Rose, Julian; Neubauer, Florian; Ankel-Peters, Jörg

Working Paper

Long-term effects of the targeting the ultra-poor program: A reproducibility and replicability assessment of Banerjee et al. (2021)

Ruhr Economic Papers, No. 1107

Provided in Cooperation with:
RWI – Leibniz-Institut für Wirtschaftsforschung, Essen

Suggested Citation: Rose, Julian; Neubauer, Florian; Ankel-Peters, Jörg (2024) : Long-term effects of the targeting the ultra-poor program: A reproducibility and replicability assessment of Banerjee et al. (2021), Ruhr Economic Papers, No. 1107, ISBN 978-3-96973-285-4, RWI - Leibniz-Institut für Wirtschaftsforschung, Essen,
https://doi.org/10.4419/96973285

This Version is available at:
https://hdl.handle.net/10419/306837

RUHR
ECONOMIC PAPERS

Julian Rose
Florian Neubauer
Jörg Ankel-Peters

Long-Term Effects of the Targeting the Ultra-Poor Program – A Reproducibility and Replicability Assessment of Banerjee et al. (2021)

RWI #1107

Imprint

Ruhr Economic Papers

Published by

RWI – Leibniz-Institut für Wirtschaftsforschung
Hohenzollernstr. 1-3, 45128 Essen, Germany

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, e-mail: W.Leininger@tu-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Ronald Bachmann, Prof. Dr. Almut Balleer, Prof. Dr. Manuel Frondel, Prof. Dr. Ansgar Wübker
RWI, Phone: +49 (0) 201/81 49-213, e-mail: presse@rwi-essen.de

Editorial Office

Sabine Weiler
RWI, Phone: +49 (0) 201/81 49-213, e-mail: sabine.weiler@rwi-essen.de

Ruhr Economic Papers #1107

Responsible Editor: Manuel Frondel

ISSN 1864-4872 (online) – ISBN 978-3-96973-285-4

Ruhr Economic Papers #1107

Julian Rose, Florian Neubauer, and Jörg Ankel-Peters

Long-Term Effects of the Targeting the Ultra-Poor Program – A Reproducibility and Replicability Assessment of Banerjee et al. (2021)

RWI

Bibliografische Informationen der Deutschen Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at **http://dnb.dnb.de**

RWI is funded by the Federal Government and the federal state of North Rhine-Westphalia.

https://dx.doi.org/10.4419/96973285
ISSN 1864-4872 (online)
ISBN 978-3-96973-285-4

Julian Rose, Florian Neubauer, and Jörg Ankel-Peters*

Long-Term Effects of the Targeting the Ultra-Poor Program – A Reproducibility and Replicability Assessment of Banerjee et al. (2021)

Abstract

Banerjee, Duflo, and Sharma (BDS, 2021a) conduct a ten-year follow-up of a randomized transfer program in West Bengal. BDS find large effects on consumption, food security, income, and health. We conduct a replicability assessment. First, we successfully reproduce the results, thanks to a perfectly documented reproduction package. Results are robust across alternative specifications. We furthermore assess the paper's pre-specification diligence and the reporting in terms of extern.

JEL-Codes: A1, O12

Keywords: Replicability; randomized controlled trial; transfer programs; research transparency

September 2024

** Julian Rose, RWI, University of Passau, and LMU; Jörg Ankel-Peters, RWI and University of Passau; Florian Neubauer, RWI.* *– We thank the original authors for patiently responding to our questions and commenting on our paper. Funding: This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), Grant No. 3473/1-1 within the DFG Priority Program META-REP (SPP 2317).* *– All correspondence to: Julian Rose, RWI, Hohenzollernstraße 1–3, 45128 Essen, Germany, e-mail: julian.rose@rwi-essen.de*

1. Introduction

Banerjee, Duflo, and Sharma (2021a), henceforth BDS, evaluate the ten-year effects of the "Targeting the Ultra Poor"-program (TUP) in West Bengal, India. The evaluation of short-term effects of TUP, after three years, were published as a multi-country study in Banerjee et al. (2015). TUP provided a large productive asset transfer to selected households alongside weekly consumption support in cash, a saving component, and training visits. The program's underlying idea was that people are stuck in a poverty trap, from which they should be released by the intervention. In West Bengal, a local NGO implemented the program in 120 village hamlets in 2007, with randomization taking place on the household level. Eligible households had to be in the bottom wealth quintile, have no credit access, and a female member to manage the asset provided by the program.

Our paper presents results of a reproducibility and replicability assessment, which acts as a pilot for a large-scale reproduction project called *Replicability and Robustness in Economics* (R2E). The assessment adheres closely to our newly developed standardized protocol (see Ankel-Peters et al. 2024). First, the assessment starts by a computational reproduction using the author's code and data. Second, we perform a robustness reproduction through multiple robustness checks. Third, we evaluate the pre-specification of the analysis. Fourth, we assess the reporting in terms of external and construct validity.

Our replicability assessment makes a proposal for how to scrutinize a paper to comprehensively cover different reproducibility and replicability dimensions. It is important to note that this assessment covers transparency dimensions that are to varying degrees common standards in economics (Christensen and Miguel 2018). Hence, not meeting the criteria we apply in this paper does not necessarily imply that common standards in economics are not met. Computational reproducibility, in fact, is now common practice at journals published by the *American Economic Association* (AEA) and in this step, we only redo what the AEA data editor does (see Vilhuber 2020). While only few other journals have a data editor, most leading journals have a data and code sharing policy and there is a clear consensus that published work should be computationally reproducible. In terms of robustness reproducibility, it is less clear what the standards are. There are growing concerns about researchers' degrees of freedom leading to robustness issues in economics (e.g., Brodeur et al. 2020; 2023; 2024a; Huntington-Klein et al. 2021), but there is no consensus on how robustness

should be demonstrated (Ankel-Peters et al. 2023; Simonsohn et al. 2020). We apply a specification curve approach, which is not the norm in economics, and show a reproducibility dashboard concisely summarizing the robustness results. Pre-specification, just as data and code sharing, has become the standard in economics for experiments (Miguel 2021; Ferguson et al. 2023), although there are ongoing debates about how vigorously this should be done (Banerjee et al. 2020, Brodeur et al. 2024b; Ofosu and Posner 2023). Not least, detailed reporting in terms of external and construct validity are uncommon in economics (Masselus et al. 2024; Peters et al. 2018), despite a wide agreement that both are important for policy makers who use the study results to inform their decisions (Gechter 2024; Pritchett and Sandefur 2015; Pritchett et al. 2013; Vivalt 2020; Vivalt and Coville 2023).

BDS build on a randomized controlled trial (RCT) originally used by Banerjee et al. (2015) to examine the program's effect on various socioeconomic outcomes. Banerjee et al. (2015) reports results from similar RCTs on TUP in six countries: West Bengal in India (under evaluation in BDS), Ethiopia, Ghana, Honduras, Pakistan, and Peru. In West Bengal, a local NGO implemented the program TUP, which originally had been designed by BRAC in Bangladesh. In 2007, in a total sample size of 978 households, 514 households were randomly assigned to the treatment group, with randomization stratified at the hamlet level. Treated households could select from different productive assets, such as livestock or non-farm microenterprise inventory, all with approximately equal monetary value. Additionally, treated households received a weekly consumption support of USD 7.60 for a maximum of 40 weeks, access to savings, and weekly training visits for 18 months. Training encompassed topics related to income generation, life skills, and health information and was executed by the implementing NGO. Most households opted for livestock over non-farm inventory (82%), while 248 households that had been selected for the treatment group declined treatment participation.

For the short term, Banerjee et al. (2015) find substantial positive effects on a variety of different outcomes in the West Bengal RCT. For the ten-year follow-up, BDS report in the abstract: "*we find positive effects on consumption (0.6 SD), food security (0.1 SD),* income *(0.3 SD), and health (0.2 SD).*" Because these four outcomes are prominently presented in the abstract, we focus our robustness reproduction on these as the main outcomes (detailed in Table 1). All four outcomes are measured with indices.

Table 1: Key results of BDS

	(1) Consumption	(2) Food Security	(3) Income and revenues	(4) Physical Health
Name of display item in BDS	Table 1	Table 1	Table 1	Table 1
Column	Column 2 - Panel D	Column 3 - Panel D	Column 4 - Panel D	Column 6 - Panel D
Estimate	0.579	0.127	0.264	0.187
Standard Error	0.175	0.062	0.08	0.04
p-value	0.001	0.04	0.001	0.00
Confidence Interval (95%)	[0.235 – 0.922]	[0.003 – 0.250]	[0.106 – 0.422]	[0.109 – 0.265]
Level of analysis	Household	Household	Household	Individual
Number of observations	880	885	885	1229
Sample	Full sample	Full sample	Full sample	Full sample
Type of variable	Standardized index	Standardized index	Standardized index	Standardized index
Unit of outcome	Baseline standard deviation	Baseline standard deviation	Baseline standard deviation	Baseline standard deviation
Estimation method	OLS	OLS	OLS	OLS
Fixed Effects	Hamlet level	Hamlet level	Hamlet level	Hamlet level
Standard Error type	Robust	Robust	Robust	Clustered at household level
Control variables	Baseline value of consumption	Baseline value of food security	Baseline value of income and revenues	Baseline value of physical health

Notes: The table displays details on the ten-year results of BDS. All information is obtained on BDS-Table 1.

In our first step, we computationally reproduce all results of the paper successfully, with only two unmeaningful discrepancies in the original BDS-Table 4.[1] The authors meticulously documented the reproduction package on the journal's website. It contains the raw data and the analysis datasets, alongside Stata do-files for data processing steps (cleaning, preparation, and analysis). Despite increasing transparency standards, the user friendliness of the reproduction package and the provision of raw data are noteworthy.

In our second step, the robustness reproduction, we subject each main outcome to three additional robustness checks[2]:

- Adjusting the index composition in multiple ways to gauge the original results' sensitivity to the index design;
- Introducing survey timing as controls in the regressions to accommodate seasonality across the sample;

[1] See the online appendix for the reproduced BDS-Table 4.

[2] We selected these robustness checks in an ad hoc manner and in a way that they complement the robustness checks provided by BDS.

- Including a full set of baseline household characteristics as controls, addressing income imbalances at baseline, and utilizing the comprehensive with typical control variables.

Our robustness checks strongly confirm the original findings. For consumption, income, and health we observe nearly identical effects regarding both significance and magnitude. Only for food security, the inclusion of control variables renders the effect insignificant at the 5%-level.

In our third step, we check pre-specification. According to BDS' acknowledgements on the paper's title page the study was pre-registered under the name "*Helping the ultra-poor use microcredit in Murshidabad, India*".[3] The pre-registration does not include or mention a pre-analysis plan. We contacted the authors and they confirmed that no PAP exists. In this correspondence, the authors emphasized that the 10-year analysis was conducted in the exact same way as Banerjee et al. (2015). In fact, all indices for the four main outcomes are created identically to Banerjee et al. (2015). Yet, this implicit pre-specification does not apply to the heterogeneity analysis and migration as a mechanism, which is prominently featured in the BDS abstract and which was not examined in Banerjee et al. (2015). Usually, such subgroup analysis calls for detailed pre-specification because it is otherwise unclear whether it is based on ex-post exploratory data analysis or ex-ante theoretical expectations (Banerjee et al. 2020).

BDS do not discuss why the West Bengal study was selected for a long-term follow up. Upon request, the authors clarified that in the 2015 publication they were solely responsible for the intervention in West Bengal and, naturally, conducted the follow-up independently of the other five interventions. Yet, especially since the West Bengal intervention was the most effective one among the six in Banerjee et al. (2015) it would have been desirable to transparently discuss this in the paper to address concerns about a specific selection bias. Barker et al. (2024) conduct a follow-up study on TUP in Ethiopia, the second most successful RCT among the six Banerjee et al. (2015) RCTs, documenting declining effects over time.

In our fourth step, our assessment of construct and external validity reveals two main concerns in the BDS reporting. First, the role and the background of the implementing NGO is not explained in BDS (or in Banerjee et al. 2015). NGOs have been found to be more effective than governmental organizations and, in general, how the treatment is delivered is important for

[3] See https://www.socialscienceregistry.org/trials/382.

generalizing to other settings (Allcot 2015; Angrist and Meager 2023; Bold et al. 2018; Mo et al. 2020; Usmani et al. 2022). Second, BDS lacks a detailed description of the multi-pronged treatment's composition. Without a profound understanding of the treatment, theory-based or reasoned-intuition-based deductive inference becomes difficult about how similar programs might work elsewhere (Basu 2014; Duflo et al. 2007; Esterling et al. 2023; White 2009). In addition, an inductive learning approach based on the accumulation of evidence on the same or similar interventions is difficult without a clear understanding of the context and the treatment (Cartwright 2011; Duflo 2020).

2. Computational reproducibility

The BDS reproduction package is available on the journals' website (BDS 2021b). Using BITSS' Levels of Computational Reproducibility, we assign BDS the highest level on a scale from one to ten (BITSS 2020). BDS meet all criteria by providing analysis code, data, cleaning code, and raw data. Their results are computationally reproducible starting from both the analysis and raw data.

Moreover, the authors provide all questionnaires, a read-me document, as well as all do-files to clean, prepare, and analyze the data alongside a master do-file that executes all do-files in the correct order. This perfectly documented reproduction package ensures a quick computational reproduction by running the master do-file (reproduced tables and figures are displayed in the appendix). We exactly reproduce BDS-Tables 1 to 3 as well as BDS-Figure 1. For BDS-Table 4, we obtain some negligible discrepancies in effect sizes (affecting the third decimal place) and number of observations (two observations).

3. Robustness reproduction

We implement several robustness checks for each main outcome to examine the sensitivity of the results. Each robustness check is conducted separately as well as in all possible combinations. For each outcome, we present the aggregated results in a Robustness Dashboard (see Figure 1; Bensch 2024) and in Specification Curves (see online appendix)[4].

[4] The online appendix is available here: https://osf.io/ag6ez.

Table 2 provides an overview of all robustness checks. The Robustness Dashboard includes the following robustness checks for each main outcome:

- Inclusion of control variable for timing of survey, as the data collection spanned five months and hence seasonality might play a role (Table 2, #2);
- Adding a set of baseline household characteristics as controls to address income imbalances at baseline (see Table A2 in appendix), with these controls including all decomposed indices (Table 2, #3).[5]
- Redefinition of outcome indices by examining alternate compositions. We assess the available variables in the data and create outcome indices that we deem to be reasonable alternatives, such as varying outlier management and including different variables in the indices (Table 2, #4);

For the specification curve analysis, we additionally deconstruct all outcome indices to evaluate effects on each individual component (Table 2, #1).

Figure 1 presents the main results from our robustness reproduction using the Robustness Dashboard proposed by Bensch (2024). The dashboard aggregates the results of the robustness checks for each main outcome of BDS into three categories: *Significant, same sign* represents the share of all robustness checks yielding a significant estimate in the same direction as the original result. This is the share of what one might call successful robustness checks. *Insignificant* depicts the share of robustness checks that turn the estimate insignificant (p-value > 0.05). *Significant, opposite sign* indicating the share of robustness checks producing a significant estimate in the opposite direction to the original result. Additional metrics in the dashboard provide detailed assessments:

- $\tilde{\beta}$: the *relative effect size* indicator puts the median effect size of the robustness checks in relation to the original effect.
- $\overline{\Delta\beta}$: the *effect size variation* indicator measures the mean absolute deviation of effect sizes of all robustness checks in relation to the original effect size.

[5] We include the following baseline controls: productive and household assets; monthly expenditures for food, non-food and durables; whether the household skipped a meal in the past 12 months, whether any adult has not eaten enough in the past 12 months, whether all household member get enough food, and whether everyone eats regularly two meals a day; wage income, formal wage income, income from self-employment, business income, agricultural income; total loans and monthly savings

- $\overline{\Delta p}$: mean change in p-values of all robustness checks leading to insignificant coefficients.

Table 2: Implemented robustness checks

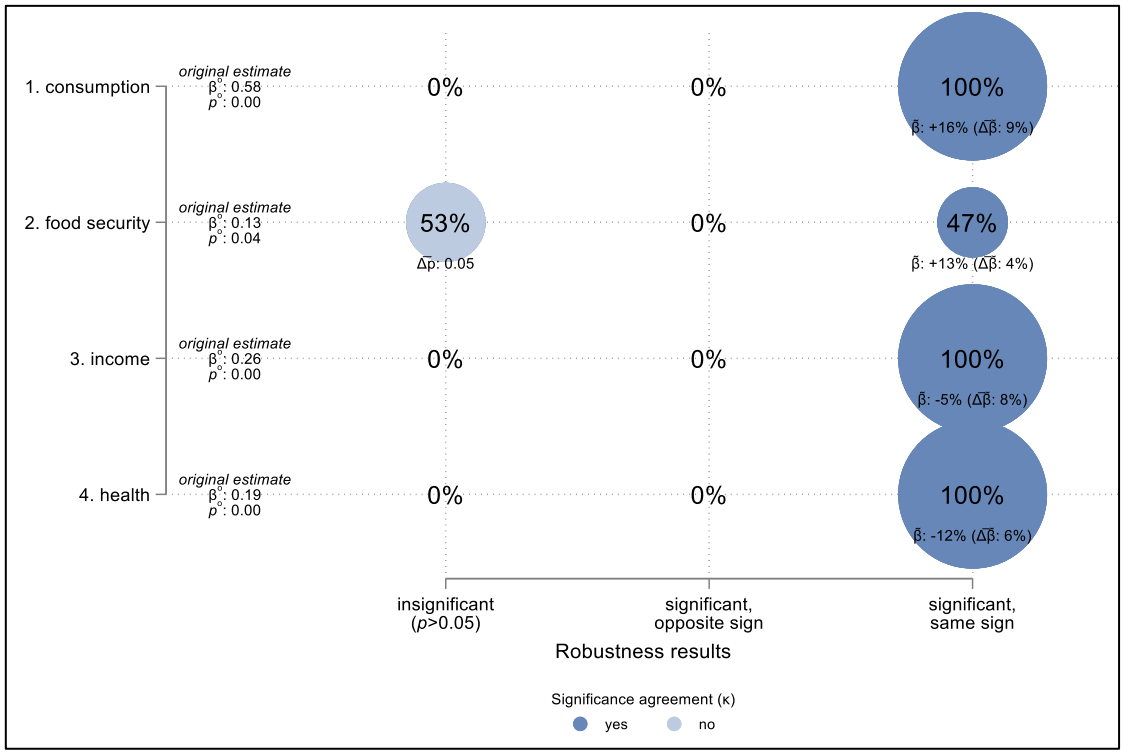
#	Robustness check	BDS choice	Alternative choice options	
			#	Description
1	Outcome variable definition	*Group outcomes into indices consisting of multiple variables*	3-5*	*Components of index as separate outcomes*
	Reflect multiple dimensions of outcomes and check for selective hypothesis testing	BDS construct indices consisting of up to five variables.		Run components of outcome indices separately to understand which ones are driving the effect
2	Controls for timing of survey	*No control for timing of survey*	1	*Control variable for month of interview*
	Account for potentially timing effects of the survey			Add dummy for survey timing to account for timespan of survey (five months).
3	Baseline controls	*No controls for baseline variables*	1	*Baseline controls for full set of baseline household characteristics*
	Account for imbalances at baseline			We add baseline household characteristics as controls to the main specification since there are some slight imbalances between treatment and control group.
4	Outcome variable definition	*Definition of outcome indices*	1-4^	*Multiple Options*
	Reflect multiple dimensions of outcomes and account for multiple hypothesis testing	BDS construct outcome indices.		We slightly vary the definition of the outcome indices based on the available data.

Notes: # = number of alternative choices. * The outcomes indices consist of three to five variables. ^ we define one to four alternative indices.

The results in Figure 1 confirm the robustness of the findings in BDS. For consumption, income, and health, all robustness checks support the original findings. The relative effect size of the robustness checks ($\tilde{\beta}$) is close to the original, with minimal variation of the robustness checks ($\overline{\Delta\beta}$). Only for food security, our robustness checks indicate some sensitivity: 53% of robustness checks render the effect insignificant at the 5%-level. Yet, $\overline{\Delta p}$ (0.05) shows that the p-values in the robustness check change only slightly, leading to p-values around the 10%-level.

In addition to the dashboard analysis, our online appendix reports specification curves for each main outcome, as advocated by Simonsohn et al. (2020).

Figure 1: Robustness Dashboard



Lastly, we assess the role of treated individuals who refused treatment. Of the 514 households offered treatment, only 266 accepted – even though the treatment came at no costs and offered gifts and training, with no perceivable monetary or non-monetary costs. BDS state in the appendix that there were rumors about the NGO aiming to convert beneficiaries to Christianity and "some wives were worried that their husband would mishandle the asset". BDS only show Intent-to-Treat (ITT) effects, comparing households assigned to treatment with those in the control group, regardless of actual treatment reception. We explore whether impacts differ in line with expectations when looking at the *Average Treatment Effect on the Treated* (ATT). Effect sizes notably increase when focusing on the treatment group with actual treatment, underpinning the BDS results (more detailed results can be found in our appendix).

4. Pre-specification

This section of our protocolled assessment examines the degree of adherence to a PAP. BDS' title page refers in the acknowledgments to a pre-registration at the AEA RCT Registry. According to the registry, the study was registered on July 13, 2015 (AEARCTR-0000382) under the title "Helping the ultra-poor use microcredit in Murshidabad, India". From what the registry provides, it is unclear why the study is registered as a microcredit intervention. The treatment comprises a savings component, but it is one of several elements. In fact, the

original authors confirmed upon request that the study was mistakenly registered as a microcredit intervention and that there was no microcredit component.

The AEA RCT Registry does not provide a PAP. Upon inquiry, the authors confirmed the absence of a PAP and noted that the main outcomes align identically with Banerjee et al. (2015). This indeed holds true for the main outcomes, but not for the mechanism analysis in chapter II.C, ("Channels of Persistence"). In Banerjee et al. (2015), there is no mention of migration as either an outcome or mechanism.[6] Here, it is therefore unclear whether the examined mechanisms derive from ex-ante or ex-post theoretical considerations. Lacking an explicit or implicit pre-specification, the mechanism analysis should be explicitly labelled as exploratory, which chapter II.C. in BDS also does. Nevertheless, migration is prominently featured in the abstract and introduction as the main channel of persistence without being labeled as an explorative analysis ("*One main channel for persistence is that treated households take better advantage of opportunities to diversify into more lucrative wage employment, especially through migration*", BDS 2021, p. 471).

In the nascent literature on long-term RCTs it is also important to consider which RCT is selected for a long-term evaluation over others (see Bouguen et al. 2019). Absent explicit pre-specification, a favourable selection of such RCTs that have proven successful in the shorter term while less promising ones are not further examined could lead to a specific type of publication bias. BDS is an interesting case in this regard, as it follows up on one of six RCTs documented in Banerjee et al. (2015). In fact, the West Bengal RCT is the most successful one in the 2015 paper, yielding by far the highest benefits-to-costs ratio at 433%, followed by Ethiopia with 260%. Pakistan and Peru experienced more modest effects, while the Honduras RCT exhibits even negative outcomes. The program in West Bengal yielded by far the highest. Upon request, the original authors explained that for the short-term evaluation published in the 2015 paper they were responsible for the West Bengal leg and selecting this for the long-term follow up occurred naturally. Out of the six RCTs documented in Banerjee et al. (2015), to our knowledge only one other RCT has been evaluated in the long-term, the Ethiopia leg (Barker et al. 2024). What is more, another RCT on TUP that was not part of Banerjee et al. (2015), now in Andhra Pradesh, India, delivered null effects three years after (Bauchet et al.

[6] When providing comments to our paper, the original authors emphasize that migration is highly relevant in West Bengal and a natural candidate for a mechanisms analysis.

2015; see more details in the next section). From a global policy learning perspective, the long-term effects documented in BDS must be embedded into the highly heterogeneous short-term impacts: the a priori odds of such a successful result as it is observed in BDS cannot be expected for a newly implemented program.

5. External validity and construct validity

The two approaches to generalization implicitly used in economics – the deductive and the inductive approach – both require a clear understanding of the context in which certain results were observed and of how the treatment was delivered (Esterling et al. 2023). This understanding is also necessary to predict whether a newly implemented TUP program is more like the successful BDS version of TUP, or rather like the much less successful versions of TUP reported in Banerjee et al. (2015) and Bauchet et al. (2015). Such predictions require adequate information about the intervention and outcome measures (construct validity) and the contextual conditions of the study (external validity). In this section, we examine whether BDS provides the necessary information to assess these dimensions. We also consider and report information that is provided in Banerjee et al. (2015) and BDS' appendix, thereby taking into account that BDS is published as a short paper format.

External validity hinges on the study population's characteristics and how the population's response might change in case the randomized intervention is scaled (see Peters et al. 2018). In BDS, the study population comprises ultra-poor households in rural areas, and the authors provide a comprehensive list of eligibility criteria, clearly defining the study population. This level of detail is valuable for policymakers and researchers, facilitating precise understanding of the study population. General equilibrium effects, in turn, are not discussed, although they are an obvious threat to external validity in case the program is scaled. For example, the livestock grants could deteriorate prices on the local market. Another dimension of external validity are John Henry and Hawthorne effects, which are not discussed in the paper and might be relevant, especially given the randomization at the household level. Related to this, BDS do not indicate whether participants were informed about being part of an experiment. Although informed consent can be expected as the norm, it still matters what participants were told exactly.

For construct validity, the treatment characteristics and how it was delivered are important. Following Esterling et al. (2023), it is essential to understand the role of the intervention as a possible 'causal agent' to pinpoint the treatment components to the observed effect. The BDS treatment is complex and comprises four different components:

1. Productive asset transfer: valued at around USD 437, involving animals or what BDS refer to as "*non-farm microenterprise inventory*". Yet, some information remains vague such as what exactly the offered non-farm enterprise inventory choices included. BDS (p.474) mention only *"…chose a productive asset from a menu of options (two cows, four goats, one cow and two goats, nonfarm microenterprise inventory, etc)*" (BDS, p. 474). Banerjee et al. (2015) do not provide more details either.
2. Weekly consumption support: Households received around USD 7.60 weekly for up to 40 weeks, equating to nearly one day's worth of calories (Banerjee et al. 2015). This is an accurate description.
3. Access to savings: BDS do not provide any information on this component. The online appendix and Banerjee et al. (2015) note that households must save around USD 1 per week during meetings with staff from the implementing NGO, yet it is unclear how this saving is processed and how it can be accessed at a later point.
4. Training visits: For 18 months, households received a weekly visit from the NGO staff "*designed to deliver training on generating income from the chosen asset, life-skills coaching, and health information*" (BDS, p. 474). Details about the training, for example its intensity and content, are lacking, though.

Pritchett (2020) reports that TUP was tweaked and developed during extensive trial and error spadework, suggesting that even minor design deviations could produce different results. Our assessment underlines the complexity of the treatment construct and suggests that its effects may change if the construct changes. Especially the lack of comprehensive information concerning the savings and training components is surprising, given the extensive literature documenting the ambiguous and inconclusive effects of these interventions (e.g., McKenzie et al. 2023). While an academic paper, especially a short format like BDS, certainly cannot cover every detail, some aspects could have been highlighted more prominently or added to the online appendix.

Another dimension of construct validity is how the treatment delivery might deviate from what would be the scaled intervention. Here, *researchers special care* plays an important role (Masselus et al. 2024; Peters et al. 2018). The treatment in BDS was implemented by Bandhan. BDS characterizes Bandhan as a "*nongovernmental organization*" (BDS, p. 474), while according to Banerjee et al. (2015) it is a private sector microfinance institution ("*local MFI*"). Neither BDS nor Banerjee et al. (2015) provide more information on Bandhan. On this note, the fact that nearly 50% of treated households refused the credit, partly attributed by BDS to misconceptions about Bandhan being a Christian organization attempting to convert them, underpins the organization's relevance for generalizability. While Banerjee et al. (2015) list the reasons for refusal, potential implications for the results are neither there nor in BDS discussed.

Regarding generalizability of the BDS results, it is notable that TUP funded by the same source was implemented at ten sites simultaneously, among them the six RCTs reported in Banerjee et al. (2015). One of the remaining four not reported in Banerjee et al. (2015) was randomized as well, in Andhra Pradesh, in India. This intervention is evaluated in Bauchet et al. (2015). Unlike Banerjee et al. (2015) and BDS, they find no impacts on their key outcomes: income, consumption, asset accumulation, and use of financial services. Bauchet et al. (2015) partly attribute the null effects to implementation problems, including a lack of consumption support and lack of customization to individual households. They also emphasize the economic context, noting that a tight labor market and high wages for the control group in wage employment influenced the outcomes. Banerjee et al. (2015) mention the other pilots in a footnote. Here, the authors also briefly state that the Bauchet et al. (2015) RCT was excluded from their paper „*due to the lack of comparability of data*", which is confirmed as one reason for null results by Bauchet et al. (2015). BDS do not mention the Bauchet et al. (2015) study and the conflicting results.

6. Conclusion

In this reproducibility and replicability assessment, we demonstrate that the study is computationally reproducible, and our robustness checks confirm the internal validity of the results. We particularly emphasize the exemplary documentation of the replication package. Our assessment of external and construct validity, though, highlights the importance of other dimensions for causal inference. By carefully examining all available information in BDS and

Banerjee et al. (2015), we identify some shortcomings in the information necessary to draw conclusions beyond the specific context of the study.

The protocol underlying our replicability assessment goes beyond current standards in economics in terms of how empirical work is being reported. This needs to be considered in the interpretation of our assessment. Yet, we propose to broaden the scope in the reporting of a study to include other dimensions of inference that are not related to internal validity. More specifically, adherence to pre-analysis plans should be scrutinized as standard practice – which is not to say that only pre-specified studies and results should be reported and published. What we call for is to transparently distinguish between prespecified and exploratory analysis. Economics papers are also notoriously silent about external validity and especially construct validity. A more diligent description of the treatment and how it was delivered would facilitate a theory-based accumulation of evidence in the academic literature, but also help policymakers to update their priors appropriately.

Appendix

Long-term Effects of the Targeting the Ultra-Poor Program – A Reproducibility and Replicability assessment of Banerjee et al. (2021)

Julian Rose, Florian Neubauer, and Jörg Ankel-Peters

Contents

A. Computational Reproduction

Only for BDS-Table 4, we obtain slightly different results. Table A1 below shows the reproduced table, with deviations highlighted in bold. The differences are marginal and do not affect the overall results. The discrepancy seems to stem from a slight variation in the number of observations, with our reproduced Table 4 containing two additional observations.

Table A1: Reproduction of BDS-Table 4

	Migration	No. of migrants	Duration	Migrates to Kolkata	Migrates to urban area	Earnings of migrant worker, typical month	Working in business or formal work
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Endline 1 (18 months)							
Treatment	**-0.015**	**0.002**	11.767	-0.004	0.002	26.326	0.042
	(0.034)	**(0.041)**	(6.798)	(0.066)	(0.055)	(19.011)	(0.046)
Control Mean	0.35	0.39	37.08	0.36	0.83	139.89	0.10
Observations	**814**	**814**	285	285	285	285	285
Panel B. Endline 2 (three years)							
Treatment	**0.029**	**0.032**	14.776	-0.095	-0.039	30.574	0.032
	(0.032)	(0.041)	(15.332)	(0.069)	(0.059)	(29.920)	(0.042)
Control Mean	0.29	0.33	125.09	0.38	0.83	231.18	0.15
Observations	**840**	**840**	256	256	256	256	256
Panel C. Endline 3 (seven years)							
Treatment	**0.045**	**0.045**	-11.078	0.067	0.012	89.788	0.017
	(0.034)	(0.047)	(12.466)	(0.058)	(0.047)	(33.619)	(0.037)
Control Mean	0.37	0.46	123.26	0.30	0.78	361.21	0.11
Observations	**844**	**844**	332	332	332	332	332
Panel D. Endline 4 (ten years)							
Treatment	**0.015**	**0.022**	25.167	-0.138	0.033	51.238	-0.029
	(0.032)	(0.046)	(12.743)	(0.059)	(0.053)	(31.215)	(0.042)
Control Mean	0.34	0.44	123.78	0.35	0.79	361.95	0.13
Observations	**861**	**861**	**308**	**309**	**309**	**309**	**309**

B. Baseline Balance

Table A2 presents the balancing test on the baseline characteristics. We find slight imbalances regarding the treatment group's income, which is 17% higher at baseline, a difference that is also statistically significant. The division into different income sources shows that the treatment group reports higher income across all sources.

Table A2: Baseline Balance

	(1) Control Mean	(2) Treatment Mean	(3) Diff (T-C)	(4) (1) vs. (2), p-value
Outcome Indices				
Asset Index	-0.04	0.04	0.08	0.24
Total per capita Consumption, standardized	0.01	-0.01	-0.02	0.78
Food Security Index	0.03	-0.02	-0.05	0.45
Financial Inclusion Index	0.01	-0.01	-0.03	0.67
Productive Asset Index	0.00	0.00	0.00	0.99
Household Asset Index	-0.04	0.04	0.08	0.23
Consumption				
Food Consumption per capita, month	26.63	26.88	0.25	0.80
Nonfood Consumption per capita, month	14.38	13.65	-0.73	0.49
Durable goods expenditure per capita, month	0.95	1.00	0.04	0.84
Food Security				
No adults skipped meals	0.09	0.09	0.00	0.88
No one in the HH went a whole day without food	0.28	0.27	-0.01	0.72
No children skipped meals	0.54	0.49	-0.05	0.15
Everyone in HH gets enough food everyday	0.11	0.11	-0.01	0.65
Everyone in the HH regularly eats 2 meals per day	0.77	0.77	0.00	0.92
Income (in USD)				
Wage income (last month)	81.21	88.31	7.10	0.08*
Formal wage income (last month)	1.00	1.92	0.92	0.15
Self-employment income (last month)	14.76	17.69	2.93	0.26
Nonfarm Microenterprise Income (last month)	17.05	26.24	9.19	0.40
Agricultural Profits (last month)	-0.12	0.10	0.22	0.27
Total Income	113.89	134.26	20.36	0.08*
Financial Situation				
Total Outstanding Loans	221.67	219.51	-2.16	0.94
Total Savings (last month)	2.95	1.54	-1.41	0.61

Notes: * implies p < .1, ** implies p < .05, *** implies p < .01.

C. Robustness Reproduction

Here we report and discuss the specification curves for consumption, food security, income, and health. Each specification curve depicts the coefficient and confidence interval for each robustness check and all possible combinations. Each specification curve presented in Figure A1 – Figure A4 is identically structured: Panel A shows the original index alongside various alternative index compositions, while Panel B displays the individual components of the outcome index as outcome variables. Circles represent effect sizes, and the grey areas denote confidence intervals. The original estimate is indicated by a diamond shape. The lower panel details each specification, with filled dots indicating the inclusion of specific robustness checks.

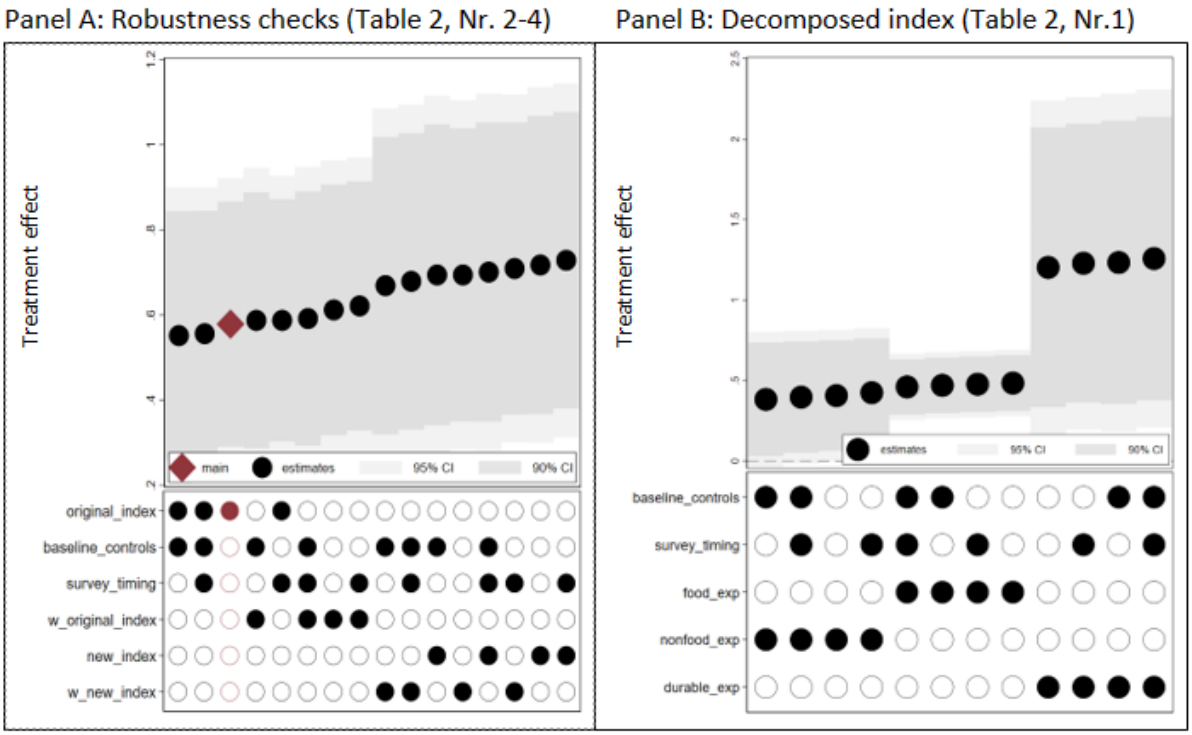
Consumption

Figure A1 presents the specification curve for consumption. We alter the composition of the index in the following ways:

- *w_original_index*: we winsorize the original index by replacing the lowest (highest) 0.01% with the next larger (smaller) value;
- *new_index*: generate a new index by adding up the individual components of the index in Stata (monthly food expenditures, monthly non-food expenditures, and monthly non-durable expenditures) instead of using the sum of those variables generated automatically in the questionnaire;
- *w_new_index*: winsorized version of the newly created index.

In Panel A of Figure 2, the original estimate (diamond-shaped) is placed at the lower end of the distribution suggesting a lower bound for the effect size and underlining the robustness. Panel B decomposes the index, estimating the effect of each component separately, showing significant contributions from each component, with durable expenditures exhibiting the largest effects. The inclusion of control variables in the analysis does not alter significance levels or the effect size of these components.

Figure A1: Specification curve analysis for consumption



Notes: Panel A employs the original index (original_index) alongside the winsorized original index (w_original_index); new_index is a slightly different index created by adding up the individual components (per capita monthly food expenditures, per capita monthly non-food expenditures, and per capita monthly non-durable expenditures) separately instead of using the variable in the questionnaire. W_new_index is the winsorized version of the new index.
Panel B displays the specification curve for each component of the original index separately: food_exp captures the per capita monthly food expenditures, nonfood_exp captures the per capita monthly non-food expenditures, and durable_exp the per capita monthly durable expenditures. Baseline_controls includes a full set of baseline characteristics and survey_timing controls for the month of the survey.

Food Security

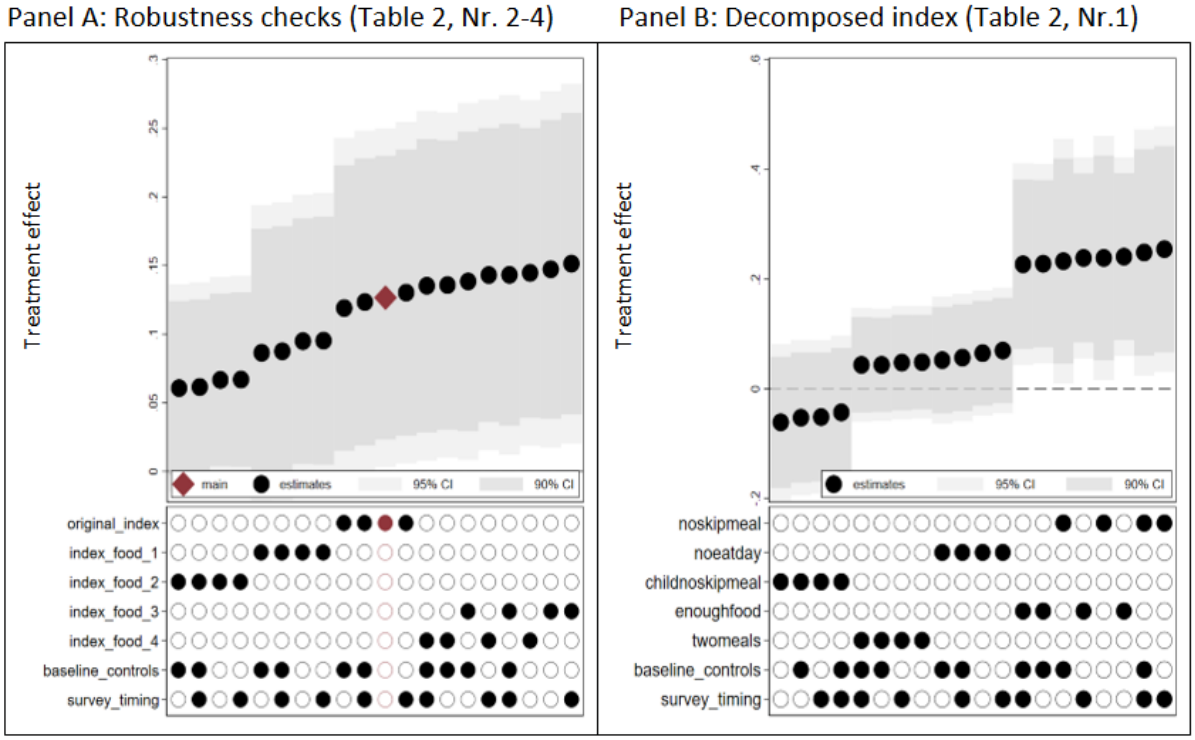
Figure A2 presents the specification curve for food security. We alter the composition of the index in the following ways:

- *index_food_1*: drops variable 1a (adult reduced portion size/skipped meal in past 12 months) from the index, since it is potentially an unprecise question given the 12 months recall period;
- *index_food_2*: drops 1c (children's meals reduced in past 12 months), again, because of the recall period but also since only half the households have children under 16;
- *index_food_3*: drops 1a only 1c jointly;
- *index_food_4*: index is only composed of 1d (all members get enough food every day) and 1e (everyone regularly eats two meals a day) since both questions capture food security at a more abstract level.

In Panel A, we observe some changes in significance levels of the effects when incorporating baseline controls. The results are robust to different definitions of the index, yet index_food_1 and index_food_2 as well as the original index are sensitive to the inclusion of control variables as the effect size reduces considerably. Our preferred index composition index_food_4 yields similar results as BDS in terms of significance level and effect size.

In Panel B, we present the decomposed index, estimating the effect of each component separately. The results demonstrate that noskipmeal and enoughfood have substantial and highly significant effects, driving the food security index's overall impact. Conversely, the remaining components are not statistically significant, with childnoskipmeal even showing negative effects. These findings remain consistent when control variables are included.

Figure A2: Specification curve analysis for food security



Notes: Panel A employs the original index (original_index) alongside with four variations: *index_food_1* drops *noskipmeal*; *index_food_2* drops *noskipmeal* and *childnoskipmeal*; *index_food_3* drops *childnoskipmeal*; *index_food_4* consitsts only of *enoughfood* and *twomeall.* Panel B displays the specification curve for each component of the original index separately: *noskipmeal* captures whether any adult in the household cut the size or skipped a meal in the past 12 months; *noeatday* captures whether any adult did not eat for a whole day in the past 12 months; *childnoskipmeal* captures whether the size of children's meal was cut or skipped in the past 12 months; *enoughfood* captures if all household members got enough food every day; *twomeal* captures whether all household members regularly eat two meals a day. Baseline_controls includes a full set of baseline characteristics and survey_timing controls for the month of the survey.

Income

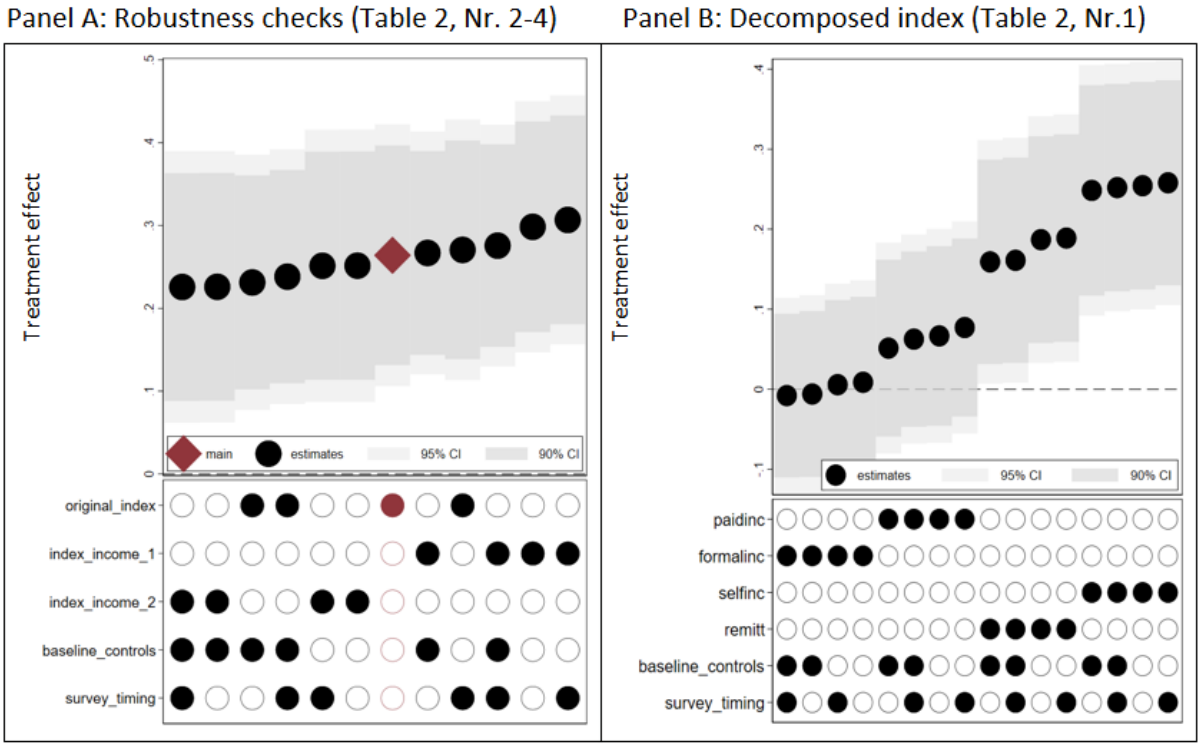
Figure A3 presents the specification curve for income. We alter the composition of the index in the following ways:

- *index_income_1*: includes all reported income sources from the household roster, including begging and remittances. Our income index is therefore generated consistently from the same part of the questionnaire, while BDS do not consider begging for their income index and calculate remittances from a different part of the questionnaire;
- *index_income_2*: replaces the reported income sources from the household roster with more detailed income sections wherever available. For livestock revenue, non-farm income, and remittances, the questionnaire contains a detailed section collecting information on costs and profits. It might be the case that respondents recall income figures more precisely when asked detailed questions about the activity.

Panel A shows that the inclusion of control variables does not alter effect size and significance levels. The alternative income indices affirm the robustness of the original results. The effect sizes remain very similar in size and retain their statistically significant. Overall, our robustness reproduction for income reinforces the findings presented in BDS.

In Panel B, we present the decomposed index, estimating the effect of each component separately. It underlines that income from self-employment and remittances drive the effects, as shown by the significant effect sizes in green on the upper left of the table. BDS conduct a similar analysis and find identical patterns in their Table 3. The finding for the decomposed index is robust to the inclusion of a control variables.

Figure A3: Specification curve analysis for income



Notes: Panel A employs the original index (original_index) alongside with four variations: *index_income_1* includes all reported income sources from the household roster including begging and remittances; *index_income_2* uses more detailed income sections from the questionnaire whenever possible. Panel B displays the specification curve for each component of the original index separately: *paidinc* captures wage income; *fomalinc* captures formal wage income; *selfinc* captures income from self-emplyoment; *remitt* captures remittances received by the household. Baseline_controls includes a full set of baseline characteristics and survey_timing controls for the month of the survey.

Physical health

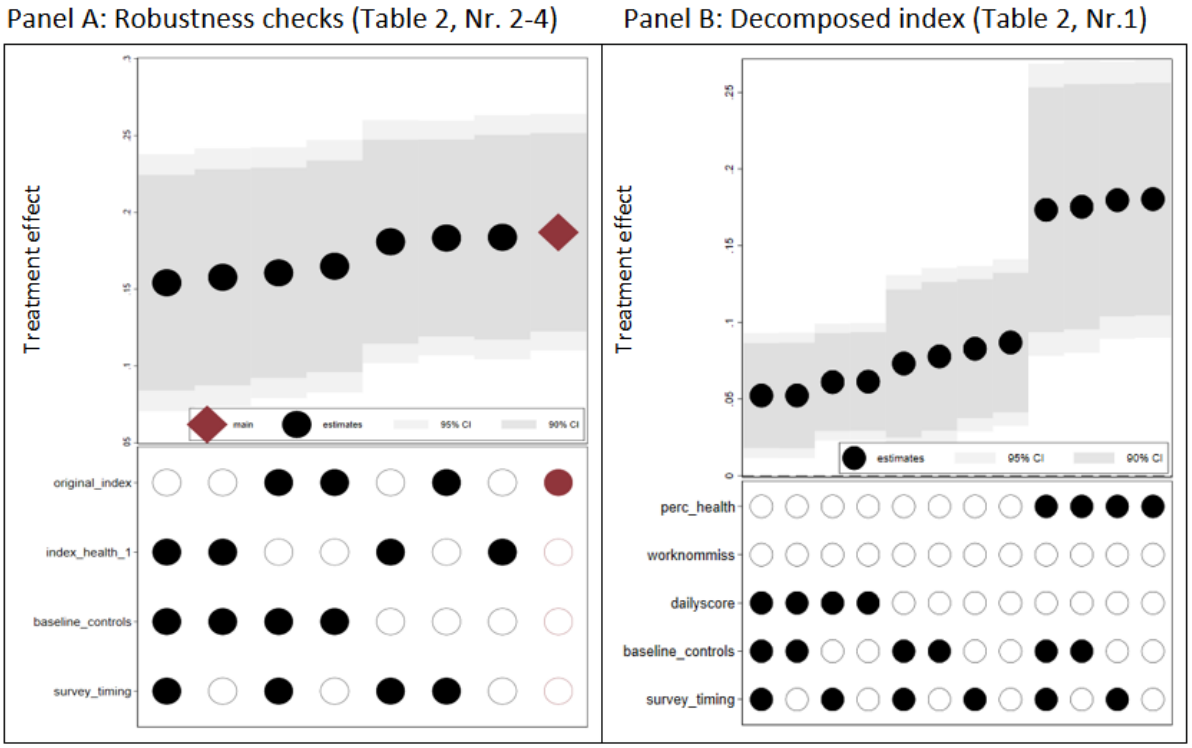
Figure A4 presents the specification curve for health. We alter the composition of the index in the following ways:

- by including a more comprehensive index for the level of difficulty to carry out daily tasks that is not limited to five kilo objects; walk five kilometers; could work a day in the field as the original index. Additionally we include the tasks: getting dressed, eating, doing light work, and squat.

Panel A shows the robustness of the results to the inclusion of control variables. The alternative outcome index performs very similar to BDS's health index and the inclusion of controls does not change this finding. Overall, the results for the health index are very robust.

In Panel B, we present the decomposed index, estimating the effect of each component separately. It demonstrates that all three components are significantly positive while *perc_health* exhibits the largest effect. Adding control variables for timing of the survey, baseline controls, and both jointly leads to an increase in coefficients.

Figure A4: Specification curve analysis for physical health



Notes: Panel A employs the original index (original_index) alongside with four variations: *index_health_1* contains a more detailed index for the ability of conducting daily tasks. Panel B displays the specification curve for each component of the original index separately: *perc_health* is a health self-assessment from 1-19; *worknomiss* captures whether the person was unable to perform daily tasks in the past 30 days; *dailyscore* is an index for the ability of conducting daily tasks. Baseline_controls includes a full set of baseline characteristics and survey_timing controls for the month of the survey.

Treatment refusal

We assess the role of treated individuals who refused treatment. Of the 514 households offered treatment, only 266 accepted – even though the treatment came at no costs and offered gifts and training, with no perceivable monetary or non-monetary costs. BDS state in the appendix that there were rumors about the NGO aiming to convert beneficiaries to Christianity and "some wives were worried that their husband would mishandle the asset". BDS only show Intent-to-Treat (ITT) effects, comparing households assigned to treatment with those in the control group, regardless of actual treatment reception. We explore whether impacts differ in line with expectations when looking at the treatment effect on those who accepted the treatment.

In Table A3, we perform two comparisons: First, we calculate the *Average Treatment Effect on the Treated* (ATT) by comparing households that *received* the treatment with the control group, expecting larger effects than for the ITT analysis (Panel A). Second, we compare households that *refused* the treatment with the control group to see if they exhibit similar effects as the pure control group (Panel B). Although this is a biased comparison, we still find the expected patterns: Effect sizes notably increase when focusing on the treatment group with actual treatment. Conversely, when comparing households from the treatment group that refused treatment to the control group, the effects disappear. Different patterns might have raised validity concerns, but this finding underpinning the BDS results in a striking manner.

Table A3: Robustness analysis for treatment refusal

	(1) Asset Index	(2) Per capita consumption	(3) Food security index	(4) Income and revenues	(5) Financial inclusion index	(6) Physical health index	(7) Mental health index	(8) Productive time use
Panel A: Only treated who accepted treatment								
Treatment	0.832***	1.054***	0.244**	0.596***	0.363	0.210***	0.325***	0.204**
	(0.000)	(0.000)	(0.001)	(0.000)	(0.057)	(0.000)	(0.000)	(0.003)
N	673	669	673	673	673	945	945	945
Panel B: Only treated who refused treatment								
Treatment	-0.269	-0.079	-0.053	-0.142	-0.250	0.140*	0.023	0.065
	(0.147)	(0.183)	(0.091)	(0.086)	(0.203)	(0.056)	(0.060)	(0.071)
N	631	629	631	631	631	851	851	851

Notes: P-values in parentheses. Regressions identical to BDS-Table 1.

References

Allcott, H. (2015). Site selection bias in program evaluation. *The Quarterly Journal of Economics, 130(3)*, 1117-1165.

Angrist, N., & Meager, R. (2023). Implementation matters: Generalizing treatment effects in education. *Available at SSRN 4487496*.

Ankel-Peters, J., Brodeur, A., Dreber, A., Johannesson, M., Neubauer, F., & Rose, J. (2024). *A Protocol for Structured Robustness Reproductions and Replicability Assessments* (No. 143). I4R Discussion Paper Series.

Ankel-Peters, J., Fiala, N. & Neubauer, F., 2023. Is economics self-correcting? Replications in the American Economic Review. *Economic Inquiry*.

Banerjee, A., Duflo, E., Finkelstein, A., Katz, L.F., Olken, B.A. & Sautmann, A., 2020. *In praise of moderation: Suggestions for the scope and use of pre-analysis plans for rcts in economics* (No. w26993). National Bureau of Economic Research.

Banerjee, A., Duflo, E., & Sharma, G. (2021a). Long-term effects of the targeting the ultra poor program. *American Economic Review: Insights*, *3*(4), 471-86.

Banerjee, A., Duflo, E., & Sharma, G. (2021b). Data and Code for: Long-term effects of the Targeting the Ultra Poor Program. Nashville, TN: American Economic Association, 2021. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2021-11-17. https://doi.org/10.3886/E130362V1

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., ... & Udry, C. (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, *348*(6236), 12607.

Barker, N., Karlan, D., Udry, C., & Wright, K. (2024). The Fading Treatment Effects of a Multifaceted Asset-Transfer Program in Ethiopia. *American Economic Review: Insights*, *6*(2), 277-294.

Basu, K. (2014). Randomisation, causality and the role of reasoned intuition. *Oxford Development Studies*, *42*(4), pp.455-472.

Bauchet, J., Morduch, J., & Ravi, S. (2015). Failure vs. displacement: Why an innovative anti-poverty program showed no net impact in South India. *Journal of Development Economics*, *116*, 1-16.

Bensch, G. (2024). *Repframe*. GitHub. https://github.com/guntherbensch/repframe.

BITSS, Berkeley Initiative for Transparency in the Social Sciences. (2020). *"Guide for Advancing Computational Reproducibility in the Social Sciences."* [Accessed: May 29,2024]. https://bitss.github.io/ACRE/.

Bold, T., Kimenyi, M., Mwabu, G., & Sandefur, J. (2018). Experimental evidence on scaling up education reforms in Kenya. *Journal of Public Economics*, *168*, 1-20.

Bouguen, A., Huang, Y., Kremer, M., & Miguel, E. (2019). Using randomized controlled trials to estimate long-run impacts in development economics. *Annual Review of Economics*, *11*, 523-561.

Brodeur, A., Carrell, S., Figlio, D. & Lusher, L., 2023. Unpacking p-hacking and publication bias. *American Economic Review*, *113*(11), pp.2974-3002.

Brodeur A., Cook N. M., Hartley J. S. & Heyes A. (2024b) Do Pre-Registration and Pre-Analysis Plans Reduce p-Hacking and Publication Bias? Evidence from 15,992 Test Statistics and Suggestions for Improvement. *Journal of Political Economy: Microeconomics,* forthcoming.

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

Brodeur, A., Mikola, D., Cook, N., Brailey, T., Briggs, R., de Gendre, A., ... & Havránek, T. (2024a). *Mass Reproducibility and Replicability: A New Hope* (No. 107). The Institute for Replication (I4R).

Cartwright, N. (2011). A philosopher's view of the long road from RCTs to effectiveness. *The Lancet*, 377(9775), 1400–1401. https://doi.org/10.1016/S0140-6736(11)60563-

Christensen, G. & Miguel, E., 2018. Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, *56*(3), pp.920-980.

Duflo, E. (2020). Field Experiments and the Practice of Policy. *American Economic Review*, 110(7), 1952–1973. https://doi.org/10.1257/aer.110.7.1952

Duflo, E., Glennerster, R. & Kremer, M., 2007. Using randomization in development economics research: A toolkit. *Handbook of Development Economics*, *4*, pp.3895-3962.

Esterling, K. M., Brady, D., & Schwitzgebel, E. (2023). *The necessity of construct and external validity for generalized causal claims* (No. 18). I4R Discussion Paper Series.

Ferguson, J., Littman, R., Christensen, G., Paluck, E. L., Swanson, N., Wang, Z., ... & Pezzuto, J. H. (2023). Survey of open science practices and attitudes in the social sciences. *Nature Communications*, *14*(1), 5401.

Gechter, M., 2024. Generalizing the Results from Social Experiments: Theory and Evidence from India. *Journal of Business & Economic Statistics*, *42*(2), pp.801-811.

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., ... & Stopnitzky, Y. (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, *59*(3), 944-960.

Masselus, L., Petrik, C., & Ankel-Peters, J. (2024). *Lost in the Design Space? Construct Validity in the Microfinance Literature*. OSF Preprint. https://doi.org/10.31219/osf.io/nwp8k

McKenzie, D., Woodruff, C., Bjorvatn, K., Bruhn, M., Cai, J., Gonzalez-Uribe, J., ... & Valdivia, M. (2021). Training entrepreneurs. *VoxDevLit*, *1*(2), 3.

Mo, D., Bai, Y., Shi, Y., Abbey, C., Zhang, L., Rozelle, S., & Loyalka, P. (2020). Institutions, implementation, and program effectiveness: Evidence from a randomized evaluation of computer-assisted learning in rural China. *Journal of Development Economics*, *146*, 102487.

Miguel, E. (2021). Evidence on research transparency in economics. *Journal of Economic Perspectives*, *35*(3), 193-214.

Ofosu G. K. & Posner D. N. (2023) 'Pre-analysis Plans: An Early Stocktaking', *Perspectives on Politics*, *21/1*: 174–90.

Peters, J., Langbein, J., & Roberts, G. (2018). Generalization in the tropics–development policy, randomized controlled trials, and external validity. *The World Bank Research Observer*, *33*(1), 34-64.

Pritchett, L. (2020). Randomizing Development: Method or Madness?". Chap. 2 In Randomized Control Trials in the Field of Development a Critical Perspective, edited by Florent Bédécarrats, Isabelle Guérin and Francois Roubaud, 79-109.

Pritchett, L., Samji, S. & Hammer, J.S. (2013). It's all about MeE: *Using Structured Experiential Learning ('e') to crawl the design space*. Center for Global Development Working Paper, (322).

Pritchett, L. & Sandefur, J. (2015). Learning from experiments when context matters. *American Economic Review*, *105*(5), pp.471-475.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208-1214.

Usmani, F., Jeuland, M., & Pattanayak, S. K. (2022). NGOs and the effectiveness of interventions. *Review of Economics and Statistics*, 1-45.

Vilhuber, L. (2020). Reproducibility and replicability in economics. *Harvard Data Science Review*, *2*(4), 1-39.

Vivalt, E., 2020. How much can we generalize from impact evaluations? *Journal of the European Economic Association*, *18*(6), pp.3045-3089.

Vivalt, E. & Coville, A., 2023. How do policymakers update their beliefs? *Journal of Development Economics*, *165*, p.103121.

White, H., 2009. Theory-based impact evaluation: principles and practice. *Journal of Development Effectiveness*, *1*(3), pp.271-284.