CES-D-10 Score | Stress Diagnosis | Depression Diagnosis |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| **Panel A: Straight-Scored Items Only** | | | | | | | | |
| Treatment Effect, Without Covariates | -0.143 | -0.425 | -0.088 | -0.142 | -0.013 | -0.062 | -0.056 | -0.02 |
| | (0.054) | (0.049) | (0.027) | (0.022) | (0.069) | (0.062) | (0.028) | (0.029) |
| | [0.009] | [0] | [0.001] | [0] | [0.849] | [0.32] | [0.047] | [0.494] |
| | {0.049} | {0} | {0.018} | {0} | {0.852} | {0.567} | {0.128} | {0.705} |
| Treatment Effect, With Covariates | -0.144 | -0.444 | -0.089 | -0.15 | 0.009 | -0.051 | -0.049 | -0.016 |
| | (0.053) | (0.048) | (0.026) | (0.022) | (0.067) | (0.061) | (0.028) | (0.028) |
| | [0.007] | [0] | [0.001] | [0] | [0.889] | [0.4] | [0.081] | [0.582] |
| | {0.042} | {0} | {0.01} | {0} | {0.897} | {0.661} | {0.199} | {0.795} |
| **Panel B: Reverse-Scored Items Only** | | | | | | | | |
| Treatment Effect, Without Covariates | -1.015 | -0.771 | -0.19 | -0.316 | -1.091 | -1.983 | -0.319 | -0.519 |
| | (0.053) | (0.047) | (0.019) | (0.024) | (0.076) | (0.058) | (0.017) | (0.021) |
| | [0] | [0] | [0] | [0] | [0] | [0] | [0] | [0] |
| | {0} | {0} | {0} | {0} | {0} | {0} | {0} | {0} |
| Treatment Effect, With Covariates | -0.986 | -0.761 | -0.182 | -0.315 | -1.078 | -1.973 | -0.318 | -0.518 |
| | (0.052) | (0.046) | (0.019) | (0.024) | (0.078) | (0.059) | (0.017) | (0.021) |
| | [0] | [0] | [0] | [0] | [0] | [0] | [0] | [0] |
| | {0} | {0} | {0} | {0} | {0} | {0} | {0} | {0} |
| Observations | 2220 | 2220 | 2220 | 2220 | 2254 | 2254 | 2254 | 2254 |

*Note*: Treatment effect estimates on the outcome displayed in the column title are shown along with standard errors clustered at the village level in parentheses, raw $p$-values in straight brackets, and Westfall-Young family-wise error rate-controlled $p$-values in curled brackets (Westfall and Young 1993). Standardized scores hold a mean of zero and a standard deviation of one in the control group. Treatment effects are based on baseline treatment status. For each panel and each treatment effect type (i.e., with covariates and without covariates), Westfall-Young $p$-values are calculated as in VEA24's Table 2 across all eight outcomes in the columns of this table, happiness, life satisfaction, future aspirations, and COVID-19 compliance. All models control for union fixed effects, and all models on PSS scores and diagnosis indicators control for baseline PSS scores and diagnosis indicators (respectively). Additional covariates include respondents' age, education, and occupation, the occupation of respondents' husbands, a dummy indicating whether the respondent reports that their household lost income during the COVID-19 pandemic, a dummy indicating whether the respondent reports that their household chores increased after the COVID-19 lockdowns, a dummy indicating whether the respondent is the head of their household, the number of people in the respondent's household, and the number of children in the respondent's household.

Table 2: Treatment Effect Estimates, Different Item Types




*Note:* Participants are classified as 'stressed' and 'depressed' based on the diagnosis cutoff scheme for sensitivity analysis discussed earlier in this section. Means by baseline treatment status and survey wave are displayed along with 95% confidence intervals, based on VEA24's original item coding scheme (see Section 3.2).

Figure 3: Mental Health of Treatment Groups, Different Item Types

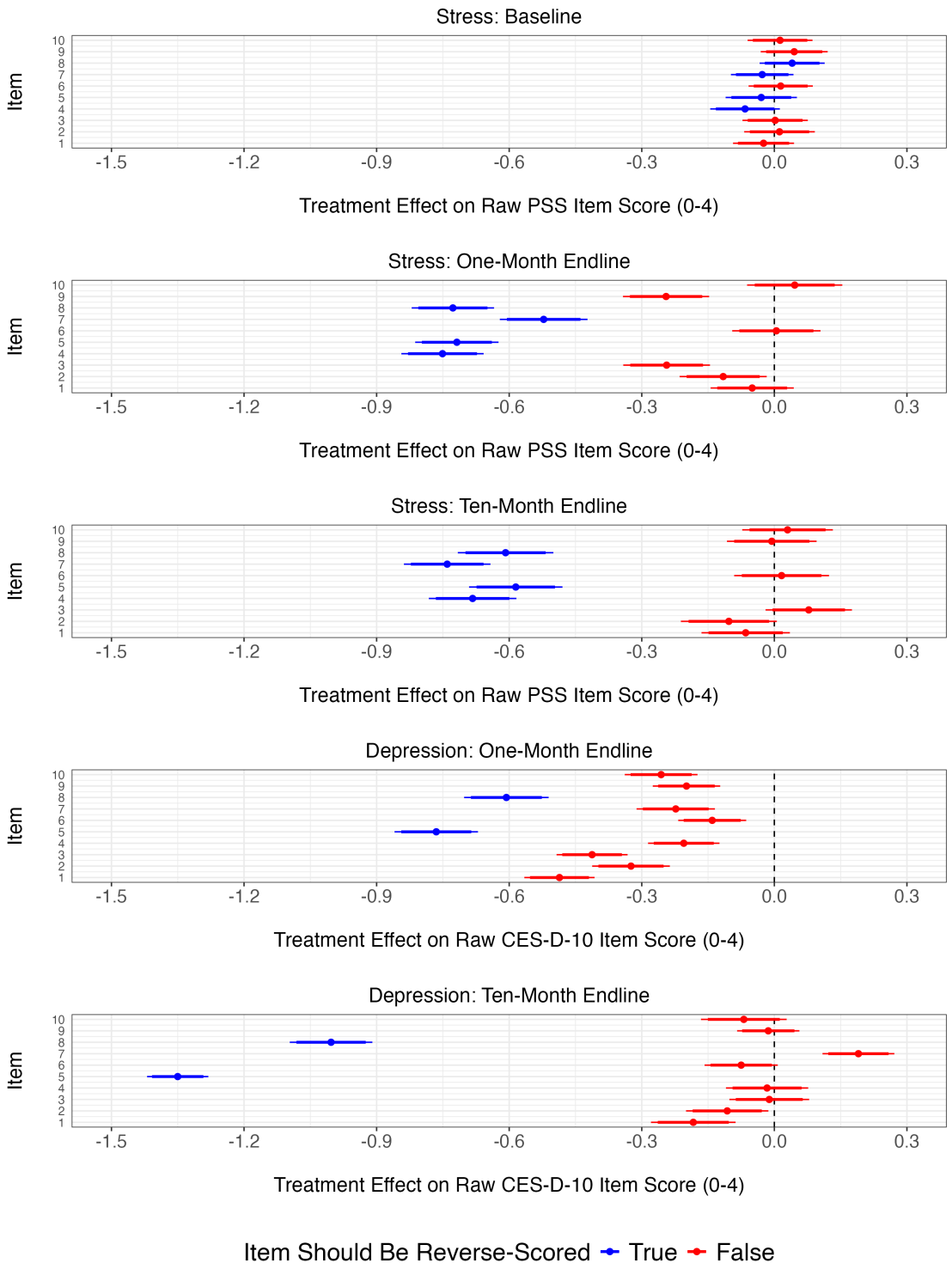
Panel B in Table 2 shows that compared to straight-scored items, reverse-scored items are disproportionately sensitive to treatment. Across both endlines, Panel B's treatment effect estimates in Table 2 on outcomes constructed exclusively from reverse-scored items are always larger than Panel A's respective estimates on outcomes constructed only from straight-scored items, being larger by 71-11,480%. In contrast, Table 2's Panel A estimates are much smaller than the main estimates in Table 1, with none of those estimates in Table 2's Panel A being statistically significantly different from zero at the ten-month endline.

Figure 4 visualizes the reverse-scored items' disproportionate sensitivity to treatment, displaying treatment effect estimates for each PSS and CES-D-10 item across each survey wave. Endline treatment effect estimates on straight-scored items cluster around zero, particularly in the ten-month endline. However, the reverse-scored items exhibit persistently and unrepresentatively large treatment effects.

These disproportionately sensitive items also exhibit far larger treatment effect estimates than thematically similar items in the same ten-month endline survey. VEA24's Online Appendix B.2 contains the item texts for the PSS and CES-D-10 scales. Consider questions eight and ten in the PSS scale. Question eight is supposed to be reverse-scored, reading "How often have you felt that you were on top of things?" Though question ten is not supposed to be reverse-scored, it is thematically similar, reading "How often have you felt difficulties were piling up so high that you could not overcome them?" Figure 4 shows that though the estimated treatment effect on responses to PSS question ten is an insignificant 0.03 points out of four, the same estimated treatment effect on responses to PSS question eight is a highly significant -0.6 points out of four, a 20-fold increase in effect size. Likewise, consider items three and eight on the CES-D-10 scale. Item three ("I was depressed") is not supposed to be reverse-scored, whereas item eight ("I was happy") is. At the ten-month endline, though treatment is estimated to decrease participants' agreement with the statement "I was depressed" by just 0.012 points out of four, it is also estimated to increase participants' agreement with the statement "I was happy" by a full point out of four. This is an 87-fold increase in effect size between two items that ask functionally identical questions at the exact same time point. The key difference is that one is subject to an undisclosed deviation in data collection procedures.

### 3.5 Other Empirical Irregularities

The Online Appendix documents many other empirical irregularities related to inconsistently reverse-scored variables. In Online Appendix A, we show that the time-




*Note:* Treatment effect estimates on scores for each item in the respective scale are displayed alongside 90% and 95% confidence intervals in thicker and thinner bands (respectively). Items that should be reverse-scored are reverse-scored in line with VEA24's original coding strategy. All estimates arise from linear regression models that regress raw item scores on baseline treatment status and union fixed effects, with standard errors clustered at the village level.

Figure 4: Treatment Effects on Raw Item Scores

intensive parental investment index treats missing values for reverse-scored questions as maximum investment instead of minimum, due to Stata's `rowtotal()` function treating missing values as zeros. Estimated effects on time-intensive parental investment are not robust to reasonable corrections for this issue. In Online Appendix B, we show that though treatment is estimated to increase the probability of recently starting new income-generating activity by over 13 percentage points after ten months, it also unexpectedly raises the probability of participants reporting that their control over their own income is 'not applicable' by 20 percentage points after one month. In Online Appendix C, we show that though treatment is estimated to significantly decrease participants' agreement that 'most decisions *in the home* should be made by men', treatment is also estimated to significantly increase participants' agreement that 'most decisions *in society* should be made by men.' This issue is obscured by VEA24's dichotomization of score variables, and in the paper, VEA24 misinterpret the coefficient sign in their Online Appendix table. Finally, in Online Appendix D, we find that the coefficients in VEA24's Figures 3 and 5 are incorrect due to a copy-pasting error from VEA24's Table 2 into the .dta file used to create the figures.

## 4 Connections to Other Papers

VEA24's data is connected to a wide range of prior papers and experiments. In Sections 4.1, 4.2, and 4.3, we document that the data from this project is connected to data used in several prior investigations in which both Asad Islam and GDRI took part. The vast majority of VEA24's data is sourced from two prior experiments (Guo et al. 2024, Ahmed et al. 2024), and VEA24's sample overlaps completely with another 2021 publication in *PLOS One* despite completely different descriptions of data collection between the two papers (Rahman, Hasnain and Islam 2021). In Section 4.4, we show that our reverse-scoring adjustment from Section 3.2 only reverses findings at the second endline for the sample originating from Ahmed et al. (2024), and does not sign-flip results at the second endline for the sample originating

from Guo et al. (2024). In Section 4.5, we find that estimated treatment effects often significantly differ depending on which sample the data is sourced from.

Disclosures about these sample connections across papers are vague and contradictory. From pg. 429:

> "To select our study sample from a list of households previously surveyed by GDRI, we first narrowed it down to households that meet the following criteria: (i) the household has a mobile phone number according to GDRI records, (ii) the phone number is valid, and (iii) the household has at least one adult (18 or above) female household member... Our partner NGO, a nonprofit research organization, has a directory of households who, in the past ten years, participated in surveys and RCTs conducted by the NGO in this region. We randomly selected our households from this directory, which is not by design representative of rural households in the region."

Any cross-paper connections are in direct conflict with this paper's AEARCT pre-registration, which specifies that "This trial does not extend or rely on any prior RCTs."[9] However, the AEARCT pre-registration is itself in conflict with the paper's ANZCTR pre-registration despite both of them being registered on the same day (15 July 2020).[10] The ANZCTR pre-registration states that

> "The baseline data have been collected by GDRI who are our local research partner in Bangladesh. The data were collected for different research and were shared the data with us [*sic*]. Please note that this data has been used as baseline data for this trial. As the survey was done for another study we are unable to change the date of participant recruitment."

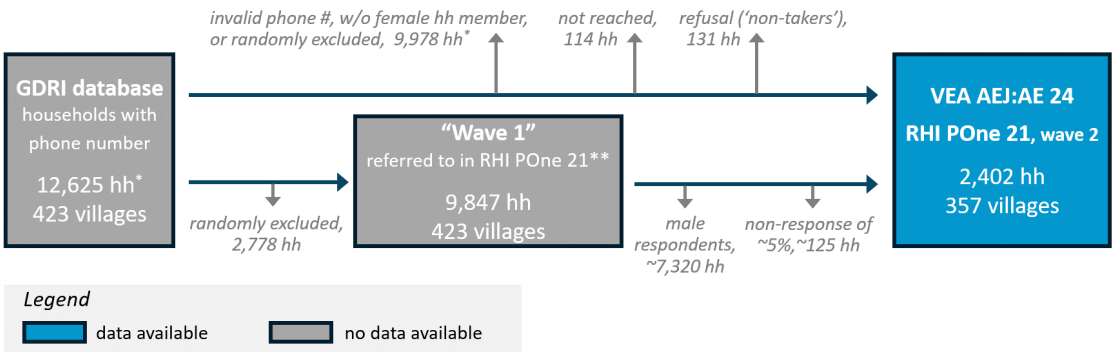
[9] AEARCT registration: https://www.socialscienceregistry.org/trials/5948.
[10] ANZCTR registration: https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=380128

### 4.1 Rahman, Hasnain, & Islam (2021, *PLOS One*)

The VEA24 sample is virtually identical to that used in Rahman, Hasnain and Islam (2021) (henceforth RHI21). RHI21 report a positive relationship between food insecurity and stress amongst Bangladeshi women during the COVID-19 pandemic, using the same PSS questionnaire and Food Insecurity Experience Scale as VEA24 (Ballard et al. 2013). The samples of 2402 women in both RHI21 and VEA24 can be linked by variable `RECORD_ID`, and these women share *identical* values of PSS scores, COVID-19 awareness, number of household members, and food insecurity (in one of the two waves collected by RHI21).

Figure 5 illustrates how the two papers describe observations' flow from the GDRI database to the final samples of VEA24 and RHI21, which is inconsistent across papers. VEA24 state that they started from the GDRI database, specifically a list of households that have a valid phone number and at least one female adult member (aged 18 or above). From the list, 2647 households were reportedly randomly selected, 2533 could be reached, and 2402 eligible women were enrolled. In contrast, RHI21 claim to have started with "a panel of households interviewed during the pre-COVID-19 period" (RHI21, pg. 4), randomly sampling 9847 households in the first wave. RHI21 do not explicitly name the GDRI database as a data source. Roughly 7320 men were then reportedly dropped, and 95% of the remaining women from the first wave were re-surveyed in the second wave. After accounting for this attrition, the final sample again consists of 2402 women from the same households. These are two fundamentally different sampling procedures that reportedly produce the exact same sample.

Furthermore, the two papers are inconsistent in reporting when specific questions were asked. What VEA24 describe as baseline data collected in May 2020 is instead reported in RHI21 as being partly elicited in their first wave (April/May 2020) and partly in their second wave (three to four weeks later). Specifically, food insecurity was collected in the first wave, while stress and COVID-19 awareness were elicited in the second wave.




*Note:* Diagrams are based on the sampling descriptions in VEA24 and RHI21. hh abbreviates households. * = these numbers are not explicitly reported, but derived from information on the GDRI database presented in Ahmed et al. (2021). ** = RHI21 and Ahmed et al. (2021) report results on these 9847 households from Wave 1.

Figure 5: Sampling Inconsistency Between VEA24 and RHI21

### 4.2 Ahmed, Hodler, & Islam (2024, *Economic Journal*)

Ahmed et al. (2024) – henceforth AHI24 – report the results of a randomized get-out-the-vote campaign conducted in the Khulna and Satkhira districts of Bangladesh in 2018 (AHI24, pg. 1309). These are the same districts studied in VEA24 (pg. 429). AHI24 find that their get-out-the-vote treatments have null overall effects on voter turnout, but have significant positive effects in villages that are strongholds for the Awami League and significant negative effects in villages that are strongholds for the Bangladesh Nationalist Party.

Drawing on AHI24's replication data (Ahmed et al. 2023), we can source at least 1192 of the 2402 individuals in our study as participants of AHI24. Online Appendix Figure A1 displays a waterfall plot that details the variables on which observations can be matched. 1483 observations in VEA24 share the same values of `RECORD_ID` as observations in AHI24, and at least 1480 of these share the same values of household incomes, household sizes, and education levels for the husband and wife who head the household. The left graph shows that most observations do not match on raw age values. Though one may reasonably attribute this to natural aging, this is actually due to an error in transferring ages from the AHI24 data into the VEA24 data. Online Appendix E details both the matching procedure between VEA24 and AHI24, as well as a correction procedure for these age transfers.

The right graph in Online Appendix Figure A1 shows that after correcting the age transfers, all 1480 individuals who match on household identifiers and other sociodemographic variables also match on age. An additional 288 do not match on village identifiers. This is related to a disclaimer provided by VEA24 in repository file `README_raw_to_final.docx`:

> "There are some minor problems in the variable Village_ID of the generated full raw dataset "Telecounseling_data_original.dta". However, we did not use this variable in our final estimation. Therefore, for replication, please ignore this variable."

Even ignoring the 288 observations whose village identifiers do not match, after age transfer corrections, we can match 1192 observations between VEA24 and AHI24.

### 4.3 Guo et al. (2024)

Guo et al. (2024) – henceforth GEA24 – report the results of a large randomized controlled trial studying two early childhood educational programs over a period of two years. One program offered daily formal preschool education, while the other involved weekly home visits by trained teachers to improve parenting practices. This trial was conducted on a sample of over 6900 children in 222 villages in rural areas of Bangladesh's Khulna and Satkhira districts (GEA24, pg. 5), the same districts as in VEA24 (pg. 429). Parents were also sampled in household surveys. The paper reports that both programs significantly improved children's cognitive and non-cognitive development and had significant positive spillover effects.

Online Appendix Figure A2 displays a waterfall plot that details the variables on which individuals can be matched between the VEA24 and GEA24 samples. 678 women in VEA24 share the same values of `RECORD_ID` as individuals in GEA24. Of these individuals, 671 can be matched to GEA24's 2017 socioeconomic survey. Seven individuals were not surveyed in 2017, but can be matched to the 2019 socioeconomic survey. At least 664 additional observations share the same values of village identifier, household incomes, household sizes, and both age and education

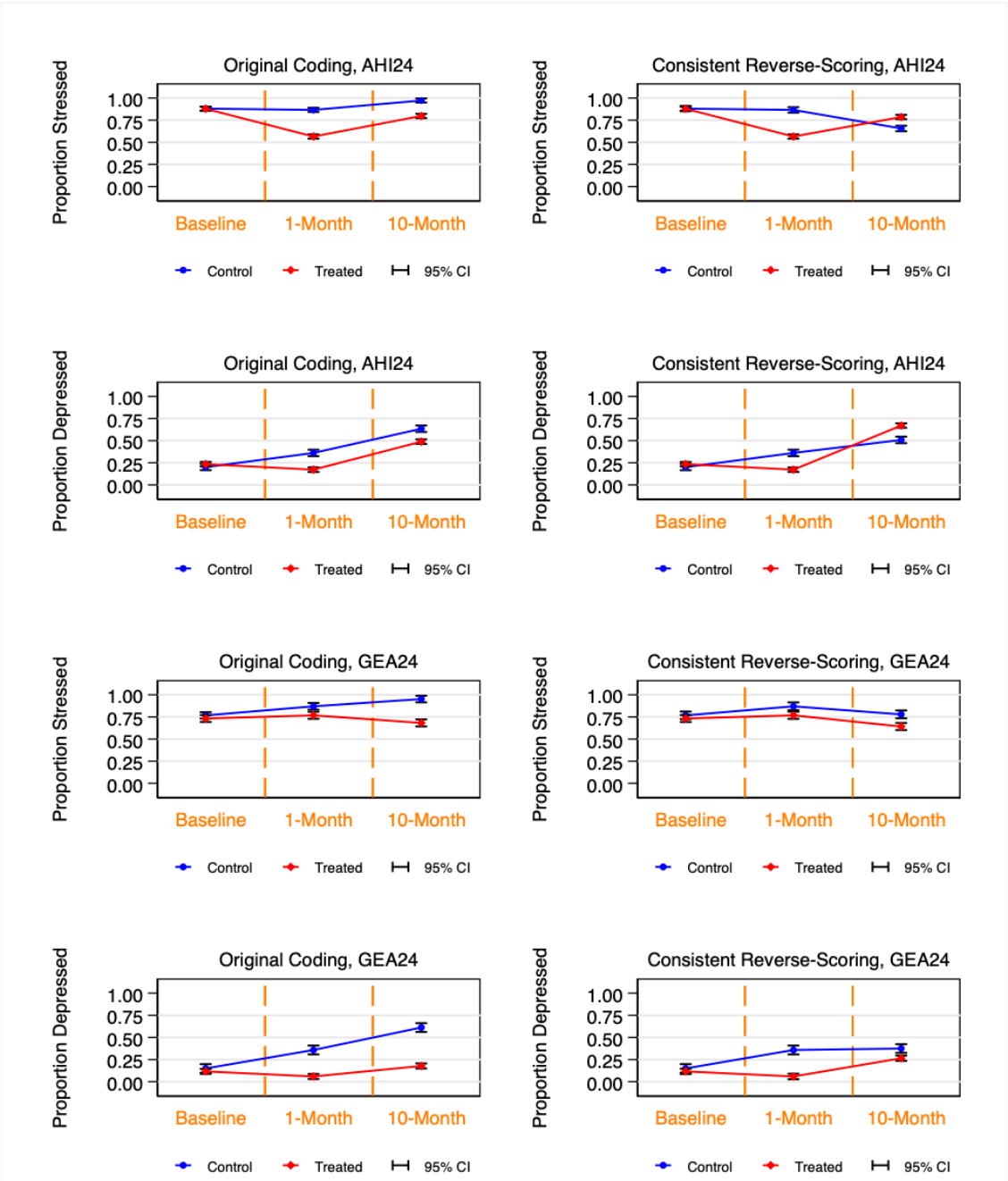
levels of both male and female heads of household. There is no overlap between the respondents matching to AHI24 and those matching to GEA24.

This matching indicates that information about household income, parent's age, and household size for a significant part of the sample (671 individuals) was already collected in 2017, contradicting VEA24's claims regarding baseline data collection in the paper. From pg. 430:

> "Between the end of May and the middle of June 2020, GDRI (the local NGO we collaborated with) surveyed the enrolled women over the phone to understand their physical and emotional state during the pandemic. Through this survey, trained enumerators (a different set of individuals from the para-counselors who carried out the intervention) gathered baseline information on some household demographics, socioeconomic characteristics, food insecurity, participants' knowledge and perception of COVID-19, how often they comply with COVID-19 health guidelines, their worries and fears, health and well-being, and their stress levels (Vlassopoulos et al. forthcoming)."

### 4.4 Differences in Reverse-Scoring Between Data Sources

The reverse-scoring corrections applied throughout Section 3 have fundamentally different impacts on VEA24's outcomes depending on whether an observation originated from AHI24 or from GEA24. Figure 6 shows how reverse-scoring corrections differentially affect estimated trajectories of, and treatment effects on, mental health outcomes in the samples of VEA24 arising from AHI24 and GEA24. As in the full VEA24 sample, estimated treatment effects on stress and depression sign-flip at the ten-month endline in the AHI24 sample after we apply our reverse-scoring corrections. However, though estimated treatment effects on stress and depression at the ten-month endline considerably attenuate in the sample arising from GEA24 after applying reverse-scoring, estimated treatment effects do not flip signs in this subsample. This implies that relative to straight-scored survey items, reverse-scored




*Note*: Participants are classified as 'stressed' and 'depressed' based on VEA24's original diagnosis cutoff scheme, as discused in Section 3.1. Means by baseline treatment status and survey wave are displayed along with 95% confidence intervals. Samples are restricted to either the subsample arising from AHI24 or the sample arising from GEA24.

Figure 6: Sampling Inconsistency Between VEA24 and RHI21

survey items are systematically more important for the subsample emerging from AHI24 than for the subsample arising from GEA24.

| | Standardized Stress, One-Month Endline, Original Scoring, (1) | Standardized Depression, One-Month Endline, Original Scoring, (2) | Standardized Stress, Ten-Month Endline, Original Scoring, (3) | Standardized Depression, Ten-Month Endline, Original Scoring, (4) | Standardized Stress, Ten-Month Endline, Consistent Scoring, (5) | Standardized Depression, Ten-Month Endline, Consistent Scoring, (6) |
|---|---|---|---|---|---|---|
| **Panel A** | | | | | | |
| Treatment | -0.852 | -0.598 | -0.507 | -0.31 | 0.833 | 0.689 |
| | (0.076) | (0.067) | (0.096) | (0.077) | (0.079) | (0.072) |
| | [0] | [0] | [0] | [0] | [0] | [0] |
| GEA24 Sample | 0.035 | -0.012 | 0.056 | 0.139 | 0.378 | 0.024 |
| | (0.085) | (0.079) | (0.09) | (0.076) | (0.082) | (0.07) |
| | [0.684] | [0.878] | [0.535] | [0.069] | [0] | [0.729] |
| Treatment × GEA24 Sample | 0.51 | -0.145 | -0.241 | -0.754 | -0.944 | -0.803 |
| | (0.116) | (0.094) | (0.15) | (0.118) | (0.117) | (0.103) |
| | [0] | [0.122] | [0.11] | [0] | [0] | [0] |
| *N* | 2220 | 2220 | 2254 | 2254 | 2254 | 2254 |
| **Panel B** | | | | | | |
| Treatment | -0.432 | -0.585 | -0.655 | -0.752 | 0.32 | 0.197 |
| | (0.076) | (0.061) | (0.108) | (0.093) | (0.089) | (0.073) |
| | [0] | [0] | [0] | [0] | [0] | [0.007] |
| AHI24 Sample | -0.037 | 0.102 | 0.249 | 0.276 | -0.114 | 0.278 |
| | (0.089) | (0.082) | (0.096) | (0.087) | (0.093) | (0.077) |
| | [0.678] | [0.215] | [0.01] | [0.002] | [0.221] | [0] |
| Treatment × AHI24 Sample | -0.449 | -0.09 | 0.115 | 0.353 | 0.393 | 0.416 |
| | (0.115) | (0.097) | (0.153) | (0.128) | (0.121) | (0.107) |
| | [0] | [0.353] | [0.452] | [0.006] | [0.001] | [0] |
| *N* | 2220 | 2220 | 2254 | 2254 | 2254 | 2254 |

*Note*: Treatment effect estimates on the outcome displayed in the column title are shown along with standard errors clustered at the village level in parentheses and raw $p$-values in straight brackets. Standardized scores hold a mean of zero and a standard deviation of one in the control group. Treatment effects are based on baseline treatment status. All models control for union fixed effects, and all models on PSS scores control for baseline PSS scores.

Table 3: Treatment Effect Interactions With Original Data Sources

### 4.5 Differences in Treatment Effect Estimates Between Data Sources

Treatment effect estimates are also extremely sensitive to the source from which observations originate. Table 3 shows that the GEA24 and AHI24 samples exhibit systematically different treatment effects than the rest of the sample. Interaction effects between treatment and dummies indicating an observation's origination from one of the two samples are statistically significant for stress at the one-month endline and at the ten-month endline (after applying consistent reverse-scoring), and for depression at the ten-month endline (regardless of reverse-scoring consistency). In these models, the effect sizes of estimated interaction effects between treatment and sample origination range from 0.353 to 0.944 standard deviations, and the interaction effect estimates in these models take sizes ranging from 47-243% of the main treatment effect estimates for observations outside the sample.

### 4.6 Siddique et al. (2024, *Review of Economics and Statistics*)

A contemporary reproduction report (Kjelsrud et al. 2025) details how the sample of VEA24 partially overlaps with that of Siddique et al. (2024) (henceforth SEA24), and how the individuals in the two studies are subject to two overlapping RCT designs on similar topics measured with similar outcome variables. Namely, SEA24 study how voice and text messages improve compliance with COVID-19 guidelines. This overlaps substantially with VEA24, who provide COVID-19 prevention advice and test for knowledge of and compliance with these prevention guidelines.

Kjelsrud et al. (2025) explain in detail how the baseline of VEA24 falls between the treatment and endline of SEA24, while the treatment period of VEA24 partially falls between the endlines of SEA24 (see Figure 2 from Kjelsrud et al. (2025) for further information). This alignment allows Kjelsrud et al. (2025) to investigate partial treatment effects and spillover effects on "new" data. The results, reported in Table 4 of Kjelsrud et al. (2025), indicate that those belonging to the 'voice only' treatment in SEA24 exhibited significantly smaller increases in compliance with COVID-19 guidelines when provided the intervention in VEA24. This suggests that there are meaningful spillovers between the two experiments. However, beyond broad disclosures about samples being recruited from prior experiments conducted by GDRI, SEA24 is not disclosed in VEA24, and *vice versa*.

## 5 Documentation Inconsistencies

This section focuses on inconsistencies between the replication data and documentation in both VEA24's pre-registrations and Online Appendix. VEA24's title page footnote points to two pre-registrations in the American Economic Association Randomized Controlled Trial (AEARCT) Registry and the Australian New Zealand Clinical Trials Registry (ANZCTR).[11] Both pre-registrations were simultaneously registered on 15 July 2020, before the collection of the one-month endline survey in

[11] AEARCT registration: https://www.socialscienceregistry.org/trials/5948. ANZCTR registration: https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=380128. Both URLs are accessed on 12 January 2024.

August 2020. The pre-analysis plan can be found in AEARCT Registry document `PAP_covid_mental_health_aearegistry.pdf`.

### 5.1 Mismatches in Survey Text

In both endline surveys, the main stress indices are constructed using Likert items that are not scored monotonically in agreement with the survey item. At baseline, repository file `Baseline_kobo form.xls` shows that the PSS questions `Q12_A` through `Q12_J` are scored such that values zero through four respectively correspond to options "Never", "*Mostly never*", "Sometimes", "Often", and "Frequently" (emphasis added). However, `Endline1_kobo form.xls` and `Endline2_kobo form.xls` show that respective PSS items `end_Q5_1` through `end_Q5_10` and `MH_1_1` through `MH_1_10` are scored such that values zero through four respectively correspond to options "Never", "*It goes without saying*", "Sometimes", "Often", and "Frequently" (emphasis added). Our official translations show that similar discrepancies emerge in the Bengali text, with scores of one on the baseline and endline scales respectively translating to "Most of the time it is not" and "It goes without saying."

For many survey items, backtranslations of the Bengali text from the raw survey files also show disparities between the question text in the raw survey files and the question text presented in VEA24's Online Appendix. Our OSF repository at https://osf.io/pvkhy/ contains official translations of the 'hint' and 'Bangla' columns in `Baseline_kobo form.xls`, `Endline1_kobo form.xls`, and `Endline2_kobo form.xls`. Comparing these backtranslations with the English texts in VEA24's Online Appendix B shows large disparities in question content for many scales in the paper that are validated in English. VEA24 make no mention of backtranslation efforts to ensure the accuracy of Bengali questions in the surveys.

### 5.2 Undisclosed Deviations from the Pre-Analysis Plan

As noted in Section 4, VEA24 are inconsistent in their disclosure regarding participant overlap with other studies. The sample is also chosen without regard to three

pre-registered inclusion criteria, which include participant age, food insecurity, and pre-existing mental health conditions. For details, see Appendix F.

### 5.3 Missing Data

Two sets of pre-registered data are missing from the replication repository's raw data, the first of which is data on physical health. Pages 5-6 of the pre-analysis plan state that participants are asked whether they have experienced each of ten common symptoms of illness in the past 15 days, with dummy variables recorded for each 'yes' response. Similar yes/no questions were supposed to be asked regarding the children and other adults in the household. VEA24 committed to dichotomizing these responses by creating dummies indicating if more than half of the symptoms were reported, separately for the participant, their children, and other adults. None of this data is available in the replication repository.

Though VEA24 claim that this physical health data was collected and dropped, the raw survey forms do not provide any space in which this data could have been collected. From pg. 431:

> "We also preregistered *Physical health* of the respondents, children, and other household members (measured using questions on the prevalence of common COVID-19 symptoms) as a health outcome but dropped it at endline because all respondents and household members did not report any symptoms at baseline."

The raw survey forms `Baseline_kobo form.xls`, `Endline1_kobo form.xls`, and `Endline2_kobo form.xls` contain no questions eliciting any of the symptoms discussed in the pre-analysis plan for any survey wave.

The second set of missing data is the randomization files. According to the AEARCT Registry, treatment was randomized using Stata. The replication repository does not contain any treatment randomization script.

## 6 Irregularities in the Raw KoboToolbox Data

The repository file "README_raw_to_final.docx" states that surveys are conducted using KoboToolbox, a free platform for creating surveys that can be collected using a web form or via the Kobo Collect app on smartphones and tablets. Two sets of files are reported to arise from these surveys. The first are the survey forms `Baseline_kobo form.xls`, `Endline1_kobo form.xls`, and `Endline2_kobo form.xls`, which respectively correspond to the baseline, first endline, and second endline surveys. These forms are in the XLS Form format used for generating and saving surveys in KoboToolbox.[12] The second set of files are the raw survey data files `Baseline_raw.xlsx`, `Endline1_raw.xlsx`, and `Endline2_raw.xlsx`, which again respectively correspond to the baseline, first endline, and second endline surveys. Repository file `README_raw_to_final.docx` describes each of these three files as being original data that is "extracted from KoboToolbox".

### 6.1 Irregular English Labels in the Survey Forms

Some question labels in `Baseline_kobo form.xls`, `Endline1_kobo form.xls`, and `Endline2_kobo form.xls` – which correspond to question text displayed in the respective Kobo surveys – are cut off before the sentence would logically end, or even mid-word. Official translations show that similar cutoffs also appear in Bengali text from the raw survey forms for some questions. E.g., question `MH_1_10` – the tenth of the PSS stress index at the ten-month endline – has English label "How often do you feel that the problem is getting worse so much that you can't g" [*sic*]. The official translation of the Bengali label is "How often do you feel that the problem is getting so bad that you can't go to the gym?" Neither corresponds to the actual tenth item in the PSS survey, which is "How often have you felt difficulties were piling up so high that you could not overcome them?" Likewise, question `Food_2_7` in the ten-month endline survey – the seventh item in the Food Insecurity Experience Scale – is titled in English: "In the last 2 to 3 weeks, someone

[12] See https://support.kobotoolbox.org/edit_forms_excel.html.

in your family was hungry but could not buy fo" [*sic*]. The official translation of the corresponding Bengali text to English is: "In the past 2 to 3 weeks, someone in your family was hungry but couldn't buy pho". Neither correspond to the actual seventh item of the Food Insecurity Experience Scale, which is "Has the following happened in the last 2-3 weeks? You or anyone in your family were hungry but you could not buy food due to lack of money?" (Ballard et al. 2013).

These cutoffs for English labels often correspond to Stata's 80-character limit for variable labels.[13] In contrast, the character limit for both a Kobo survey question and for an Excel cell is 32,767 characters.[14] Online Appendix Tables A8, A9, and A10 list the questions with labels that we identify as irregular in the baseline, one-month endline, and ten-month endline survey forms (respectively), either due to the text itself or because of known truncations of established survey items.[15] Table A8 documents six irregular question labels in the baseline survey form, including one component of the COVID-19 compliance index and five items in the PSS stress scale. Two of these questions' English labels are 79 or 80 characters long. Table A9 documents 16 irregular question labels in the one-month endline survey form, including seven of the eight items in the Food Insecurity Experience Scale, three of the ten items in the PSS scale, one item in VEA24's 'future aspirations' scale, and all five items in the COVID-19 confidence index. Table A10 identifies 18 irregular question labels in the ten-month endline survey form, including four items in the PSS scale, seven items in the Food Insecurity Experience Scale, one of the ten items in the 'following advice' index, and all items underpinning VEA24's variables on borrowing money, contacting public offices, husband's workload, social desirability bias, risk preferences, and social preferences. All irregular question labels that we identify in both endline survey forms are 79 or 80 characters long.

[13]Source: https://www.stata.com/manuals13/dlabel.pdf.

[14]Source: https://support.microsoft.com/en-us/office/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3.

[15]For instance, question `Q12_D` reads "In last 7 day- how often have you felt confident about your ability" [*sic*]. This reads as a plausible sentence, and is only 67 characters long. However, this question corresponds to the fourth question in the PSS survey, which VEA24's Online Appendix B.2 writes as "In the past 7 days... how often have you felt confident about your ability to handle your personal problems?"

### 6.2 Ordering Irreguralities in `sharedStrings` in Raw Data Excel Files

The .xlsx files in which the raw data is stored allow us to use `sharedStrings` files to track the order of data entry into the spreadsheets. .xlsx files are actually zipped archives of multiple .xml files, which store the document's underlying data and properties[16]. These files can be accessed by unzipping the .xlsx file.[17]

In .xlsx files containing strings (e.g., column titles), there are two ways in which an .xlsx file can save strings contained in the spreadsheet. The first option is to use inline strings, where string data are recorded directly in each cell in the sheet's `.xml` file in the folder `xl/worksheets`.[18] A second option is to use `sharedStrings.xml` in folder `xl` to store the strings. The strings are then linked to the cell in the sheet's `.xml` file in the folder `xl/worksheets`.[19]

Saving strings as `sharedStrings` is the default behavior of the Microsoft Excel desktop app as well as some other programs, such as the `export excel` function in Stata. On the other hand, some software libraries choose not to store string data in `sharedStrings`. Based on our testing, KoboToolbox appears to be among the software that does not export data into `sharedStrings` by default. However, when a file not containing `sharedStrings.xml` is opened and (auto)saved using the Microsoft Excel desktop app, file `sharedStrings.xml` will be created.

When manually editing documents using Microsoft Excel, strings from the .xlsx document are recorded in `sharedStrings.xml` in the order in which these strings were first entered into the spreadsheet. In contrast, in the event that strings are entered into an .xlsx file programatically (e.g., by Stata), or when they are encoded by Excel for the first time, strings appear in `sharedStrings.xml` ordered by their row-column or column-row position in the spreadsheet; the specific order differs by

[16] For applications in research, see https://datacolada.org/109.

[17] This can be done by changing the .xlsx file's extension from .xlsx to .zip; on Microsoft devices, one can directly open the resulting .zip file by pointing and clicking, whereas on Apple devices, one must type `unzip "path-to-excel-filename.zip" -d extracted_folder` into the Terminal to access the file's contents.

[18] Source: https://learn.microsoft.com/en-us/dotnet/api/documentformat.openxml.spreadsheet.cell

[19] This is done to save space and optimize speed when working with data containing repeated strings; see https://learn.microsoft.com/en-us/office/open-xml/spreadsheet/working-with-the-shared-string-table.

software. If new strings are entered manually into a programatically-constructed .xlsx file, then those new strings appear in the `sharedStrings.xml` file after all the programatically-entered strings (in order of manual entry).

For the raw baseline data in file `Baseline_raw.xlsx`, the `sharedStrings.xml` file indicates that column titles 'Instruction' and 'participation' are out of order compared to what would be ordinarily expected from a programatically-constructed .xlsx file. Strings 'Instruction' and 'participation' appear only once each in the spreadsheet for `Baseline_raw.xlsx`, respectively in the first rows of columns K and L. Column K is empty except for first row 'Instruction', and column L is a numeric column except for first row 'participation'. There are no other strings in `Baseline_raw.xlsx` except for those in the first row (i.e., the column names). 'Instruction' and 'participation' should appear as the 11th and 12th strings in the `sharedStrings.xml` file for `Baseline_raw.xlsx` if `sharedStrings.xml` reflects the row-column or column-row order of strings in the spreadsheet with no additional modifications. However, these strings instead appear as the last two strings in the `sharedStrings.xml` file for `Baseline_raw.xlsx`.

### 6.3 Incompatibility of First Endline Survey Format

The first endline survey file `Endline1_kobo form.xls` has two questions with names `VILLAGE_ID` and `Village_ID`. Uploading `Endline1_kobo form.xls` as a survey form in KoboToolbox without adjustments and attempting to deploy the survey results in KoboToolbox refusing to deploy, displaying the following message:

> unable to deploy
>
> your form cannot be deployed because it contains errors:
>
> There are more than one survey elements named 'village_id' (case-insensitive) in the section named 'a7St4rGbGw3CsXemMYTVvK'.[20]

This implies that `Endline1_kobo form.xls` cannot be the form used to conduct the first endline survey in Kobo. If `Endline1_kobo form.xls` is created in Excel

[20] This last string is random and will change in subsequent replications.

and then uploaded to KoboToolbox, the survey will not deploy. If the survey is created in KoboToolbox using the Kobo Form Builder and exported to an XLS Form, then the question's name is changed automatically in the exported form to resolve this duplication, with "_001" appended to the question name. The variable names `VILLAGE_ID` and `Village_ID` cannot coexist in a deployed Kobo survey.

### 6.4 Languages Shown in the Kobo Survey

Forms in the XLS Form format used by KoboToolbox can accommodate multiple languages.[21] In the names of columns such as "label" (for question titles or answer labels) or "hint" (for smaller text displayed under the question title), languages can be indicated using the operator "::". For example, an XLS Form with columns "label::English" and "label::Bengali" uploaded to KoboToolbox will create a survey with two languages that can be selected from a dropdown menu; if "English" is selected from this dropdown menu, then question titles would display in English. If no heading contains the "::" operator, then no dropdown menu will exist in the Kobo survey to offer language choices.

The three XLS Forms provided in `Baseline_kobo form.xls`, `Endline1_kobo form.xls`, and `Endline2_kobo form.xls` lack operators for offering multiple languages. Each of these forms is divided into three sheets. First, sheet "survey" lists English and Bengali question titles in columns "label" and "hint" (respectively). Second, sheet "choices" contains English and Bengali text for answers to multiple choice questions in columns "label" and "Bangla" (respectively). Third, sheet "settings", an optional sheet that can be used to for version control, is empty for all three surveys. The missing "::" operator in the "survey" and "choices" sheets prevents language selection in KoboToolbox. As a result, surveys display questions in large English text with smaller Bengali hints, while multiple-choice answers appear only in English, with Bengali translations unused. We do not know if the survey enumerators were all proficient in English.

[21] See https://support.kobotoolbox.org/getting_started_xlsform.html, as well as https://xlsform.org/en/#multiple-language-support.

### 6.5 Consent, Skip Logic in the Baseline Form, and Missing Values

In `Baseline_kobo form.xls`, question `participation` records whether participants agree to participate in the survey. It is not the first question; ten precede it, seven of which may be elicited from participants.[22] Data is retained for these first ten questions in the public replication repository even for the 131 observations who did not consent. This includes information regarding age and education for the respondent and their spouse as well as information about the number of members, the number of children, and the income for respondents' households. Informed consent questions do not appear to be included in either `Endline1_kobo form.xls` or `Endline2_kobo form.xls`.

In the baseline survey, the variable `participation` is used in skip logic for later questions in the survey. In XLS Forms used for Kobo surveys, (groups of) questions can be designated to display only if a logical 'relevant' condition holds. `Baseline_kobo form.xls` shows that questions in the third question group only are displayed if `participation` $= 1$.[23] This skip logic is not included for questions in the fourth question group, and so questions in this group display regardless of whether `participation` $= 1$.[24]

This skip logic implies that there is missing raw data for survey questions that are both not skipped and required. In the `Baseline_kobo form.xls`, the 131 participants who did not consent should not encounter any questions in the third group, and responses for those questions are indeed all empty for these 131 participants. However, data is also missing for all questions in the fourth group for the same participants. If `Baseline_kobo form.xls` directly corresponds to the Kobo survey, and if `Baseline_raw.xlsx` is the raw data directly extracted from KoboToolbox

[22]The first three questions concern village ID numbers, unique respondent IDs, and union council identifiers.

[23]See row 18 in `Baseline_kobo form.xls`, which denotes a question type of `begin_group` with a hint of `Group-3`.

[24]See row 47 in `Baseline_kobo form.xls`, with question type `end_group`, and row 48 with question type `begin_group`, neither of which contain skip logic in the 'required' column. Uploading `Baseline_kobo form.xls` to KoboToolbox also produces a survey in which questions in the third group, but not the fourth group, are hidden when `participation` $= 0$.

for this survey, then this should not be possible. As previously discussed, questions in the fourth group are not skipped when `participation` = 0, and `Baseline_kobo form.xls` indicates that all but one of these questions is required. However, in the raw data, no data exists for any of the variables in the fourth question group.[25]

### 6.6 Impossible Values in the Raw Data

Several variables in the raw data contain values that are not possible to obtain from surveys corresponding to the Kobo forms in the replication repository. This can be true in two ways. First, questions of type `select_one` are multiple choice questions, where only one value can be selected from a list of predetermined options. Impossible values arise for one of these questions if a stored answer in the raw data is not among the available choices for the question. Second, a required question in a Kobo survey cannot be skipped, and the survey cannot be completed unless answers are provided for such questions. Finishing a Kobo survey requires that some data is inputted for all required questions. Impossible values arise for these questions if there are empty cells in the raw data where answers to these questions should be.

Four `select_one` questions in the baseline survey contain multiple choice answers that are not among the lists of available answers indicated in `Baseline_kobo form.xls`. Online Appendix Table A6 shows each of these four questions, along with possible answers for each question. The question `occ_respondent`, which records respondents' occupations, is listed by `Baseline_kobo form.xls` as having options 1-8 (corresponding to different occupations). In `Baseline_raw.xlsx`, variable `occ_respondent` contains zeros for 2059 observations and value 98 for 182 observations. Likewise, `occ_husband` records occupations of respondents' husbands (all participants in the survey are married women). `Baseline_kobo form.xls` lists `occ_husband` as having options 1-7. `Baseline_raw.xlsx` contains 48 observations where `occ_husband` = 0. Question `Q13_3` contains options 0-2 in `Baseline_kobo`

[25] This cannot be explained by the raw data files being cleaned of any data for observations who did not consent to data collection for this project. As aforementioned, data is still available for all questions asked prior to the asking of the `participation` question (including income, education, and household structure) across all participants for which `participation` = 0.

`form.xls`; seven observations in `Baseline_raw.xlsx` contain values of three for this variable. Finally, question `Q14_1` – a non-mandatory question asking participants whether they are afraid that they or their family members may be infected with COVID-19 – is listed in `Baseline_kobo form.xls` as having options zero (no), one (yes), 98 (I don't want to say) or 99 (don't know). `Baseline_raw.xlsx` lists integer values from 0-10 for this variable, and 2351 observations have values that are not possible given the list in `Baseline_kobo form.xls`.[26]

17 required questions across the two endline surveys contain missing values in the raw data, which should not be possible. Online Appendix Table A7 displays each such variable from these surveys, the survey it belongs to, and the number of observations for which data is missing for that variable. In the one-month endline survey, this includes four of the seven questions in the COVID-19 compliance scale, all of the questions that form the time-intensive parental investment scale, and all of the questions in the gender empowerment index. In the ten-month endline survey, this includes all of the items for the variables on borrowing money, contacting public offices for assistance, husband's workload, and time-intensive parental investments. For some variables, hundreds of values are missing.

## References

Ahmed, F., Hodler, R. and Islam, A.: 2023, Replication package for: "Partisan effects of information campaigns in competitive authoritarian elections: Evidence from Bangladesh". [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10059860.

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.

Ahmed, F., Islam, A., Pakrashi, D., Rahman, T. and Siddique, A.: 2021, Deter-

[26] This count excludes missing values, which are possible given that `Q14_1` is a non-mandatory question. We do not observe any missing values for observations for which `participation` = 1.

minants and dynamics of food insecurity during COVID-19 in rural Bangladesh, *Food Policy* **101**, 102066.

Andresen, E. M., Malmgren, J. A., Carter, W. B. and Patrick, D. L.: 1994, Screening for depression in well older adults: Evaluation of a short form of the CES-D, *Prev Med* **10**, 77–84.

Ballard, T. J., Kepple, A. W. and Cafiero, C.: 2013, The food insecurity experience scale: Development of a global standard for monitoring hunger worldwide. FAO Technical Paper.

Baranov, V., Bhalotra, S., Biroli, P. and Maselko, J.: 2020, Maternal depression, women's empowerment, and parental investment: Evidence from a randomized controlled trial, *American Economic Review* **110**(3), 824–859.

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

Bryant, R. A., Schafer, A., Dawson, K. S., Anjuri, D., Mulili, C., Ndogoni, L., Koyiet, P., Sijbrandij, M., Ulate, J., Harper Shehadeh, M. et al.: 2017, Effectiveness of a brief behavioural intervention on psychological distress among women with a history of gender-based violence in urban kenya: A randomised clinical trial, *PLoS Medicine* **14**(8), e1002371.

Cohen, S., Kamarck, T. and Mermelstein, R.: 1983, A global measure of perceived stress, *Journal of Health and Social Behavior* pp. 385–396.

Cohen, S., Kessler, R. C. and Gordon, L. U.: 1997, *Measuring Stress: A Guide for Health and Social Scientists*, Oxford University Press.

Cohen, S. and Williamson, G. M.: 1988, *Perceived stress in a probability sample of the United States*, SAGE Publications, p. 31–67.

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A. and Dobson, K. S.: 2013, A meta-analysis of cognitive-behavioural therapy for adult depres-

sion, alone and in comparison with other treatments, *The Canadian Journal of Psychiatry* **58**(7), 376–385.

Cuijpers, P., Smit, F., Bohlmeijer, E., Hollon, S. D. and Andersson, G.: 2010, Efficacy of cognitive–behavioural therapy and other psychological treatments for adult depression: Meta-analytic study of publication bias, *The British Journal of Psychiatry* **196**(3), 173–178.

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**.

Harris, K. M., Gaffey, A. E., Schwartz, J. E., Krantz, D. S. and Burg, M. M.: 2023, The Perceived Stress Scale as a measure of stress: Decomposing score variance in longitudinal behavioral medicine studies, *Annals of Behavioral Medicine* **57**(10), 846–854.

Kjelsrud, A., Kotsadam, A., Rogeberg, O. and Brodeur, A.: 2025, A comment on "Raising health awareness in rural communities: A randomized experiment in Bangladesh and India" by Siddique et al. (2024). I4R Discussion Paper.

Linardon, J., Cuijpers, P., Carlbring, P., Messer, M. and Fuller-Tyszkiewicz, M.: 2019, The efficacy of app-supported smartphone interventions for mental health problems: A meta-analysis of randomized controlled trials, *World Psychiatry* **18**(3), 325–336.

McGuire, J., Kaiser, C. and Bach-Mortensen, A.: 2020, The impact of cash transfers on subjective well-being and mental health in low-and middle-income countries: A systematic review and meta-analysis. Unpublished.

Mohr, D. C., Vella, L., Hart, S., Heckman, T. and Simon, G.: 2008, The effect of telephone-administered psychotherapy on symptoms of depression and attrition: A meta-analysis., *Clinical Psychology: Science and Practice* **15**(3), 243.

Rahman, T., Ahmed, F., Pakrashi, D., Siddique, A. and Islam, A.: 2021, Income loss and wellbeing during COVID-19 lockdown in rural Bangladesh: Evidence from large household surveys, *Bangladesh Development Studies* p. forthcoming.

Rahman, T., Hasnain, M. G. and Islam, A.: 2021, Food insecurity and mental health of women during COVID-19: Evidence from a developing country, *PLoS One* **16**(7), e0255392.

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

Strayhorn, J. M. and Weidman, C. S.: 1988, A parent practices scale and its relation to parent and child mental health, *Journal of the American Academy of Child & Adolescent Psychiatry* **27**(5), 613–618.

Vlassopoulos, M., Siddique, A., Rahman, T., Pakrashi, D., Islam, A. and Ahmed, F.: 2024a, Data and code for: Improving women's mental health during a pandemic. Nashville, TN: American Economic Association [publisher], 2024. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2024-02-28. https://doi.org/10.3886/E167601V1.

Vlassopoulos, M., Siddique, A., Rahman, T., Pakrashi, D., Islam, A. and Ahmed, F.: 2024b, Improving women's mental health during a pandemic, *American Economic Journal: Applied Economics* **16**(2), 422–455.

Westfall, P. H. and Young, S. S.: 1993, *Resampling-based multiple testing: Examples and methods for p-value adjustment*, Wiley.

## Online Appendix

## A Time-Intensive Parental Investment

In exploratory analyses, VEA24 assess treatment effects on time-intensive parental investments, or "how frequently respondents spend time with their children to help out with their studies and play" (pg. 433). This outcome is the sum of two five-point Likert items modified from Strayhorn and Weidman (1988). The index is measured at both the one-month and ten-month endlines.

Missing values in this index are treated inconsistently across survey waves. On lines 243 and 415 in repository file `Generate_variables.do`, VEA24 generate the index at the one-month and ten-month endlines (respectively) by summing the response on both Likert items using the `rowtotal()` function in Stata's `egen` command, conditional on at least one item being non-missing; the index is coded as missing if both Likert items are missing. Because `rowtotal()` treats missing values as zeros, participants with one missing value will have that missing value imputed as zero. This imputation is done equivalently for both endlines. The inconsistency arises because the parental investment items are reverse-coded at the one-month endline (where higher scores indicate *less* investment) relative to their coding in the ten-month endline data (where higher scores indicate *more* investment). However, line 247 of `Generate_variables.do` reverse-scores the one-month endline index *after* the raw Likert items have been summed on line 243. As a result, missing values are imputed as 'maximum investment' if they occur in the one-month endline, but 'minimum investment' if they occur in the ten–month endline.

This imputation is undisclosed in VEA24's paper and Online Appendix, and affects a substantial share of the data. Exactly one of the two relevant Likert scales is missing for 360 observations at the one-month endline, and for 373 observations at the ten-month endline. Excluding the observations that are missing both Likert scales, over 20% of observations at the one-month endline and over 18% of observations at the ten-month endline are affected by these imputations.

Consistent codings of this imputation strategy do not always yield significant treatment effect estimates. Table A3 shows several robustness checks for this imputation strategy. Panel A directly replicates the results for the time-intensive parental investment index in VEA24's Table 3, except for the observation counts, which are considerably smaller in our reproductions. Imputing missing values consistently as zeros reduces treatment effect estimates at the one-month endline by more than half, rendering those estimates not statistically significantly different from zero. However, a battery of other robustness checks yield treatment effect estimates that are statistically significantly different from zero. These include robustness checks that add dummies indicating that an observation has had zeros imputed into one of the two items underpinning the time-intensive parental investment index, that impute missing values as maximums consistently, that impute missing values as means consistently, or that treat missing values in one of the two items as missing values for the entire index.

## B 'Not Applicable' Answers to Gender Empowerment Questions

Another exploratory outcome analyzed in VEA24 measures gender empowerment. In the one-month endline survey, participants are asked who has control over a series of nine choices concerning household finances and mobility. Based on the repository file `Endline1_kobo form.xls`, participants can answer that these choices are their "Own" (value 0), that they are made by the "Husband" (value 1), that they are "joint" decisions (value 2), that some "Other" person makes the decision (value 4), or that the choice is "Not applicable" (value 99). The gender empowerment index is constructed using the following code in `Generate_variables.do`:

```
254: //Gender empowerment
255: forvalues c = 1(1)9 { //when joint or women's decision
as empowerment
256: gen emp_`c'=1 if (end_Q14_`c'==0|end_Q14_`c'==2)
257: replace emp_`c'=0 if (end_Q14_`c'==1|end_Q14_`c'==4)
```

```
258: replace emp_`c'=. if end_Q14_`c'==99
259: }
261: egen end_empowerment = rowtotal(emp_1 emp_2 emp_3 emp_4
emp_5 emp_6 emp_7 emp_8 emp_9) if (emp_1!=.|emp_2!=.|emp_3!=.
|emp_4!=.|emp_5!=.|emp_6!=.|emp_7!=.|emp_8!=.|emp_9!=.) & treat2!=.
262: drop emp_1-emp_9
263: label variable end_empowerment "Endline Empowerment (score
0-9) where higher values means empowered"
```

Each of the nine gender empowerment items is dichotomized on lines 255-259. Women's choices are coded as their own when they are made by the women or jointly with their husband (values 0 or 2), but are coded as not their own if they are made by the husband or by another party (values 1 or 4).

"Not applicable" responses (value 99) are coded as missing when each Likert item is dichotomized, but are effectively treated as zeros in the broader gender empowerment index. This is because line 261 creates the gender empowerment index for a given participant by summing that participant's entire row of dichotomized indicators using `rowtotal()` command in Stata. As discussed in Online Appendix A, `rowtotal()` treats missing values as zeros, meaning that "Not applicable" answers are effectively imputed as zeros in the gender empowerment index.

This raises doubts about the interpretation of treatment effects on the gender empowerment index, as treatment appears to have significant effects on respondents' probability of reporting "Not applicable" for the items underpinning the gender empowerment index. For example, over 47% of participants untreated at the one-month endline report that their control over their own income is "Not applicable", but this proportion increases to over 67% for treated participants. This imputation thus affects many participants in the survey. Focusing just on one component of the index, over 58% of all participants report that their control over their own income is "Not applicable".

Irregularities arise when comparing these treatment effects on reporting "Not

applicable" responses to VEA24's reported treatment effects on other key mechanisms. Column 6 of VEA24's Table 5 reports significantly positive treatment effects on a dummy variable indicating whether the participant has started new income-generating activity in the last six months at the ten-month endline. A participant's control over her income being "Not applicable" presumably implies that the participant does not have her own income.

In principle, treatment effects on participants reporting that control over their own income is "Not applicable" should hold the opposite sign of treatment effects on women recently starting new income-generating activity; they do not. Table A4 displays treatment effect estimates for these two indicators. These estimates imply that receiving the telecounseling intervention decreases the probability of women having income by over 20 percentage points after one month, but increases the probability of recently starting new income-generating activity by over 13 percentage points after ten months.

We also find that responses to questions on income-generating activity do not match within the same survey wave *for the same participant* for several observations. Kobo questions of type `select_multiple` are multiple choice questions where multiple answers may be selected. Item `Eco_3_4` is a `select_multiple` question in the second endline survey that asks women which of a list of five new income-generating activities (if any) they have started in the past six months. However, both `Endline2_kobo form.xls` and `Endline2_raw.xlsx` indicate that immediately after `Eco_3_4` is asked, the survey form includes six required numeric input questions (`Eco_3_40` through `Eco_3_45`) that ask for numeric inputs to questions concerning whether women started each of the five income-generating activities in the past six months (or none at all). This means that the second endline survey requires this question to be filled out twice consecutively, first as a multiple-choice question and thereafter as a series of numeric inputs (that are all inputted as either zero or one); neither can be skipped. If these questions capture the same information, then they should yield identical results for all participants; this does not

hold. For four observations, answers to `Eco_3_4` do not match those from `Eco_3_40` through `Eco_3_45`.[1]

## C Attitudes Toward Gender Norms

In exploratory analyses, VEA24 examine treatment effects on 'attitudes toward gender norms' at the one-month endline. This is coded as a five-item index of Likert items where participants rate their agreement with statements related to women's rights and skills relative to men. Repository file `Endline1_kobo form.xls` shows that Likert items `end_Q15_1` and `end_Q15_2` are straight-scored, in the sense that higher values of these items correspond to more agreement that women can or should have the same rights and decision-making authority as their husbands (or men more generally). Likert items `end_Q15_3` through `end_Q15_5` are in this sense reverse-scored.

VEA24's 'attitudes toward gender norms' index is constructed by summing dummies that dichotomize responses to the index's five items. The dichotomization and index generation process can be found in file `Generate_variables.do`:

```
265: //Attitude toward gender norms
266: forvalues c = 1(1)2 {
267: gen genatt_`c'=1 if (end_Q15_`c'==3|end_Q15_`c'==4)
268: replace genatt_`c'=0 if
(end_Q15_`c'==0|end_Q15_`c'==1|end_Q15_`c'==2)
269: replace genatt_`c'=. if end_Q15_`c'==.
270: }
272: forvalues c = 3(1)5 { //reverse scoring for positive
questions
273: gen genatt_`c'=1 if
(end_Q15_`c'==0|end_Q15_`c'==1|end_Q15_`c'==2)
274: replace genatt_`c'=0 if (end_Q15_`c'==3|end_Q15_`c'==4)
```

[1] These include observations with `Record_ID` values of 17023, 502614, 505408, and 601530.

```
275: replace genatt_`c'=. if end_Q15_`c'==.
276: }
278: egen end_gender_attitude = rowtotal(genatt_1 genatt_2
genatt_3 genatt_4 genatt_5) if
(genatt_1!=.|genatt_2!=.|genatt_3!=.|genatt_4!=.|genatt_5!=.)
& treat2!=.
279: drop genatt_1 genatt_2 genatt_3 genatt_4 genatt_5
280: label variable end_gender_attitude "Endline attitudes
toward gender norms (between 0 and 5)"
```

This code forms the index used in the paper, `end_gender_attitude`, in two steps. The first step is dichotomization (lines 266-270 and 273-276) – attitudes for item `` `i' `` are coded into dummy `` genatt_`i' ``, which indicates whether respondents provide favorable responses to item `` `i' `` (i.e., responses that indicate sufficiently high agreement that women can/should have the same authority as men). The second step sums these `` genatt_`i' `` indicators for all five items in the index (line 278). Lines 266-276 of the above code are functionally identically reproduced in lines 854-864 of repository file `Appendix_figures_and_tables.do`, immediately before producing VEA24's Online Appendix Table A15.

The dichotomization process is inconsistent across the index's items. Line 274 of the above code shows that a participant who answers the third, fourth, or fifth item in this index with a value of two (displayed in `Endline1_kobo form.xls` as "Neutral") will be coded as having favorable attitudes on that item, and thus receive a value of one in their `` genatt_`i' `` indicator for that item. In contrast, line 268 shows that participants who provide such "Neutral" responses to the first or second item in this index will be coded as having unfavorable attitudes for that item, and thus receive a zero in their `` genatt_`i' `` indicator for that item. This 'moving threshold' issue is similar to that for the stress and depression diagnosis indicators discussed in Section 3.1, and deviates from VEA24's verbal description of how these variables are created. From pg. A23:

> "Dependent variables [...] are binary (=1 if response to the respective question is the maximum two points implying improved attitudes and 0 otherwise, all answered on a 5-point response scale)."

VEA24's Online Appendix Table A15 displays treatment effect estimates on the `genatt_`i'` indicators for the index's five items, which VEA24 discuss on pg. 445; we conduct two robustness checks. First, we similarly estimate treatment effects on the *raw* Likert scores for each item in the gender attitudes index. We reverse-score the index's third through fifth items to reflect the fact that higher values of the `genatt_`i'` indicators correspond to stronger support of gender equality. Second, we estimate treatment effects on *standardized* Likert scores, again reverse-scoring the third through fifth items.

Our robustness checks show that VEA24's coding choices obscure irregularities in the raw items on gender attitudes. Table A5 shows the results of these robustness checks. Panel A adopts VEA24's original coding, displaying treatment effects on the `genatt_`i'` indicators. This replicates the results in VEA24's Online Appendix Table A15. However, Panel B shows treatment effect estimates on the raw item scores, which reveal a considerable inconsistency. Though treatment is estimated to significantly *decrease* participants' agreement that 'most decisions *in the home* should be made by men', it is also estimated to significantly *increase* participants' agreement that 'most decisions *in society* should be made by men.' VEA24's original coding sign-flips the estimated treatment effect on the second item, making that estimate appear to share the same sign as the treatment effect estimate on the first item, both of which are positive (supporting VEA24's original conclusions about the intervention's effects on gender norm attitudes). The estimated negative effect on the second item is sizable, with Panel C showing that treatment is estimated to increase agreement that men should make most decisions in society by over 0.24 standard deviations compared to the control group.

These results also show that VEA24 misinterpret the sign of treatment effect estimates on one of the items in the scale. From pg. 445:

> "... in terms of attitudes and opinions toward gender norms, we find that treated women had improved opinions about female decision-making power in households and society ($p < 0.01$ and $p < 0.05$, respectively) *and were more likely to believe that they can make better calculative decisions than men* ($p < 0.01$)." (emphasis added)

The emphasized portion of this quote references the third column of Online Appendix Table A15, which reports treatment effect estimates on agreement with the statement 'You can make better calculations and decisions than your husband.' VEA24 interpret this as a positive effect, but the estimated effect is actually significantly negative. This is visible in all panels of Table A5, including Panel A. Because Panel A directly replicates Online Appendix Table A15, this was already visible to anyone who checked VEA24's Online Appendix. This negative treatment effect estimate is also substantially large. Panel A implies that receiving the intervention decreases participants' probability of believing that they can make better calculations and decisions than their husband by 18 percentage points. Panel C shows that this estimated treatment effect is larger than half a standard deviation.

## D Transcription Errors in Figures 3 and 5

Figures 3 and 5 in the paper are incorrect due to transcription errors arising from manual copy-pasting of values into plot data. Repository file `README.pdf` states that treatment effect coefficients, $p$-values and t-values from VEA24's Tables 2 and 3 in the paper are directly copy-pasted into repository files `Coefplot_fig3.dta` and `Coefplot_fig5.dta` to construct Figures 3 and 5 in VEA24 figures. A simple comparison of these values in the repository files reveals that there are many substantial mismatches between the paper tables and the values used for these figures. For example, the `Coefplot_fig3.dta` contains the value of 0.0722246 as a coefficient for the effect of treatment on future aspirations. However, the corresponding coefficient from Table 2 equals 0.374. As a second example, `Coefplot_fig5.dta` assigns the treatment effect on time preferences (called "Delay gratification" in the

paper) a value of 0.0141392, which mismatches the coefficient reported in Table 3 with a value of 0.003.

## E Matching Procedure for VEA24 and Source Data

VEA24 use previously collected information, including household background characteristics (see page B7 of VEA24's Online Appendix). While VEA24 do not specify the source datasets, we identified two potential source datasets with data collected in the study area: first, data from Ahmed et al. (2024), and second, data collected for an early childhood development intervention called *Investing in our Future* (*IIOF*).[2] Both datasets are publicly available, either as replication package (Ahmed et al. 2024) or through UK Data Service's online repository.[3]

We check whether these source datasets match the VEA24 data using eight different variables:

1. Household identifier
2. Village identifier
3. Household size
4. Household income

5.-6. Ages of the main male and female household members

7.-8. Years of schooling of the main male and female household members

While this information was readily available in the VEA24 data and Ahmed et al. (2024) data, the raw variables in *IIOF* required processing. Specifically, we calculated household size and income by aggregating individual household members and their incomes, and we converted categorical education levels into years of schooling.[4]

[2] see https://gtr.ukri.org/projects?ref=ES%2FN010221%2F1

[3] See https://reshare.ukdataservice.ac.uk/855543/

[4] From the observations that matched on other variables, we inferred that the original authors consistently applied a recoding algorithm in which education levels from "No school at all" up to "Bachelor/Master's" were converted into years of schooling. For instance, individuals who completed primary school were assigned five years of schooling (see our replication code).

Eventually, we were able to match data for 2,161 of the 2,402 households based on household identifiers – 1,483 from the Ahmed et al. (2024) data, and 678 from the *IIOF* data.[5] Among the matches with data from Ahmed et al. (2024), we identified a coding error that occurred when transferring the data to VEA24. The error stemmed from the fact that the male and female household members' data was stored as respondent and spouse data, and that respondents partly changed between Ahmed et al. (2024) and VEA24. Female respondents in VEA24 were partly spouses of male respondents in the Ahmed et al. (2024) data. Accordingly, respondent data from Ahmed et al. (2024) partly had to be converted into spouse data in VEA24, and *vice versa*. The coding error occurred with households where the respondent in both datasets was female. In these cases, the respondent's age from Ahmed et al. (2024) should have been directly transferred to VEA24. However, instead of the respondent's age, the spouse's age from Ahmed et al. (2024) was incorrectly transferred. This coding error could easily be corrected.

## F Pre-Registration Deviations

We find three inclusion criteria-related deviations from the ANZCTR pre-registration (the AEARCT pre-registration does not list any inclusion criteria), the first of which relates to the ANZCTR pre-registration's claim that "Females aged below 18 years or over 49 years" will be excluded from the sample. Over 10% of the sample is aged 50 years or older. These participants are never excluded from any analyses on these grounds.

Second, participants are chosen without regard to baseline food insecurity. The ANZCTR pre-registration lists its key inclusion criteria as "Female inhabitants of reproductive age group of study areas who have experienced moderate to severe food insecurity on Food Insecurity Experience Scale (FIES)" [*sic*]. Criteria related to food insecurity are unclear. The FIES is an individual-level survey that cap-

[5] 228 of the remaining 241 observations seem to originate from a third RCT that is referenced in other papers using the same data (Rahman, Ahmed, Pakrashi, Siddique and Islam (2021), Ahmed et al. (2021)), but which does not yet appear to be publicly available.

tures participants' perceptions of food insecurity, which is elicited in questions `Q6_1` through `Q6_8` in the baseline survey and questions `Food_2_1` through `Food_2_8` in the ten-month endline survey (Ballard et al. 2013). FIES scores correspond to individuals, not study areas, so FIES scores can only practically be used to include or exclude participants, not entire regions. However, we cannot find any evidence that VEA24 include or exclude participants based on responses to `Q6_1` through `Q6_8`, nor based on responses to `Food_2_1` through `Food_2_8`, for any analysis.

Third and finally, participants with pre-existing mental conditions are included in the sample. The ANZCTR pre-registration states that "Females known to have any pre-existing mental health condition" will be excluded from the study. As discussed in Section 3.1, VEA24 diagnose participants as having a moderate to severe stress condition at baseline if they exhibit baseline PSS scores strictly above 13. However, over 82% of participants are diagnosed with such a condition at baseline, and none are excluded from the analysis on these grounds. In fact, VEA24 directly examine treatment effect heterogeneity for participants with high baseline stress.

## G Appendix Tables and Figures

| | end_Q5_4 | end_Q5_5 | end_Q5_7 | end_Q5_8 | end_Q6_5 | end_Q6_8 | MH_1_4 | MH_1_5 | MH_1_7 | MH_1_8 | Depr_5 | Depr_8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| end_Q5_4 | 1.000 | 0.555 | 0.252 | 0.527 | 0.311 | 0.277 | -0.150 | -0.111 | -0.164 | -0.117 | -0.247 | -0.239 |
| end_Q5_5 | 0.555 | 1.000 | 0.249 | 0.478 | 0.304 | 0.322 | -0.156 | -0.156 | -0.189 | -0.137 | -0.246 | -0.236 |
| end_Q5_7 | 0.252 | 0.249 | 1.000 | 0.248 | 0.182 | 0.133 | -0.051 | -0.024 | -0.065 | -0.043 | -0.148 | -0.102 |
| end_Q5_8 | 0.527 | 0.478 | 0.248 | 1.000 | 0.253 | 0.226 | -0.161 | -0.154 | -0.197 | -0.159 | -0.296 | -0.234 |
| end_Q6_5 | 0.311 | 0.304 | 0.182 | 0.253 | 1.000 | 0.431 | -0.135 | -0.114 | -0.127 | -0.116 | -0.216 | -0.231 |
| end_Q6_8 | 0.277 | 0.322 | 0.133 | 0.226 | 0.431 | 1.000 | -0.127 | -0.120 | -0.109 | -0.086 | -0.170 | -0.233 |
| MH_1_4 | -0.150 | -0.156 | -0.051 | -0.161 | -0.135 | -0.127 | 1.000 | 0.681 | 0.548 | 0.678 | 0.344 | 0.305 |
| MH_1_5 | -0.111 | -0.156 | -0.024 | -0.154 | -0.114 | -0.120 | 0.681 | 1.000 | 0.522 | 0.667 | 0.314 | 0.318 |
| MH_1_7 | -0.164 | -0.189 | -0.065 | -0.197 | -0.127 | -0.109 | 0.548 | 0.522 | 1.000 | 0.510 | 0.337 | 0.319 |
| MH_1_8 | -0.117 | -0.137 | -0.043 | -0.159 | -0.116 | -0.086 | 0.678 | 0.667 | 0.510 | 1.000 | 0.327 | 0.307 |
| Depr_5 | -0.247 | -0.246 | -0.148 | -0.296 | -0.216 | -0.170 | 0.344 | 0.314 | 0.337 | 0.327 | 1.000 | 0.503 |
| Depr_8 | -0.239 | -0.236 | -0.102 | -0.234 | -0.231 | -0.233 | 0.305 | 0.318 | 0.319 | 0.307 | 0.503 | 1.000 |

*Note:* Items `end_Q5_*` and `end_Q6_*` respectively correspond to PSS and CES-D-10 items that are supposed to be reverse-scored, as stored in the raw one-month endline data. Items `MH_1_*` and `Depr_*` respectively correspond to PSS and CES-D-10 items that are supposed to be reverse-scored, as stored in the raw ten-month endline data.

Table A1: Item-Level Correlations Across Waves

| | One-Month Endline | | Ten-Month Endline | |
|---|---|---|---|---|
| | Standardized PSS Score (1) | Standardized CES-D-10 Score (2) | Standardized PSS Score (3) | Standardized CES-D-10 Score (4) |
| **Panel A: Original Coding** | | | | |
| Treatment Effect, Baseline PSS Score < Median | -0.548 | -0.538 | -0.602 | -0.494 |
| | (0.071) | (0.059) | (0.101) | (0.079) |
| | [0] | [0] | [0] | [0] |
| Treatment Effect, Baseline PSS Score ≥ Median | -0.837 | -0.752 | -0.481 | -0.505 |
| | (0.082) | (0.072) | (0.087) | (0.081) |
| | [0] | [0] | [0] | [0] |
| Treatment ×𝟙[Baseline PSS Score ≥ Median] | -0.256 | -0.21 | 0.167 | 0.03 |
| | (0.097) | (0.086) | (0.112) | (0.098) |
| | [0.009] | [0.016] | [0.135] | [0.759] |
| Treatment × Baseline PSS Score | -0.031 | -0.022 | 0.01 | 0.001 |
| | (0.011) | (0.01) | (0.013) | (0.012) |
| | [0.006] | [0.023] | [0.45] | [0.92] |
| **Panel B: Consistent Reverse-Coding** | | | | |
| Treatment Effect, Baseline PSS Score < Median | -0.548 | -0.538 | 0.552 | 0.464 |
| | (0.071) | (0.059) | (0.078) | (0.073) |
| | [0] | [0] | [0] | [0] |
| Treatment Effect, Baseline PSS Score ≥ Median | -0.837 | -0.752 | 0.602 | 0.491 |
| | (0.082) | (0.072) | (0.09) | (0.079) |
| | [0] | [0] | [0] | [0] |
| Treatment ×𝟙[Baseline PSS Score ≥ Median] | -0.256 | -0.21 | 0.053 | 0.046 |
| | (0.097) | (0.086) | (0.101) | (0.095) |
| | [0.009] | [0.016] | [0.603] | [0.631] |
| Treatment × Baseline PSS Score | -0.031 | -0.022 | 0.007 | -0.003 |
| | (0.011) | (0.01) | (0.012) | (0.012) |
| | [0.006] | [0.023] | [0.574] | [0.777] |
| **Panel C: Straight-Scored Items Only** | | | | |
| Treatment Effect, Baseline PSS Score < Median | -0.125 | -0.4 | 0.011 | -0.022 |
| | (0.061) | (0.06) | (0.086) | (0.079) |
| | [0.042] | [0] | [0.902] | [0.783] |
| Treatment Effect, Baseline PSS Score ≥ Median | -0.142 | -0.485 | 0.012 | -0.09 |
| | (0.075) | (0.066) | (0.082) | (0.08) |
| | [0.06] | [0] | [0.888] | [0.26] |
| Treatment ×𝟙[Baseline PSS Score ≥ Median] | -0.007 | -0.09 | 0.022 | -0.049 |
| | (0.083) | (0.081) | (0.099) | (0.097) |
| | [0.934] | [0.272] | [0.828] | [0.612] |
| Treatment × Baseline PSS Score | -0.016 | -0.013 | -0.005 | -0.016 |
| | (0.012) | (0.012) | (0.014) | (0.015) |
| | [0.208] | [0.274] | [0.708] | [0.276] |
| **Panel D: Reverse-Scored Items Only** | | | | |
| Treatment Effect, Baseline PSS Score < Median | -0.934 | -0.769 | -1.153 | -2.015 |
| | (0.069) | (0.063) | (0.121) | (0.084) |
| | [0] | [0] | [0] | [0] |
| Treatment Effect, Baseline PSS Score ≥ Median | -1.038 | -0.767 | -1.043 | -1.97 |
| | (0.064) | (0.062) | (0.091) | (0.073) |
| | [0] | [0] | [0] | [0] |
| Treatment ×𝟙[Baseline PSS Score ≥ Median] | -0.131 | -0.03 | 0.043 | 0.004 |
| | (0.081) | (0.083) | (0.132) | (0.1) |
| | [0.107] | [0.72] | [0.744] | [0.965] |
| Treatment × Baseline PSS Score | -0.024 | 0 | -0.005 | 0.003 |
| | (0.016) | (0.015) | (0.026) | (0.018) |
| | [0.126] | [0.986] | [0.846] | [0.861] |

*Note*: Regression estimates on the outcome displayed in the column title are shown along with standard errors clustered at the village level in parentheses and raw $p$-values in straight brackets. Each estimate, standard error, and $p$-value comes from a different regression model. Standardized scores hold a mean of zero and a standard deviation of one in the control group. Treatment effects are based on treatment status indicators that account for attrition in each wave. All models control for respondents' age, education, and occupation, the occupation of respondents' husbands, a dummy indicating whether the respondent reports that their household lost income during the COVID-19 pandemic, a dummy indicating whether the respondent reports that their household chores increased after the COVID-19 lockdowns, a dummy indicating whether the respondent is the head of their household, the number of people in the respondent's household, the number of children in the respondent's household, and union fixed effects.

Table A2: Heterogeneity Analysis, Using Different Index Scoring

| | One-Month Endline | | Ten-Month Endline | |
|---|---|---|---|---|
| | Standardized Time-Intensive Parental Investment (1) | Standardized Time-Intensive Parental Investment (2) | Standardized Time-Intensive Parental Investment (3) | Standardized Time-Intensive Parental Investment (4) |
| **Panel A: Original Coding** | | | | |
| Treatment Effect | 0.227 | 0.22 | 0.232 | 0.192 |
| | (0.055) | (0.057) | (0.05) | (0.049) |
| | [0] | [0] | [0] | [0] |
| | {0} | {0} | {0} | {0} |
| Observations | 1790 | 1790 | 1978 | 1978 |
| **Panel B: Missing Imputed as Zero** | | | | |
| Treatment Effect | 0.107 | 0.086 | 0.232 | 0.192 |
| | (0.06) | (0.061) | (0.05) | (0.049) |
| | [0.076] | [0.157] | [0] | [0] |
| | {0.294} | {0.163} | {0} | {0} |
| Observations | 1790 | 1790 | 1978 | 1978 |
| **Panel C: Missing Imputed as Zero, Control for Dummy Indicating Imputation** | | | | |
| Treatment Effect | 0.18 | 0.163 | 0.242 | 0.218 |
| | (0.051) | (0.053) | (0.04) | (0.04) |
| | [0] | [0.002] | [0] | [0] |
| | {0.005} | {0.001} | {0} | {0} |
| Observations | 1790 | 1790 | 1978 | 1978 |
| **Panel D: Missing Imputed as Maximum** | | | | |
| Treatment Effect | 0.227 | 0.22 | 0.313 | 0.287 |
| | (0.055) | (0.057) | (0.053) | (0.052) |
| | [0] | [0] | [0] | [0] |
| | {0} | {0} | {0} | {0} |
| Observations | 1790 | 1790 | 1978 | 1978 |
| **Panel E: Missing Imputed as Mean** | | | | |
| Treatment Effect | 0.2 | 0.187 | 0.304 | 0.271 |
| | (0.059) | (0.06) | (0.051) | (0.051) |
| | [0.001] | [0.002] | [0] | [0] |
| | {0.007} | {0.003} | {0} | {0} |
| Observations | 1790 | 1790 | 1978 | 1978 |
| **Panel F: Missing One Imputed As Missing Both** | | | | |
| Treatment Effect | 0.282 | 0.251 | 0.267 | 0.249 |
| | (0.062) | (0.064) | (0.057) | (0.055) |
| | [0] | [0] | [0] | [0] |
| | {0} | {0} | {0} | {0} |
| Observations | 1430 | 1430 | 1605 | 1605 |
| Covariates | | X | | X |

*Note*: Regression estimates on the outcome displayed in the column title are shown along with standard errors clustered at the village level in parentheses, raw $p$-values in straight brackets, and Westfall-Young family-wise error rate-adjusted $p$-values in curled brackets (Westfall and Young 1993). For each panel and each treatment effect type (i.e., with covariates and without covariates), Westfall-Young $p$-values are calculated as in VEA24's Table 3 across the time-intensive parental investment outcomes at both endlines, food insecurity at both endlines, COVID-19 confidence at the one-month endline, gender empowerment at the one-month endline, attitudes toward gender norms at the one-month endline, attitudes toward intimate partner violence at the one-month endline, and vaccination status, risk preferences, social preferences, and time preferences at the ten-month endline. All models control for union fixed effects. Additional covariates include respondents' age, education, and occupation, the occupation of respondents' husbands, a dummy indicating whether the respondent reports that their household lost income during the COVID-19 pandemic, a dummy indicating whether the respondent reports that their household chores increased after the COVID-19 lockdowns, a dummy indicating whether the respondent is the head of their household, the number of people in the respondent's household, and the number of children in the respondent's household. All models in Panel C additionally control for a dummy indicating whether a zero is imputed into a missing value for one of the items in the time-intensive parental investment index.

Table A3: Robustness Checks on Time-Intensive Parental Investment

| | Participant Started New Income-Generating Activity in Last Six Months (1) | Participant Started New Income-Generating Activity in Last Six Months (2) | Participant's Control Over Her Own Income is 'Not Applicable' (3) | Participant's Control Over Her Own Income is 'Not Applicable' (4) |
|---|---|---|---|---|
| Treatment Effect | 0.139 | 0.137 | 0.206 | 0.212 |
| | (0.028) | (0.028) | (0.024) | (0.022) |
| | [0] | [0] | [0] | [0] |
| Survey Wave | Ten-Month Endline | Ten-Month Endline | One-Month Endline | One-Month Endline |
| Covariates | | X | | X |
| Observations | 2254 | 2254 | 2402 | 2402 |
| $R^2$ | 0.075 | 0.091 | 0.061 | 0.249 |

*Note*: Regression estimates on the outcome displayed in the column title are shown along with standard errors clustered at the village level in parentheses and raw $p$-values in straight brackets. Treatment effects are based on baseline treatment status. All models control for union fixed effects. Additional covariates include respondents' age, education, and occupation, the occupation of respondents' husbands, a dummy indicating whether the respondent reports that their household lost income during the COVID-19 pandemic, a dummy indicating whether the respondent reports that their household chores increased after the COVID-19 lockdowns, a dummy indicating whether the respondent is the head of their household, the number of people in the respondent's household, and the number of children in the respondent's household.

Table A4: Robustness Checks on Treatment Effects for Income-Generating Activity

| | Disagreement w/ 'Most decisions in the home should be made by men' (1) | Disagreement w/ 'Most decisions in society should be made by men' (2) | Agreement w/ 'You can make better calculations and decisions than your husband' (3) | Agreement w/ 'Women and men in general should have equal rights' (4) | Agreement w/ 'A woman can disagree with any decision of her husband' (5) |
|---|---|---|---|---|---|
| **Panel A: Original Coding** | | | | | |
| Treatment Effect | 0.259 | 0.038 | -0.18 | 0.02 | 0.012 |
| | (0.022) | (0.015) | (0.022) | (0.011) | (0.022) |
| | [0] | [0.013] | [0] | [0.078] | [0.601] |
| $R^2$ | 0.13 | 0.074 | 0.086 | 0.046 | 0.041 |
| **Panel B: Raw Item Scores** | | | | | |
| Treatment Effect | 0.665 | -0.241 | -0.54 | 0.494 | 0.359 |
| | (0.068) | (0.048) | (0.056) | (0.04) | (0.06) |
| | [0] | [0] | [0] | [0] | [0] |
| $R^2$ | 0.113 | 0.073 | 0.104 | 0.127 | 0.062 |
| **Panel C: Standardized Item Scores** | | | | | |
| Treatment Effect | 0.607 | -0.242 | -0.539 | 0.57 | 0.347 |
| | (0.062) | (0.049) | (0.056) | (0.046) | (0.058) |
| | [0] | [0] | [0] | [0] | [0] |
| $R^2$ | 0.113 | 0.073 | 0.104 | 0.127 | 0.062 |
| Observations | 2219 | 2217 | 2197 | 2216 | 2207 |

*Note*: Regression estimates on the outcome displayed in the column title are shown along with standard errors clustered at the village level in parentheses and raw $p$-values in straight brackets. Each estimate, standard error, and $p$-value comes from a different regression model. Standardized scores hold a mean of zero and a standard deviation of one in the control group. Treatment effects are based on treatment status indicators that account for attrition in the one-month endline. All models control for respondents' age, education, and occupation, the occupation of respondents' husbands, a dummy indicating whether the respondent reports that their household lost income during the COVID-19 pandemic, a dummy indicating whether the respondent reports that their household chores increased after the COVID-19 lockdowns, a dummy indicating whether the respondent is the head of their household, the number of people in the respondent's household, the number of children in the respondent's household, and union fixed effects.

Table A5: Robustness Checks on Treatment Effects for Gender Norm Attitudes

Table A6: Impossible Values in the Raw Baseline Data

| **Question** | **Answer Options** | **Issue** |
|---|---|---|
| occ_respondent | 1 Housewives<br>2 Farmers<br>3 Agricultural workers<br>4 Day laborers<br>5 Business<br>6 Public service<br>7 Private service<br>8 Other | The data includes value "0" 2059 times and value "98" 182 times. |
| occ_husband | 1 Farmers<br>2 Agricultural workers<br>3 Day laborers<br>4 Business<br>5 Public service<br>6 Private service<br>7 Other | The data includes value "0" 48 times. |
| Q13_3 | 0 False<br>1 True<br>2 Don't know | The data includes value "3" 7 times. |
| Q14_1 | 0 No<br>1 Yes<br>98 I don't want to say<br>99 Don't know | The data include values 0 – 10, with 2351 observations including values that should not be possible. |

Table A7: Impossible Missing Values in the Raw Endline Data

| **Question** | **Endline** | **# Missing Values** |
|---|---|---|
| end_Q3_2 | One-Month | 1 |
| end_Q3_5 | One-Month | 10 |
| end_Q3_6 | One-Month | 2 |
| end_Q3_7 | One-Month | 9 |
| end_Q12_2 | One-Month | 506 |
| end_Q12_3 | One-Month | 714 |
| end_Q15_1 | One-Month | 1 |
| end_Q15_2 | One-Month | 3 |
| end_Q15_3 | One-Month | 23 |
| end_Q15_4 | One-Month | 4 |
| end_Q15_5 | One-Month | 13 |
| Eco_3_1 | Ten-Month | 12 |
| Eco_3_2 | Ten-Month | 21 |
| Eco_3_3 | Ten-Month | 31 |
| Phy_599 | Ten-Month | 187 |
| Edu_6_1 | Ten-Month | 334 |
| Play_6_2 | Ten-Month | 591 |

Table A8: Irregular Question Labels, Baseline Survey

| **Question** | **English Label** | **Bengali Label, Official Translation** | **# Characters, English Label** |
|---|---|---|---|
| Q4_1 | In the last 7 days, apart from using toilet, I washed my hands at least 5 times | In the last 7 days, I have washed my hands at least 5 times, without using the toilet. | 79 |
| Q12_B | In last 7 day-how often have you felt that you were unable to control | How often in the past 7 days have you felt unable to control yourself? | 69 |
| Q12_D | In last 7 day- how often have you felt confident about your ability | In the last 7 days - how often have you felt confident about your abilities? | 67 |
| Q12_F | In last 7 day-how often have you found that you could not cope with | How many times in the past 7 days have you found yourself unable to cope? | 67 |
| Q12_I | In last 7 day-how often have you been angered because of things that were beyond | How many times have you been angry in the last 7 days because of things outside of it? | 80 |
| Q12_J | In last 7 day-how often have you felt difficulties were piling up so high | How often in the past 7 days have you felt that problems were getting too much? | 73 |

Table A9: Irregular Question Labels, One-Month Endline Survey

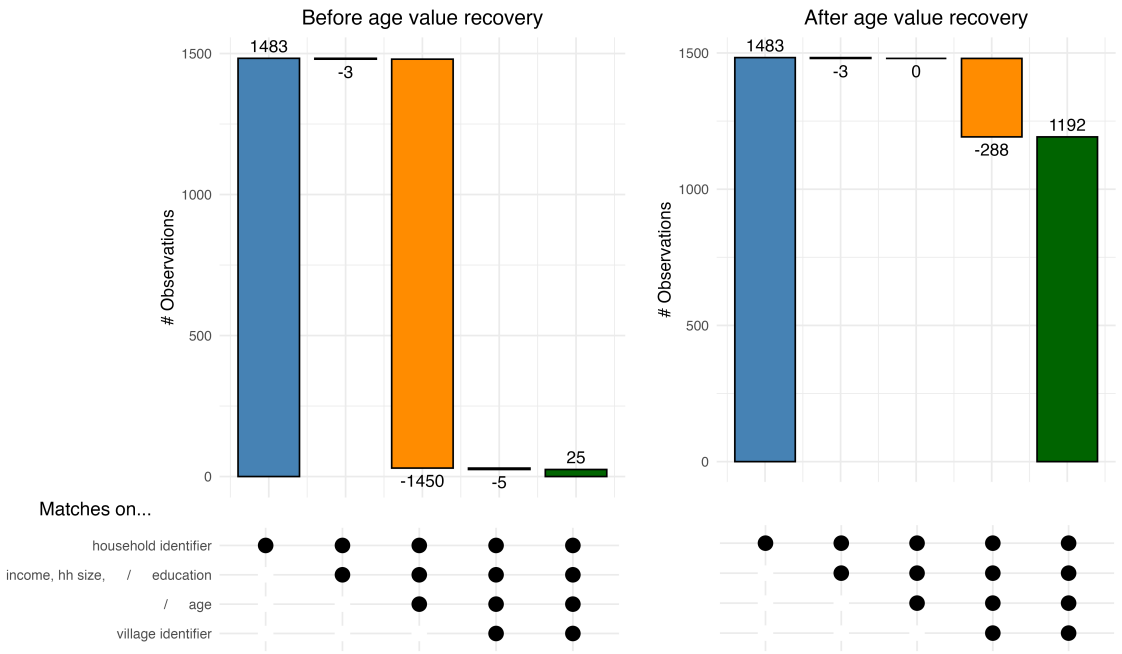
| **Question** | **English Label** | **Bengali Label, Official Translation** | **# Characters, English Label** |
|---|---|---|---|
| end_Q4_1 | Lack of money will not have 3 meals for everyone in the house - were you worried | Lack of money will not have 3 meals for everyone in the house - were you worried? | 80 |
| end_Q4_2 | Please think again about the last 2 to 3 weeks, was there a time when you or any | Please think back to the last 2 to 3 weeks, was there a time when you or someone else in the family could not eat healthy and nutritious food because of lack of money? | 80 |
| end_Q4_3 | Has it ever happened in the last 2 to 3 weeks that for lack of money you or some | In the past 2 - 3 weeks, has it ever happened that you or someone else in the family ate only a few types of food (repeatedly the same / few items) because of lack of money? | 80 |
| end_Q4_4 | What has happened in the last 2 weeks that you have not had enough money to feed | What has happened in the last 2 weeks that I have had to go without a meal for someone in the family because I don't have enough money? | 80 |
| end_Q4_5 | Please think back to the last 2 to 3 weeks, was there a time when you ate less t | Please think back to the last 2 to 3 weeks, was there a time when you ate less than you should have, even though you played for 3 hours due to lack of money? | 80 |
| end_Q4_7 | In the last 2 to 3 weeks you / someone in the family was hungry but could not bu | Has it happened in the last 2 to 3 weeks that you / someone in the family was hungry but could not buy food due to lack of money? | 80 |
| end_Q4_8 | Has it ever happened in the last 2 to 3 weeks that you / family have not eaten a | Has there ever been a time in the last 2 to 3 weeks that you / anyone in the family went without food for a whole day due to lack of money? | 80 |
| end_Q5_2 | How often have you felt that you were unable to control the important things in | How many times have you felt that you were unable to control important things? | 79 |

| Question | English Label | Bengali Label, Official Translation | # Characters, English Label |
|---|---|---|---|
| end_Q5_6 | how often have you found that you could not cope with all the things that you ha | How many times have you found that you can't cope with all the things? | 80 |
| end_Q5_10 | How often have you felt difficulties were piling up so high that you could not o | How many times have you felt that the problems are becoming so overwhelming that you just can't handle them? | 80 |
| end_Q9_2 | How optimistic are you that you or your husband will be able to earn a living ag | How optimistic are you that you or your spouse will be able to earn a living? | 80 |
| end_Q10_1 | How confident are you in protecting yourself and your family from the coronaviru | How confident are you to protect yourself and your family from the corona virus? | 80 |
| end_Q10_2 | How confident are you that you are following the rules to protect yourself and y | How confident are you that you are following the rules for yourself and your safety? | 80 |
| end_Q10_3 | How confident are you that everyone in your family is following the rules to pro | How confident are you that everyone in your family is following the rules to save themselves? | 80 |
| end_Q10_4 | How confident are you that you will be able to cope if someone in your family is | How confident are you that you will be able to cope if someone in your family is affected? | 80 |
| end_Q10_5 | How are you confident that you know where you have to go and ask for help to dea | How confident are you that you know where to go and seek help if someone is affected? | 80 |

Table A10: Irregular Question Labels, Second Endline Survey

| **Question** | **English Label** | **Bengali Label, Official Translation** | **# Characters, English Label** |
|---|---|---|---|
| MH_1_4 | How often do you feel that you have the ability to solve problems for yourself a | How often do you feel that you have the ability to solve problems for yourself? | 80 |
| MH_1_5 | How often have you feel that all things are going in the way that you wish or co | How many times have you felt that everything was going or cooperating as you wished? | 80 |
| MH_1_6 | How often do you feel that you are not doing what you need to do in the current | How often do you feel that you are not doing what needs to be done at the moment? | 79 |
| MH_1_10 | How often do you feel that the problem is getting worse so much that you can't g | How often do you feel that the problem is getting so bad that you can't go to the gym? | 80 |
| Food_2_1 | Lack of money will not be enough for 3 meals for everyone in the house - were yo | Lack of money for everyone in the house and not enough for lunch - were you? | 80 |
| Food_2_2 | Please think again about the last 2 to 3 weeks, was there any time when you / an | Please think back over the last 2 weeks, was there a time when you / | 80 |
| Food_2_3 | Has there ever been a time in the last 2 to 3 weeks when you / someone in the fa | Has there ever been a time in the last 2 to 3 weeks when you/someone was in FA? | 80 |
| Food_2_5 | Please think again about the last 2 to 3 weeks, was there a time when you ate le | Please think back over the last 2 weeks, were there any times when you ate "lay"? | 80 |
| Food_2_6 | In the last 2 to 3 weeks, has your family been in food crisis due to lack of mon | During the past 2 to 3 weeks, has your family experienced food insecurity due to lack of food? | 80 |
| Food_2_7 | In the last 2 to 3 weeks, someone in your family was hungry but could not buy fo | In the past 2 to 3 weeks, someone in your family was hungry but couldn't buy pho | 80 |
| Food_2_8 | Has it ever happened in the last 2 to 3 weeks that you / anyone in your family h | Has it ever happened in the last 2 to 3 weeks that you/someone in your family? | 80 |

| **Question** | **English Label** | **Bengali Label, Official Translation** | **# Characters, English Label** |
|---|---|---|---|
| Eco_3_1 | In the last 10 months, how often have you borrowed money from your relatives, ne | In the last 10 months, how many times have you borrowed money from your relatives, no | 80 |
| Eco_3_2 | In the last 10 months, due to lack of food at home, how often have you contacted | In the last 10 months, how many times have you been contacted for lack of food at home? | 80 |
| Eco_3_3 | How much work has your husband done lately (such as in the last 10 months) compa | How much has your husband worked recently (eg. in the last 10 months) | 80 |
| Feel_HR | I want to be the one in my village that everyone will respect- tell me how true | How true is it that I want to be a person that everyone respects in my village? | 79 |
| Phy_55 | Breathing Exercises - Hold the breath for 5 seconds and slowly exhale, doing thi | Breathing Exercise Hold the breath for 5 seconds and exhale slowly. | 80 |
| end2_risk _pref | In general, are you a person who is fully prepared to take risks? Or are you som | In general, are you a person who is fully prepared to take risks? | 80 |
| end2_social _pref | Suppose today you suddenly got 5000 BDT, how much money will you donate from thi | Suppose today you suddenly got 5000 rupees, how much money will you donate from this money? | 80 |




*Note*: Orange-shaded bars show the number of observations lost when an additional matching criterion is specified. Green bars show the number of observations remaining when all matching criteria are met. Matching criteria include household identifiers, household incomes, household sizes, education levels of husbands and wives, ages of husband and wives, and village identifiers. The left graph shows matching on the raw values from AHI24, whereas the right graph shows matching after correcting transferrals of ages from the AHI24 data; see Online Appendix E for details.

Figure A1: Sample Matching Between VEA24 and AHI24

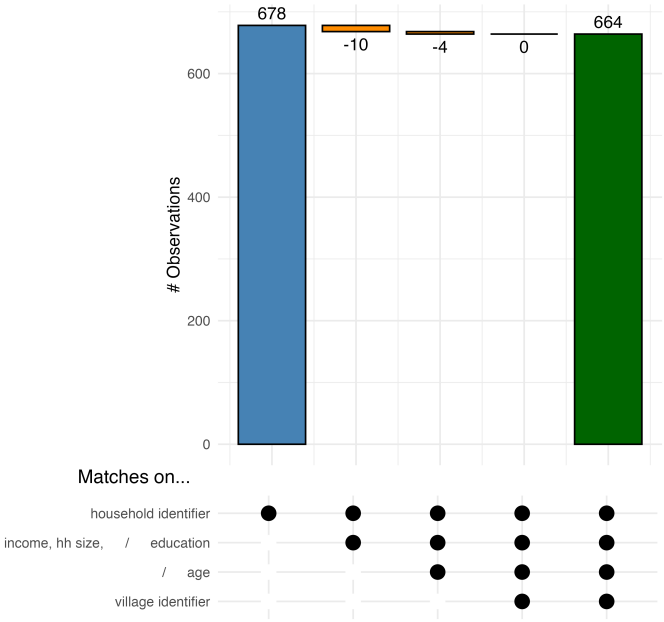



Figure A2: Sample Matching Between VEA24 and GEA24

*Note*: Orange-shaded bars show the number of observations lost when an additional matching criterion is specified. Green bars show the number of observations remaining when all matching criteria are met. Matching criteria include household identifiers, household incomes, household sizes, education levels of husbands and wives, ages of husband and wives, and village identifiers.