Online Appendix

Cognitive Behavioral Therapy Among Ghana's Rural Poor Is Effective Regardless of Baseline Mental Distress

Nathan Barker, Gharad Bryan, Dean Karlan, Angela Ofori-Atta and Christopher Udry

Appendix A: Sample Construction and Randomization Procedures

The study was conducted in 14 districts in five regions of Ghana: Northern, Upper East, Ashanti, Bono, and Bono East. In each district, IPA and Heifer International met with District Assembly staff (i.e. local government) to identify each community in the district, and to select communities that (a) had at least 50 compounds,[25] (b) were accessible by road from the district capital (to allow staff based in the district capital to travel to the communities), and (c) did not have programs similar to Heifer's graduation program already in operation.

In each community that fulfilled these initial criteria, IPA administered a census of all households, in which we collected contact information and administered a proxy means test. In total 68,309 households in 366 communities were part of the census.

Surveying and the intervention took place in two waves, divided by the two ecological zones in which our study took place. In the Northern Belt (Northern and Upper East Regions), the census took place from January through March, 2016. In the Middle Belt (Bono, Bono East and Ashanti Regions), the census took place from May through June, 2016.

Following the census, communities were deemed eligible if they had at least 45 compounds. Following the determination that a community was eligible, we randomized communities into treatment or control status (among treatment communities, we also randomized whether a subsequent economic program would take place).

In each community, we selected the 40 compounds with the lowest average household proxy means test score, and for each compound, randomly chose one household to include in our eligible sample, which consisted of 40 households in each community.

In order to maximize statistical power, the number of households we surveyed (and subsequently preserved in our sample) differed by community-level treatment status. In particular, we targeted 17 households in control communities, and either 20 or 40 in treatment communities (depending on whether they were set to receive a subsequent economic program). We randomly selected which of the 40 eligible households would be surveyed and would thus remain in our sample.

[25] A compound is a cluster of households living in separate dwellings clustered within a single structure.

Following the selection of households into our sample, we administered our baseline survey. In the Northern Belt, the baseline survey was administered from April through June, 2016. In the Middle Belt, the baseline survey was administered from September through November, 2016. Following the completion of the baseline survey, we randomized communities to either "Male CBT" or "Female CBT" communities (whether the single-gendered CBT group in the community would be for men or women), and then randomized which households (and thus individuals) would be offered CBT.

In the Northern Belt, CBT was administered from July to September 2016. For the Middle Belt, CBT was administered from January through March 2017.

The endline survey was implemented in the Northern Belt between November and December 2016, and in the Middle Belt between April and May, 2017.

For each of the randomizations described above, we performed a re-randomization stratification procedure. We randomized a predetermined 10,000 times, tested for balance on a vector of characteristics (listed in Appendix Table 2) and picked the randomization with the maximum minimum p-value. This procedure was applied to both the community-level randomizations and the within-community randomizations.

Data and code to replicate the analysis can be found in Barker et al. (2022*b*).

Appendix B: Description of Randomization Inference Procedure

The multi-stage nature of our randomization procedure (community-level randomization, followed by randomization to determine which households in pure control communities are included in the final sample, followed by randomization of the gender of CBT in a community, and individual level-randomization into CBT or control) motivates our use of randomization inference. We therefore follow the general procedure laid out by Young (2019), adapted to the specifics of our randomization procedure. In particular, we implemented the following procedure, for each of 2,000 simulations, (again following Young (2019), who finds "no appreciable change in rejection rates beyond 2,000 draws"):

1. Using community-level data obtained from the census, re-randomize 100 times to assign placebo treatments, test for balance on characteristics (listed in Appendix Table 2, Panel A), choose the randomization with the maximum minimum p-value.

2. Assign sample weights to households based on their placebo community assignment, reflecting the fact that a smaller number of households were included in the sample in pure control and CBT only communities in the true randomization assignment. For example, an household assigned to a pure control community in the real randomization (in which we randomly selected 17 of the 40 eligible households to include in our study) but was assigned to be in a full program community in the placebo randomization (in which all 40 eligible households were chosen) would be given a sample weight of (40/17).

3. Using household and adult-level data, re-randomize 100 times to assign placebo (i) CBT gender in a given community, and (ii) individual assignment into CBT or control, test for balance on characteristics (listed in Appendix Table 2, Panel C), choose the randomization with the maximum minimum p-value.

4. With the placebo treatment assignments, regress our outcome variables of interest on the placebo treatment, and in tests of heterogeneity, on the interaction between the true baseline outcome (distress, gender) and the placebo treatment. Store estimates of the coefficients of (i) average treatment effects, and for tests of heterogeneity, of (ii) sub-group treatment effects, and (iii) the difference in coefficient between the two sub-groups.

Our inference involves comparing the true point estimates (and in cases of heterogeneity, the difference in coefficients) to the empirical distribution of coefficients (differences) from our 2,000 simulations. Our "RI p-value" is equal to the share of the 2,000 simulations in which the absolute value of the coefficient (difference) is larger than the absolute value (absolute value of the difference) of the results from our true randomization assignment.

We implement stages 1-3 (i.e. the placebo randomizations) separately for the "Northern Belt" (Northern and Upper East) and "Middle Belt" (Ashanti, Bono, Bono East) parts of our sample, reflecting the way in which we performed the actual randomization (we completed data collection activities and conducted randomizations for the Northern Belt before proceeding to the Middle Belt). For stage 4, the regression, we pool the full sample.

One aspect to note is that in our initial randomization procedure, we re-randomized a pre-determined 10,000 times to determine the maximum minimum p-value, where in these simulations we re-randomize a pre-determined 100 times. This adjustment was made for computational reasons, as given our nested randomizations we face the curse of dimensionality in exactly replicating our procedure.[26] We found when comparing simulations with 10,000 to 100 re-randomizations that our balance did not seem to differ appreciably.

[26] We estimate that exactly replicating the procedure would take approximately 160 days for the code to run.

Appendix C: Structured Ethics Appendix

For more explanation of each question, see Asiedu et al. 2020.

1. Policy Equipoise

Is there policy equipoise? That is, is there uncertainty regarding participants' net benefits from each arm of the study relative to the other arms and to the best possible policy to which participants could have access? If not, ethical randomization requires two conditions related to scarcity: (1) Was there scarcity, i.e., did the inclusion of multiple arms change the expected aggregate value of the programs delivered? (2) Do all ex-ante identifiable participants have equal moral or legal claims to the scarce programs?

If there is no reasonable expectation that one arm of the study produces more benefits to participants than any other arm or than the best possible alternative policy, then randomization is ethically unproblematic. If not, then excluding some participants from the superior treatment arm can only be justified by scarcity. Scarcity conditions are twofold: (1) resources are not sufficient, given constraints, to include all participants in the superior treatment arm; (2) no ex-ante identifiable participants are excluded from the superior arm and have a greater claim to those resources than any participant assigned to the superior arm. See MacKay 2018 for more complete discussions of policy equipoise.

The treatment arm provides group CBT therapy to a general population of the poor, rather than to individuals with a common identified mental health difficulty. There was no consensus among experts regarding the effectiveness of this form of CBT for a general population, so the control and treatment arms were in policy equipoise. Furthermore, for those in mental distress at the time of the intervention, we believe that there is equipoise given limited evidence of effectiveness in this setting and with CBT delivered in groups by lay counsellors. Regardless, should there not be equipoise, there was scarcity in that the program had a limited budget for delivering CBT to communities.

2. Role of researchers with respect to implementation

Are researchers "active" researchers, i.e. did the researchers have direct decision making power over whether and how to implement the program? If YES, what was the disclosure to participants and informed consent process for participation in the program? Providing IRB approval details may be sufficient but further clarification of any important issues should be discussed here. If NO, i.e., implementation was separate, explain the separation.

A researcher should be considered "active" if, for example, the implementing staff are employed by an institution at which the PI is employed, and the staff report either

directly or indirectly to the PI at this institution with regard to this project. Or if researchers control funding for implementation, or have direct decision-making power over key implementation decisions.

Some key factors that help illuminate whether the researchers are "active" or not (here "researchers" are defined as the PIs and the staff that report directly or indirectly to the PIs): Did researchers directly provide any of the interventions, or parts thereof, to participants? Did researchers interact directly with participants and implicitly endorse one or more of the interventions?

The research team played an active role in the design of the program, but the program was implemented by a third party. IRB approval was received from the University of Ghana Medical School, IPA, Yale University and Northwestern University. Informed consent from participants was limited to consent to take part in a survey, and not the intervention. Lack of informed consent for the intervention aspect is justified because of the voluntary nature of the intervention; the independent purpose of the intervention as a non-research service for those in the community; and, the fact that the participants were not a vulnerable population seeking advice from the research team.

3. Potential harms to participants or nonparticipants from the interventions or policies

Does the intervention, policy or product being studied pose potential harm to participants or non-participants? Related, are participants or likely affected non-participants particularly vulnerable? Also related, are participants' access to future services or policies changed because of participation in the study? If yes to any of the above, what is being done to mitigate such risks

It may be important to consider whether the researchers are "active" (see above) or not for this discussion. If the researchers are "active", then they are responsible for the potential harms, and thus a robust discussion is appropriate. If the researchers are not "active", then while they may not be responsible for potential harms, a discussion of this would be appropriate here.

There will almost always be some potential harms, if nothing else because of complementary investments such as time that participants in an intervention necessarily redirect from one activity to another. Quantifying these risks and complementary investments may be difficult ex-ante, but a discussion of what they are here would help the reader assess their likely importance relative to the potential benefits of the tested intervention. Also note that measuring any harms ex-post may be the exact reason for the study, particularly when the intervention is common.

If risks to nonparticipants exist, discuss the mechanisms through which the risk arises

from the study and provide an estimate of the magnitude of the risk and the probability of harm.

The IRB reviewed protocols for the CBT program, participation in which was voluntary and from which individuals were always free to withdraw. Protocols were in place for responding to sensitive issues and distress that emerged during or as a result of the sessions. In particular, anyone identified in surveys as in distress was directed to the community psychiatric nurse for help regardless of which arm they were randomized into.

The sessions did require participation, effort and time, but these costs were small in magnitude, and always under the control of the participants. Participants were not required to attend sessions, and there was no consequence to them for non-attendance

4. Potential harms to research participants or research staff from data collection (e.g., surveying, privacy, data management) or research protocols (e.g., random assignment)

Are data collection and/or research procedures adherent to privacy, confidentiality, risk-management, and informed consent protocols with regard to human subjects? Are they respectful of community norms, e.g., community consent not merely individual consent, when appropriate? Are there potential harms to research staff from conducting the data collection that are beyond "normal" risks?

Example of sub-questions to consider as part of the broad question: Are there any risks that could ensue because of the data collection process or storage, e.g. discomfort to being asked certain questions or breach of confidentiality? If so, what are the mitigation strategies? Are there costs to the participant for the data collection process, such as their time, and if so, what is the strategy or rationale for offsetting this cost?

Because these are all issues covered by most IRB processes, a sufficient explanation for a "yes" response may be to provide the IRB approval numbers for all IRBs that have approved the project. However, if there are particular issues that are important to discuss, please do so here.

Harms to research staff could include, e.g., exposure to political violence, exposure to unusual levels of a communicable disease, mistrust due to lack of perceived lack of community consent, or emotional wellbeing from surveying about difficult subject matters. This would not include, e.g., traffic accidents.

Data collection procedures were in adherence with human subjects protocols and respectful of community norms. There were no special risks to research staff.

5. Financial and reputational conflicts of interest Do any of the researchers have financial conflicts of interest with regard to the results of the research? Do any of the researchers have potential reputational conflicts of interest?

We define financial conflicts of interest as that used by the researcher's institutional (e.g., their university) guidelines. We define a reputational conflict of interest as one in which prior writing or advocacy could be contradicted by specific results pursued in this study, and such contradiction would pose reputational risks to the author.

None.

6. Intellectual freedom

Were there any contractual limitations on the ability of the researchers to report the results of the study? If so, what were those restrictions, and who were they from?

This could include, for example, approval of release of the paper and restrictions on data release, but does not include things such as a "comment period" during which interested parties have a right to review and provide comments prior to release but not to control the outputs of the study.

No restrictions.

7. Feedback to participants or communities

Is there a plan for providing feedback on research results to participants or communities? If yes, what is the plan? If not, why not?

Engaging in post-study feedback is a way of acknowledging the agency of participants and communities, and is thus a desired practice. However, it may be impractical due to costs, timing, challenges communicating the results, or potential harms if such communication may itself change behavior in undesirable ways.

We hope to provide feedback as part of the closing procedure for the overall Escaping Poverty research program, of which this is part.

8. Foreseeable misuse of research results

Is there a foreseeable and plausible risk that the results of the research will be misused and/or deliberately misinterpreted by interested parties to the detriment of other interested parties? If yes, please explain any efforts to mitigate such risk.

In settings with strong imbalances of power between interested parties, there may be foreseeable risks that a powerful party could use deliberately selected research findings to

their advantage and to the harm of participants or non-participants, including for general public policy. For example, if the research might reveal the vulnerability of some that can be exploited for the gain of the more powerful party, what steps does the researcher plan to mitigate this risk?

None.

9. Other Ethics Issues to Discuss

None.

Appendix Tables

Appendix Table 1: Comparison of study to Haushofer et al. (2021)

	This paper (1)	Haushofer et al. (2021) (2)
Panel A. Study Context		
Country	Ghana	Kenya
Country GDP per capita	4993	4204
Location within country	Upper East, Northern, Brong Ahafo and Ashanti Regions	Nakuru County
Poverty / Income-Level of Study Area relative to National Levels	Regions in the study (weighted by study sample size) have a poverty rate of 27.9% (per the Ghana Statistical Service's classification); the national rate is 23.6%.	Nakuru County's is the 2nd-wealthiest of 47 counties, with GDP per capita of 6403 USD PPP
Panel. B: Intervention		
Years of Program Activities	2016-2017	2017-2018
Therapy Type	Cognitive Behavioral Therapy	Problem Management Plus, a psychotherapy developed by the World Health Organization, based on Cognitive Behavioral Therapy
Group or individual	Group, target 10 of same gender per group	Individual
Number of sessions	12	5
Length of each session	90 minutes	90 minutes
Counselor characteristics	37 counselors (and assistants) with a Bachelor's Degree, most commonly in pscyhology or development studies	72 Community Health Workers: volunteers who had completed secondary school
Training offered to counselors	Two weeks of classroom training, one week of practice sessions (delivered in communities excluded from study based on size)	9 days of classroom training, 5 supervised training sessions with clients
Panel C. Research Design		
Number of Communities	258	233
Community Selection Criteria	District Assemblies identified communities with high poverty levels, road access, no existing graduation programs, census verified 45+ compounds in community	Partner NGO selected villages in which they were prepared to work, Nakuru County chosen due to high levels of poverty, high baseline rates of poor mental health, and existing NGO presence
Sample Size	7227	5756
Number of individuals receiving therapy	1290	1018
Household Selection Criteria	Households in 40 poorest compounds,in census	Households without brick, stone, or metal walls
Panel D. Results		
Length of time between end of intervention and endline survey	1-3 months	2-23 months (mean 12.63, median 13)
Outcome Variable(s) used to Measure Distress	Kessler Psychological Distress Scale (K10)	12-item General Health Questionnaire (GHQ-12), Perceived Stress Scale (Cohen)
Average Treatment Effects	0.17 (Randomization Inference p-val = 0.000)	0.03 for GHQ-12 (SE = 0.06), 0.02 for Cohen (SE = 0.06)

This table compares our study and the study reported in Haushofer et al (2021)'s "The Comparative Impact of Cash Transfers and a Psychotherapy Program on Psychological and Economic Well-being." Measures of GDP PPP per capita come from the World Bank, numbers and descriptors of Haushofer et al (2021) come directly from the working paper text, accessed January 2022

Appendix Table 2: Variables Used in Re-Randomization Procedures

Panel A: Variables in Re-Randomization to Determine Community-Level Assignment
District-level dummies
Mean proxy means test score
SD of proxy means tests in community
Paved road connected to village
Electricity in village
Distance from nearest market
Number of compounds in community

Panel B: Variables in Re-Randomization to Determine Final Sample of Households
Male head of household
Number of co-resident co-wives
Proxy means test score
Age of household head
Average proxy means score among HHs in compound
Number of households in compound

Panel C: Variables used in Re-Randomization to Determine CBT Treatment Assignment
Presence of male adult in household
Presence of female adult in household
Age of household head
Number of children under 5
Household size
Cash savings balance
Land owned
Business profits
Any adult skipped meals last month
Total asset value
Total livestock value
Kessler Score, baseline
Missing Kessler Score, baseline
No male head of household present

This table lists the variables used in our re-randomization procedure to determine (A) whether a community is pure control, pure CBT, or full program, (B) which households in pure control and pure CBT communities to sample and include in our study, and (C) which individuals in pure CBT or full program communities were offered the CBT program

Appendix Table 3: Attrition

	(1) Individual Attrited from Sample
Panel A. Attrition by Treatment Status	
Individual Assigned to CBT	0.013
	(0.011)
Sample - Treatment and Control	0
Sample Mean	0.00
Panel B. Correlates of Attrition	
Individual Assigned to CBT	0.0132
	(0.0110)
Household Head Age	-0.0010
	(0.0004)
Number of children under 5 in household	0.0000
	(0.0039)
Household size	-0.0060
	(0.0017)
Household Savings (/1000)	-0.0018
	(0.0101)
Acres owned of land (/1000)	-0.0780
	(0.7980)
Business Profits (/1000)	-0.0211
	(0.0210)
Any Adults Skipped Meals	-0.0002
	(0.0001)
Asset Value (/1000)	0.0072
	(0.0041)
Livestock Value (/1000)	-0.0017
	(0.0016)
All Adults are Female	-0.0552
	(0.0154)
Male Kessler Score	0.0009
	(0.0006)
Female Kessler Score	-0.0005
	(0.0006)
Panel C. Test of Differences in Attrition Correlates by Treatment	
Treatment Status	0.0199
	(0.015)
Baseline Characteristics?	Yes
Baseline Characteristics interacted with treatment?	Yes
F-Stat: Treatment + Treatment Interactions Jointly Equal 0	0
p-value: Treatment + Treatment Interactions Jointly Equal 0	0

Panel A reports regression results of whether or not an individual attrited from the sample on treatment status, with attrition as the dependent variable. Panel B regresses attrition on several correlates, again including treatment status. Panel C reports the joint F-Test from a regression of attrition on the correlates in Panel B interacted with treatment. In all cases, standard errors are clustered at the village level.

Appendix Table 4: CBT Heterogeneous Treatment Effects: Interaction with Baseline Kessler Score - Health

	Control Mean	Individual Received CBT	Received CBT * Baseline Kessler (standardized)
	(1)	(2)	(3)
Panel A: Mental Health Outcomes			
Mental Health Index	0.00	0.14	0.03
RI p-value		[0.000]	[0.349]
Kessler Score	21.53	-1.25	-0.24
RI p-value		[0.000]	[0.439]
No distress (Kessler < 20)	0.46	0.05	0.00
RI p-value		[0.014]	[0.795]
No moderate or severe distress (Kessler < 25)	0.68	0.06	0.00
RI p-value		[0.003]	[0.902]
No severe distress (Kessler <30)	0.84	0.03	0.03
RI p-value		[0.018]	[0.022]
Mental Health Self Rating (1/4)	2.90	0.07	0.01
RI p-value		[0.068]	[0.799]
30 minus days in month with poor mental health	24.85	0.51	0.27
RI p-value		[0.113]	[0.382]
Panel B: Perceived Physical Health and Effects on Labor			
Perceived Physical Health and Labor Index	0.00	0.12	0.02
RI p-value		[0.000]	[0.615]
Physical Health Self-Rating (1/4)	3.05	0.11	0.02
RI p-value		[0.000]	[0.513]
30 minus days in month with poor physical health	24.73	0.83	0.28
RI p-value		[0.003]	[0.387]
30 minus days in month in which poor mental or physical health limited labor or normal activities	26.09	0.32	-0.120
RI p-value		[0.204]	[0.667]

Each row for Columns 2-3 are from a single specification with between 6,723 and 6,767 observations, in which the outcome is regressed on treatment status, a continuous measure of the baseline Kessler Score (standardized to mean 0, standard deviation 1) and the interaction between the Kessler Score and treatment. The coefficients reported here are (a) the coefficient on treatment status, and (b) the interaction between treatment and baseline distress. Both p-values (in columns 3 and 4) are calculated via randomization inference, in which we re-run our full randomization procedure to assign placebo treatments, and compare our true estimates to the placebo distribution of estimates; the full procedure is described in Appendix B.

Appendix Table 5: CBT Heterogeneous Treatment Effects: Interaction with Baseline Kessler Score - Bandidth and Economic Perceptions

	Control Mean	Individual Received CBT	Received CBT * Baseline Kessler (standardized)
	(1)	(2)	(3)
Panel A: Socioemotional Skills			
Socioemotional Skill Index	0.00	0.26	0.02
RI p-value		[0.204]	[0.667]
Generalized Self-Efficacy Score	0.00	0.29	0.00
RI p-value		[0.000]	[0.686]
Grit Score	0.00	0.19	0.02
RI p-value		[0.000]	[0.916]
Self-Control Score	0.01	0.11	0.02
RI p-value		[0.000]	[0.623]
Panel B: Cognition			
Cognition Index	0.00	0.08	-0.03
RI p-value		[0.009]	[0.667]
Raven's Progressive Matrices, Indexed	0.00	0.04	0.001
RI p-value		[0.016]	[0.331]
Digit Span: Forwards, Indexed	-0.01	0.07	-0.05
RI p-value		[0.411]	[0.981]
Digit Span: Backwards, Indexed	0.00	0.06	0.003
RI p-value		[0.045]	[0.147]
Executive Function Test, Indexed	0.01	0.05	-0.03
RI p-value		[0.072]	[0.938]
Panel C: Economic Self-Perception			
Economic Index	0.01	0.17	-0.04
RI p-value		[0.215]	[0.398]
Self-Reported Economic Status	3.06	0.38	-0.12
RI p-value		[0.000]	[0.242]
Projected Economic Status in 5 years	5.73	0.32	-0.07
RI p-value		[0.000]	[0.146]

Each row for Columns 2-3 are from a single specification with between 6,758 and 6,767 observations, in which the outcome is regressed on treatment status, a continuous measure of the baseline Kessler Score (standardized to mean 0, standard deviation 1) and the interaction between the Kessler Score and treatment. The coefficients reported here are (a) the coefficient on treatment status, and (b) the interaction between treatment and baseline distress. Both p-values (in columns 3 and 4) are calculated via randomization inference, in which we re-run our full randomization procedure to assign placebo treatments, and compare our true estimates to the placebo distribution of estimates; the full procedure is described in Appendix B.

Appendix Table 6: CBT Treatment Effects by Gender- Health Outcomes

	Control Mean	CBT Average Treatment Effect, Female	CBT Average Treatment Effect, Male	p-value from Test: Homogenous Treatment Effect by Gender, 2=3
	(1)	(2)	(3)	(4)
Panel A: Mental Health Outcomes				
Mental Health Index	0.00	0.13	0.16	
RI p-value		[0.009]	[0.006]	[0.757]
Kessler Score	21.53	-1.32	-1.33	
RI p-value		[0.002]	[0.003]	[0.983]
No distress (Kessler < 20)	0.46	0.06	0.05	
RI p-value		[0.028]	[0.094]	[0.854]
No moderate or severe distress (Kessler < 25)	0.68	0.06	0.06	
RI p-value		[0.018]	[0.032]	[0.919]
No severe distress (Kessler <30)	0.84	0.04	0.03	
RI p-value		[0.061]	[0.122]	[0.799]
Mental Health Self Rating (1/4)	2.90	0.03	0.12	
RI p-value		[0.606]	[0.033]	[0.227]
30 minus days in month with poor mental health	24.85	0.76	0.33	
RI p-value		[0.073]	[0.480]	[0.506]
Panel B: Perceived Physical Health and Effects on Labor				
Perceived Physical Health and Labor Index	0.00	0.14	0.12	
RI p-value		[0.004]	[0.016]	[0.790]
Physical Health Self-Rating (1/4)	3.05	0.10	0.14	
RI p-value		[0.015]	[0.001]	[0.470]
30 minus days in month with poor physical health	24.73	1.04	0.73	
RI p-value		[0.004]	[0.083]	[0.577]
30 minus days in month in which poor mental or physical health limited labor or normal activities	26.09	0.499	0.206	
RI p-value		[0.139]	[0.595]	[0.556]

Each row for Columns 2-3 are from a single specification with between 7,205 and 7,253 observations, which include a dummy indicator for female, and interactions between (a) female and being offered CBT, and (b) male and being offered CBT. Column 4 reports the p-value from the test that the coefficients in columns 2 and 3 are equal. All p-values (in each of columns 2, 3, and 4) are calculated via randomization inference, in which we re-run our full randomization procedure to assign placebo treatments, and compare our true estimates to the placebo distribution of estimates; the full procedure is described in Appendix B.

Appendix Table 7: CBT Treatment Effects by Gender - Bandwidth and Economic Perceptions

	Control Mean	CBT Average Treatment Effect, Female	CBT Average Treatment Effect, Male	p-value from Test: Homogenous Treatment Effect by Gender, 2=3
	(1)	(2)	(3)	(4)
Panel A: Socioemotional Skills				
Socioemotional Skill Index	0.00	0.25	0.28	
RI p-value		[0.139]	[0.595]	[0.556]
Generalized Self-Efficacy Score	0.00	0.27	0.31	
RI p-value		[0.000]	[0.000]	[0.724]
Grit Score	0.00	0.20	0.17	
RI p-value		[0.000]	[0.000]	[0.643]
Self-Control Score	0.01	0.09	0.14	
RI p-value		[0.000]	[0.007]	[0.758]
Panel B: Cognition				
Cognition Index	0.00	0.04	0.11	
RI p-value		[0.141]	[0.025]	[0.544]
Raven's Progressive Matrices, Indexed	0.00	0.06	-0.03	
RI p-value		[0.391]	[0.040]	[0.329]
Digit Span: Forwards, Indexed	-0.01	0.02	0.12	
RI p-value		[0.280]	[0.639]	[0.301]
Digit Span: Backwards, Indexed	0.00	0.03	0.09	
RI p-value		[0.634]	[0.025]	[0.215]
Executive Function Test, Indexed	0.01	0.00	0.10	
RI p-value		[0.419]	[0.095]	[0.452]
Panel C: Economic Self-Perception				
Economic Index	0.01	0.18	0.21	
RI p-value		[0.951]	[0.075]	[0.170]
Self-Reported Economic Status	3.06	0.42	0.45	
RI p-value		[0.002]	[0.002]	[0.747]
Projected Economic Status in 5 years	5.73	0.30	0.39	
RI p-value		[0.001]	[0.002]	[0.873]

Each row for Columns 2-3 are from a single specification with between 7,247 and 7,253 observations, which include a dummy indicator for female, and interactions between (a) female and being offered CBT, and (b) male and being offered CBT. Column 4 reports the p-value from the test that the coefficients in columns 2 and 3 are equal. All p-values (in each of columns 2, 3, and 4) are calculated via randomization inference, in which we re-run our full randomization procedure to assign placebo treatments, and compare our true estimates to the placebo distribution of estimates; the full procedure is described in Appendix B.

Appendix Table 8: Average Treatment Effects on Mental Health, by Control Group Definition

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mental Health Index	Kessler Score	No distress (Kessler < 20)	No moderate or severe distress (Kessler < 25)	No severe distress (Kessler <30)	Mental Health Self Rating (1/4)	30 minus days in month with poor mental health
Panel A: Treatment vs All Control							
Assigned to CBT	0.15	-1.36	0.06	0.06	0.04	0.07	0.53
	(0.029)	(0.271)	(0.017)	(0.016)	(0.012)	(0.029)	(0.249)
Observations	7,227	7,221	7,221	7,221	7,221	7,227	7,195
R-squared	0.10	0.09	0.06	0.05	0.03	0.04	0.03
Control mean	0.00	21.5	0.46	0.68	0.84	2.9	24.9
Panel B: Treatment vs Pure Control							
Assigned to CBT	0.12	-1.18	0.051	0.053	0.031	0.060	0.30
	(0.044)	(0.385)	(0.023)	(0.023)	(0.016)	(0.045)	(0.337)
Observations	3,523	3,523	3,523	3,523	3,523	3,523	3,511
R-squared	0.10	0.09	0.06	0.05	0.03	0.04	0.03
Control mean	0.04	21.3	0.47	0.69	0.85	2.9	25.2
Panel C: Spillover Treatment vs Spillover Control							
Assigned to CBT	-0.05	0.28	-0.003	-0.017	-0.011	-0.012	-0.49
	(0.047)	(0.413)	(0.023)	(0.023)	(0.017)	(0.043)	(0.352)
Observations	5,952	5,946	5,946	5,946	5,946	5,952	5,925
R-squared	0.10	0.09	0.06	0.05	0.04	0.04	0.03
Control mean	0.01	21.6	0.44	0.69	0.84	2.9	25.3
Panel D: Restricting In-Village Control to Same Gender as CBT Recipients in Community							
Assigned to CBT	0.13	-1.22	0.05	0.05	0.03	0.07	0.42
	(0.032)	(0.291)	(0.018)	(0.017)	(0.013)	(0.033)	(0.264)
Observations	5,287	5,284	5,284	5,284	5,284	5,287	5,268
R-squared	0.10	0.08	0.06	0.04	0.03	0.04	0.03
Control mean	0.02	21.3	0.47	0.69	0.84	2.9	25.0

Panel A presents results using the sample from our main analysis, in which we include all control individuals (both individuals in control villages, and control individuals in pure CBT and full program communities). Panel B restricts the control group to individuals in control villages (i.e. fully eliminating the possibility of within-village spillovers, at the cost of a reduced sample). Panel C tests for in-village spillovers, by comparing individuals in pure CBT or full program communities who did not receive the program to individuals in pure control communities. Panel D restricts our control sample to the same gender as the CBT program in a given community (ie omits male control individuals in female CBT communities, and vice versa). In all specifications, standard errors are clustered at the village level.

Appendix Table 9: Average Treatment Effects on Physical Health, by Control Group Definition

	(1)	(2)	(3)	(4)
	Physical Health Index	Physical Health Self Rating (1/4)	Days in month without poor physical health	30 minus days in month in which poor mental or physical health limited labor or normal activities
Panel A: Treatment vs All Control				
Assigned to CBT	0.13	0.12	0.89	0.34
	(0.029)	(0.024)	(0.227)	(0.221)
Observations	7,227	7,227	7,199	7,179
R-squared	0.11	0.10	0.06	0.04
Control mean	0.00	3.05	24.73	26.09
Panel B: Treatment vs Pure Control				
Assigned to CBT	0.10	0.09	0.75	0.28
	(0.035)	(0.031)	(0.256)	(0.264)
Observations	3,523	3,523	3,515	3,505
R-squared	0.09	0.10	0.05	0.04
Control mean	0.02	3.06	24.86	26.17
Panel C: Spillover Treatment vs Spillover Control				
Assigned to CBT	-0.04	-0.04	-0.22	-0.10
	(0.032)	(0.030)	(0.239)	(0.237)
Observations	5,952	5,952	5,928	5,912
R-squared	0.11	0.10	0.06	0.05
Control mean	-0.02	3.05	25.47	26.69
Panel D: Restricting In-Village Control to Same Gender as CBT Recipients in Community				
Assigned to CBT	0.11	0.11	0.81	0.26
	(0.031)	(0.025)	(0.241)	(0.231)
Observations	5,287	5,287	5,273	5,258
R-squared	0.10	0.10	0.05	0.04
Control mean	0.01	3.06	24.82	26.21

Panel A presents results using the sample from our main analysis, in which we include all control individuals (both individuals in control villages, and control individuals in pure CBT and full program communities). Panel B restricts the control group to individuals in control villages (i.e. fully eliminating the possibility of within-village spillovers, at the cost of a reduced sample). Panel C tests for in-village spillovers, by comparing individuals in pure CBT or full program communities who did not receive the program to individuals in pure control communities. Panel D restricts our control sample to the same gender as the CBT program in a given community (ie omits male control individuals in female CBT communities, and vice versa). In all specifications, standard errors are clustered at the village level.

Appendix Table 10: Average Treatment Effects on Socio-Emotional Skills, by Control Group Definition

	(1) Socioemotional Skill Index	(2) Generalized Self-Efficacy Score	(3) Grit Score	(4) Self-Control Score
Panel A: Treatment vs All Control				
Assigned to CBT	0.27	0.29	0.19	0.12
	(0.035)	(0.033)	(0.033)	(0.036)
Observations	7,226	7,226	7,223	7,218
R-squared	0.09	0.06	0.06	0.07
Control mean	0.00	0.00	0.00	0.01
Panel B: Treatment vs Pure Control				
Assigned to CBT	0.33	0.39	0.20	0.15
	(0.052)	(0.046)	(0.050)	(0.051)
Observations	3,523	3,523	3,523	3,523
R-squared	0.10	0.08	0.06	0.07
Control mean	-0.05	-0.11	-0.01	-0.02
Panel C: Spillover Treatment vs Spillover Control				
Assigned to CBT	0.10	0.16	0.01	0.06
	(0.050)	(0.044)	(0.049)	(0.047)
Observations	5,951	5,951	5,948	5,943
R-squared	0.08	0.06	0.05	0.07
Control mean	0.05	0.07	0.02	0.03
Panel D: Restricting In-Village Control to Same Gender as CBT Recipients in Community				
Assigned to CBT	0.28	0.32	0.19	0.13
	(0.038)	(0.036)	(0.036)	(0.039)
Observations	5,286	5,286	5,284	5,280
R-squared	0.09	0.07	0.06	0.07
Control mean	-0.01	-0.03	0.00	0.01

Panel A presents results using the sample from our main analysis, in which we include all control individuals (both individuals in control villages, and control individuals in pure CBT and full program communities). Panel B restricts the control group to individuals in control villages (i.e. fully eliminating the possibility of within-village spillovers, at the cost of a reduced sample). Panel C tests for in-village spillovers, by comparing individuals in pure CBT or full program communities who did not receive the program to individuals in pure control communities. Panel D restricts our control sample to the same gender as the CBT program in a given community (ie omits male control individuals in female CBT communities, and vice versa). In all specifications, standard errors are clustered at the village level.

Appendix Table 11: Average Treatment Effects on Cognition, by Control Group Definition

	(1)	(2)	(3)	(4)	(5)
	Cognition Index	Raven's Progressive Matrices, Indexed	Digit Span: Forwards, Indexed	Digit Span: Backwards, Indexed	Executive Function Test, Indexed
Panel A: Treatment vs All Control					
Assigned to CBT	0.08	0.03	0.080	0.071	0.051
	(0.034)	(0.034)	(0.032)	(0.031)	(0.034)
Observations	7,227	7,222	7,222	7,222	7,227
R-squared	0.09	0.06	0.06	0.07	0.00
Control mean	0.00	0.00	0.00	0.01	0.00
Panel B: Treatment vs Pure Control					
Assigned to CBT	0.082	0.008	0.084	0.093	0.047
	(0.046)	(0.054)	(0.045)	(0.039)	(0.044)
Observations	3,523	3,521	3,523	3,523	3,523
R-squared	0.10	0.08	0.06	0.07	0.00
Control mean	-0.05	-0.11	-0.01	-0.02	0.00
Panel C: Spillover Treatment vs Spillover Control					
Assigned to CBT	-0.01	-0.04	0.013	0.032	-0.009
	(0.039)	(0.048)	(0.042)	(0.032)	(0.037)
Observations	5,952	5,947	5,947	5,947	5,952
R-squared	0.08	0.06	0.05	0.07	0.00
Control mean	0.05	0.07	0.02	0.03	0.00
Panel D: Restricting In-Village Control to Same Gender as CBT Recipients in Community					
Assigned to CBT	0.10	0.03	0.100	0.095	0.063
	(0.036)	(0.038)	(0.035)	(0.032)	(0.037)
Observations	5,287	5,284	5,285	5,285	5,287
R-squared	0.09	0.07	0.06	0.07	0.00
Control mean	-0.01	-0.03	0.00	0.01	0.00

Panel A presents results using the sample from our main analysis, in which we include all control individuals (both individuals in control villages, and control individuals in pure CBT and full program communities). Panel B restricts the control group to individuals in control villages (i.e. fully eliminating the possibility of within-village spillovers, at the cost of a reduced sample). Panel C tests for in-village spillovers, by comparing individuals in pure CBT or full program communities who did not receive the program to individuals in pure control communities. Panel D restricts our control sample to the same gender as the CBT program in a given community (ie omits male control individuals in female CBT communities, and vice versa). In all specifications, standard errors are clustered at the village level.

Appendix Table 12: Average Treatment Effects on Economic Perceptions, by Control Group Definition

	(1)	(2)	(3)
	Economic Index	Self-Reported Economic Status (1/10)	Projected Economic Status in 5 years (1/10)
Panel A: Treatment vs All Control			
Assigned to CBT	0.20	0.44	0.36
	(0.038)	(0.076)	(0.100)
Observations	7,227	7,227	7,227
R-squared	0.06	0.05	0.05
Control mean	0.01	3.06	5.73
Panel B: Treatment vs Pure Control			
Assigned to CBT	0.21	0.51	0.35
	(0.054)	(0.106)	(0.144)
Observations	3,523	3,523	3,523
R-squared	0.07	0.06	0.06
Control mean	0.01	3.02	5.79
Panel C: Spillover Treatment vs Spillover Control			
Assigned to CBT	0.03	0.13	-0.03
	(0.050)	(0.093)	(0.138)
Observations	5,952	5,952	5,952
R-squared	0.05	0.04	0.05
Control mean	-0.01	3.03	5.80
Panel D: Restricting In-Village Control to Same Gender as CBT Recipients in Community			
Assigned to CBT	0.20	0.44	0.35
	(0.041)	(0.083)	(0.109)
Observations	5,287	5,287	5,287
R-squared	0.06	0.05	0.05
Control mean	0.02	3.07	5.76

Panel A presents results using the sample from our main analysis, in which we include all control individuals (both individuals in control villages, and control individuals in pure CBT and full program communities). Panel B restricts the control group to individuals in control villages (i.e. fully eliminating the possibility of within-village spillovers, at the cost of a reduced sample). Panel C tests for in-village spillovers, by comparing individuals in pure CBT or full program communities who did not receive the program to individuals in pure control communities. Panel D restricts our control sample to the same gender as the CBT program in a given community (ie omits male control individuals in female CBT communities, and vice versa). In all specifications, standard errors are clustered at the village level.