Estimating Impact with Surveys versus Digital Traces: Evidence from Randomized Cash Transfers in Togo[*]

Emily Aiken Suzanne Bellue Joshua E. Blumenstock
Dean Karlan Christopher Udry

November 12, 2024

Abstract

We study whether program impacts can be estimated using a combination of digital trace data and machine learning. In a randomized controlled trial of cash transfers in Togo, endline survey data indicate positive treatment effects on food security, mental health, and perceived economic status. However, estimates of impact based solely on predicted endline outcomes (generated using trace data and machine learning, which do successfully predict baseline poverty) are smaller and noisier, and generally not statistically significant. When post-treatment outcome data are used in conjunction with predictions to estimate treatment effects, predicted impacts are similar to those estimated using surveys.

JEL Codes: C55, I32, I38

[*] Aiken: emilyaiken@berkeley.edu, University of California Berkeley; Bellue: suzanne.bellue@ensae.fr, CREST - ENSAE; Blumenstock: jblumenstock@berkeley.edu, University of California Berkeley; Karlan: karlan@kellogg.northwestern.edu, Northwestern University; Udry: christopher.udry@northwestern.edu, Northwestern University. Shikhar Mehra and Nathaniel Ver Steeg provided excellent research assistance and Isabel Oñate provided excellent research management. This project was possible through the dedication of our project partners in Togo, especially C. Lawson, S. Bakari, L. Mills, K. Ekouhoho, M. Koudeka and A. Byll. We are grateful for support from Institut National de la Statistique et des Études Economiques et Démographiques (INSEED), led by S. Telou and R. Ogoumedi; as well as the team at GiveDirectly, especially H. S. Chia, M. Cooke, K. Lee, A. Nawar and D. Quinn. Simón Ramirez Amaya and Ethan Ligon provided helpful feedback. We thank the World Bank for supporting the phone surveys and data collection under the WURI program. Aiken gratefully acknowledges financial support from a Microsoft Research PhD Fellowship. Bellue gratefully acknowledges financial support from the German Academic Exchange Service (DAAD) and the German Research Foundation (through the CRC-TR-224 project A03). Blumenstock thanks the National Science Foundation for support under CAREER Grant IIS-1942702. This trial was pre-registered with the American Economic Association Registry #7590. IRB approval was obtained from the U.C. Berkeley Committee for the Protection of Human Subjects (Protocol #2020-05-13281).

1 Introduction

Reliable estimates of post-program outcomes are essential to impact evaluation. In low- and middle-income countries (LMICs), such outcomes are traditionally measured through surveys. However, a new paradigm is emerging for estimating living standards based on the application of machine learning algorithms to nontraditional data from mobile phones (e.g. Blumenstock et al., 2015; Blumenstock, 2018), satellites (e.g. Jean et al., 2016; Yeh et al., 2020), and other digital sources (e.g. Fatehkia et al., 2020; Sheehan et al., 2019). These data and methods can generate estimates for large populations more rapidly and cheaply compared to traditional surveys.

We ask whether welfare outcomes estimated from "digital trace" data produce the same estimates of program impact as those obtained from traditional survey-based measures of welfare. If possible, this could open up new opportunities for low-cost program monitoring and impact evaluation. We study these questions in the context of Togo's *Novissi* program, which provided five monthly cash transfers of USD $13-15 to poor individuals in rural Togo during the COVID-19 pandemic. The program was rolled out starting in late 2020 as an individual-level randomized controlled trial (RCT). We compare two sources of data for estimating the treatment effects of the Novissi program: a large phone survey we conducted shortly after cash transfers were provided to the treatment group, and the complete mobile phone transaction logs of all consenting program participants.

First we use the phone surveys to estimate welfare impacts.[1] Transfers increased food security by 0.06 standard deviations (SD) with a standard error (se) of 0.02, mental health by 0.07 SD (se=0.02), and perceived economic status by 0.04 SD (se=0.02). Effects on other outcomes were positive but not statistically significant. The effect on a composite index of welfare was 0.06 SD (se=0.02).

Second, we develop intuition for how mobile phone data might be used to estimate treatment effects. Analyzing millions of phone transaction records, we find that households that received cash transfers used their phones differently than those that did not — for instance, beneficiaries made more calls to more distinct people. Across 824 distinct "features" of mobile phone use, cash transfers statistically significantly impacted 35% of features ($p < 0.05$). We also establish the extent to which machine learning algorithms applied to mobile phone data can accurately predict survey-based measures of welfare. Here, we find that machine learning (ML) algorithms can produce relatively accurate estimates of a proxy means test ($R^2 = 0.05$-0.14, which is comparable to prior work), but do not accurately predict other

[1] Analysis adheres to a pre-registered analysis plan, AEA Registry #7590.

more focused welfare outcomes such as food security, mental health, and perceived economic status ($R^2 = 0.00$ - 0.05).

Last, we test whether *predicted* welfare outcomes — generated by applying ML to mobile phone data — can be used to estimate the welfare effects of the Novissi program. When welfare impacts are based exclusively on out-of-sample predictions of endline outcomes (either because the ML model was trained exclusively on pre-treatment surveys, or because the estimator is not exposed to the endline data that were used to train the ML model), we do not observe statistically significant effects on three of the four outcomes that were statistically significant when treatment effects were directly estimated from surveys. When we use a framework for prediction-powered inference (PPI, Angelopoulos et al., 2023), which combines both survey data and ML predictions, the PPI-estimated impacts are similar in magnitude and precision to those using only endine surveys.

Subsequent analysis suggests that impact estimates based on ML predictions differed from those based on surveys because of the difficulty of predicting non-economic outcomes such as food security and perceived economic status from mobile phone data. While we cannot conclusively diagnose the source of this difficulty, we find that food security and other vulnerability measures in Togo are less geographically concentrated than poverty, which likely makes them more difficult to predict from mobile phone data. Finally, we show that — even if the predictive models had been more accurate — impact evaluation using phone data could still be complicated by issues of model drift, and by the difficulty of inferring impacts that were modest in magnitude. We conclude with a discussion of how these results can inform the broader conversation around the use of digital data for monitoring and impact evaluation.

Our research contributes to two main literatures. The first documents the impacts of unconditional cash transfers on a range of welfare outcomes, including expenditures, food security, health, education, savings, and financial inclusion (for reviews, see Bastagli et al. (2016) and Crosta et al. (2024)). Several more recent papers document the welfare impacts of cash transfers distributed in response to the COVID-19 pandemic (Banerjee et al., 2020; Londoño-Vélez and Querubin, 2022; Karlan et al., 2022; Bottan et al., 2021)[2]. Broadly, this literature shows modest, positive, and statistically significant impacts of cash transfers on food security and mental health. The first portion of our analysis contributes to this literature by documenting the impacts of pandemic cash transfers in Togo, using an extensive cash transfer program where treatment was randomly assigned at the individual level. While the

[2] see Karlan et al. (2022) for a summary of cash transfer RCTs during COVID.

cash transfers we study are smaller ($13-15.50 per month) than most of the other programs studied ($15-50 per month), we document comparable effect sizes (0.04-0.07 SD).

The second, more nascent literature explores the use of digital data sources for measuring welfare and evaluating programs and policies. While early work in this space focused on documenting the potential for measuring welfare from mobile phone (Blumenstock et al., 2015; Blumenstock, 2018; Aiken et al., 2022a) and satellite data (Jean et al., 2016; Yeh et al., 2020; Chi et al., 2022), more recent work has begun to ask whether program impacts can be estimated using these data sources. In particular, two recent papers find that estimates of the impact of large-scale development interventions, estimated using satellite imagery, are similar to but noisier than estimates based on surveys (Huang et al., 2021; Ratledge et al., 2022). We build on that work by examining the predictive power of mobile phone data for measuring the short-term impacts of much smaller scale cash transfer program.[3]

2 The GiveDirectly-Novissi Program

2.1 Program Design and RCT

The GiveDirectly-Novissi program (GD-Novissi), implemented jointly by the Togolese Ministry of Digital Transformation and GiveDirectly, provided monthly cash transfers to 138,589 individuals in rural parts of Togo between November 2020 and August 2021. Eligible women received 8,620 FCFA (USD $15.50 = $38 PPP) per month, and eligible men received 7,450 FCFA (USD $13 = $33 PPP) per month for five months.

Eligibility for GD-Novissi was based on geographic and poverty-related criteria. First, beneficiaries had to be registered to vote in one of the 100 poorest cantons in the country. Second, poverty estimates for each registered subscriber were derived from their pre-program mobile phone records; only subscribers estimated to be living on less than $1.25/day (the poorest 29% of subscribers) were eligible. Aiken et al. (2022b) provides a full description of GD-Novissi's targeting approach and discusses the extent to which these eligibility criteria created systematic exclusions from the program. While the primary focus of Aiken et al. (2022b) is to understand if phone data could be used for targeting the Novissi program, the primary focus of the present work is to understand if and how phone data can be used for impact evaluation.

[3] Barriga-Cabanillas et al. (2021) uses a regression discontinuity design to study the impact of a cash transfer program in Haiti. Preliminary results, consistent with our own, suggest that phone-data-based estimates of food security are too noisy to detect the impact of cash transfers.

The program launched in November 2020; after three months, 181,028 individuals had registered, of whom 49,083 were eligible. Prior to registration, eligible individuals were randomly assigned to treatment (N=27,673) and control (N=21,410) groups. Upon registration, subscribers in the treatment group immediately received the first of their five monthly cash transfers.[4] Subscribers in the control group were not told they would later receive transfers (N=21,410).[5]

2.2 Endline Survey

To evaluate the impact of GD-Novissi cash transfers, we conducted an "endline" phone survey with both treatment and control individuals in May 2021, between zero and two months after members of the treatment group had received their final cash transfer (and before any of the control group subscribers had received a transfer — see Figure S1 for the project timeline). Our survey included modules on food security, financial health, financial inclusion, mental health, perceived socioeconomic status, labor supply, health care access, and labor supply, along with a proxy-means test (PMT).[6]

The sample frame for the endline survey was subscribers who enrolled in GD-Novissi RCT in the earlier portion of the program between November 2020 and January 2021 (N=49,083). We then randomly sampled 49% from our sample frame to survey, stratified by treatment status and geography. We stratified geographically to account for the fact that the government had distributed a one-time cash transfer in the Savanes region, where 70% of the GD-Novissi participants reside (see Appendix C.2). In total, we completed surveys with 9,511 individuals, with a survey response rate of 39%; there was no differential attrition between treatment and control groups (Table S1). Appendix A provides more details on the endline survey, and Table S2 provides summary statistics and balance checks for the impact evaluation sample.[7] We observe small and generally not statistically significant differences in treatment assignment by gender, age, occupation, and place of residence.

[4] During our evaluation, in February 2021, the government launched a separate cash transfer program that targeted people in the Savanes region, following a travel restriction intended to contain COVID-19. This program provided a one-time cash transfer of USD $8-10 to residents of Savanes. We discuss the implications of this second program to our main evaluation in Appendix C.2.

[5] Subscribers in the control group received the same total transfer amount, but the transfer was not pre-announced and was delivered after all surveys were completed.

[6] Table S3 reports the components of each outcome index. Our analysis deviates from the registry in that we merge 'Adoption and use of mobile money services' with 'Financial inclusion,' since the financial inclusion index is defined using the fraction of bank accounts and mobile money usage in households.

[7] By design, only a few observations overlap between the baseline and endline samples, so we conduct the balance checks on time-invariant variables of the endline survey.

2.3 Pre-treatment Survey

While our impact evaluation with survey data relies primarily on the endline survey conducted post-treatment, portions of our analysis use a pre-treatment phone survey conducted in September 2020, prior to the roll-out of the GD-Novissi program. Phone survey respondents were drawn from among active mobile subscribers whose primary home location was in those 100 poorest cantons, using geographic information available in the mobile phone data (see Appendix B). In total, we completed 9,484 pre-treatment surveys.

The primary objective of the pre-treatment survey was to collect PMT data to then train the machine learning algorithms used to identify eligible GD-Novissi beneficiaries (Aiken et al., 2022b). As such, it differed from the endline survey in two key respects. First, it was more focused on the PMT; omitted a mental health module; and, had fewer food security questions (Table S4). Second, the population was designed to be representative of *all* active mobile phone subscribers in Togo's 100 poorest cantons, not just those subscribers predicted to be below the poverty threshold. As shown in the first two columns of Table S2, the pre-treatment sample was still quite poor (average estimated daily per capita consumption of \$1.49, SD = \$0.74), but less homogeneously poor than in the endline survey (average consumption \$1.31, SD = \$0.49). Additional details on the pre-treatment survey are provided in Appendix B.

2.4 Mobile Phone Metadata

We obtained comprehensive mobile phone metadata from Togo's two mobile network operators for the duration of the GD-Novissi program. These data include detailed metadata about each phone call and text message sent or received on the mobile networks, including the phone number of the caller and recipient, the timestamp, the duration of calls, and the cell tower through which the call was placed. The data also include mobile data usage, including the phone number of the subscriber, the timestamp, and the amount of mobile data used for each mobile data transaction.[8]

We obtained informed consent from each respondent in the pre-treatment and endline surveys to match their survey responses to their mobile phone records.[9] We then generated

[8] Although the dataset shared by the mobile network operators also includes records of mobile money use, we do not use mobile money transactions in our main analysis since the treatment itself was delivered via mobile money and thus mechanically (and dramatically) changed mobile money usage patterns for the treatment group. However, we explore the inclusion of mobile money data in Section 5.1.

[9] Following the data protection procedures described in our IRB protocol, we pseudonymized or removed all personally identifying information, including phone numbers, prior to linking these two datasets.

sets of mobile phone *features* describing how each survey respondent used their mobile phone in the period preceding the survey. Features were generated using open source library `cider`[10] following the procedure described in Aiken et al. (2022b). Appendix E lists all 824 features calculated from mobile phone data, relating to calling patterns, contact networks, mobility, location, data usage, international transactions, and more. We generated features for two time periods: one corresponding to the six months preceding the pre-treatment survey (April - September 2020); and one for the six months preceding the endline survey (November 2020 - April 2021).

3 Program Impacts Estimated With Survey Data

Our first set of results uses the endline survey to estimate the causal impact of GD-Novissi. These results are based on weighted regressions of each of the seven outcomes on treatment status and include strata, enumerator, and week of the survey fixed effects. To account for multiple hypotheses, we include p-values adjusted for the False Discovery Rate (Anderson, 2008) for our seven pre-specified outcome indices.

Results in Table 1 Panel A indicate that GD-Novissi increased food security (by 0.06 SD, $p = 0.003$), mental health (by 0.07 SD, $p < 0.001$), and self-perceived socioeconomic status (by 0.04 SD, $p = 0.074$). These results are broadly consistent with studies of the effects of cash transfers during the COVID-19 pandemic in other contexts (Banerjee et al., 2020; Londoño-Vélez and Querubin, 2022; Karlan et al., 2022; Bottan et al., 2021).[11] GD-Novissi does not decrease individual labor supply (the coefficient is positive but not statistically significant, with a point estimate close to zero), consistent with evidence on the effect of cash transfers in other contexts (Banerjee et al., 2017, 2022; Crosta et al., 2024). We observe no statistically significant effects on our indices of financial health, financial inclusion, or healthcare access, although the coefficient estimates are positive.[12] The last column of Table 1 Panel A indicates that GD-Novissi increases an aggregate welfare index by 0.06 standard deviations ($p = 0.008$), where the aggregate index is constructed as an aggregated normalized index of the seven underlying outcome indices.[13]

[10] https://global-policy-lab.github.io/cider-documentation/

[11] Appendix D compares our survey-based results to impact evaluations of cash transfers in other settings during the COVID-19 pandemic.

[12] Our financial inclusion index measures the fraction of bank accounts and mobile money usage in households, excluding mobile money accounts of the respondents.

[13] In robustness tests, we test whether results change if we exclude the 17% of households with more than one GD-Novissi beneficiary. We find results are almost identical: food security increases by 0.06 SD, mental health by 0.07 SD, perceived economic status by 0.04 SD, and the composite welfare index by 0.06 SD.

In Appendix C, we test for treatment effect heterogeneity on four dimensions: gender, poverty, occupation, and region of residence (in or outside Togo's northernmost region, Savanes). Treatment effects are not heterogeneous across any of these dimensions except for geography: treatment effects on food security and mental health were statistically significantly larger for beneficiaries in the Savanes region in the far North of Togo — see Appendix C.2 for a discussion of this geographic heterogeneity.

4 Program Impacts Estimated With Phone Data

Our main analysis explores whether the welfare impacts observed in survey data can be estimated from mobile phone data alone. We test whether treatment effects on *predicted outcomes*, generated by applying machine learning models to the mobile phone data, are the same as those observed in the endline survey. If successful, such an approach could enable new paradigms for impact evaluation, since digital trace data can be obtained at much lower cost than traditional surveys, and from populations that might be difficult or impossible to reach with surveys.

4.1 Preliminaries: Intuition For Our Approach

Prior to estimating treatment effects on measures of welfare *predicted* from mobile phone data, we develop two important intuitions. First, we show that phone use changed significantly in response to the cash transfer program. Second, we assess the accuracy with which survey outcomes can be predicted from phone data.

4.1.1 Treatment Effects on Mobile Phone Use

The GD-Novissi program affected how people used their phones. Across the 824 different metrics of phone use (enumerated in Appendix E and calculated from phone transactions between November 2020 and April 2021, during the treatment group's transfers), there are statistically significant differences ($p < 0.05$) between treatment and control for 35% of metrics. Table 2 provides the standardized impacts of cash transfers on several easily interpretable dimensions of phone use: the cash transfer treatment increases calls by 0.02 SD (se = 0.009), contacts by 0.03 SD (se = 0.009), active days by 0.06 SD (se = 0.009), and unique prefectures (admin-2 units) visited by 0.04 SD (se = 0.009).

The cash transfers most consistently impact dimensions of phone use relating to calling: 55% of the 313 features related to calling patterns (such as number and duration of calls and diversity of call contacts) differ significantly between the treatment and control groups at the 0.05 level. Only 11% of text-related features are significantly different, 17% of mobile data usage-related features, and 38% of features related to mobility and location. No features relating to international transactions differ statistically significantly between treatment and control groups. The largest (in magnitude) treatment effects are on the share of contacts that make up 80% of an individual's calling time (-0.07 SD, se = 0.009), the share of contacts that make up 80% of an individual's calls (-0.07 SD, se = 0.009), and the number of unique days an individual uses their mobile phone (0.06 SD, se = 0.009).

4.1.2 Predicting Welfare Levels from Phone Data

Table 3 reports the accuracy with which outcomes from a single wave of survey data can be predicted from mobile phone data. This analysis closely follows the methodology developed in Blumenstock et al. (2015) and adapted to Togo in Aiken et al. (2022b) – see Appendix E for details. In Panel A, which shows results for predicting pre-treatment survey outcomes (using six months of pre-treatment phone data), predictive accuracy is highest for the pre-treatment PMT ($R^2 = 0.143$, or a pearson correlation coefficient of $r = 0.381$) — a finding that replicates results previously documented in Aiken et al. (2022b). However, we are unable to accurately predict any of the other welfare indices (food security, financial health, perceived status, and labor supply) from mobile phone features ($R^2 = 0.002$ - 0.046).[14] In Panel B, which shows results for predicting endline survey outcomes (using six months of post-treatment phone data), predictive accuracy for the PMT is lower ($R^2 = 0.049$, or a pearson correlation coefficient of 0.253). This difference is likely due in part to the more homogeneous population represented in the endline survey: while both surveys focus on the same rural areas, the endline survey was also restricted to *program-eligible* mobile subscribers, where eligibility was determined by predicted poverty (Section 2.2). Other endline outcomes are not accurately predicted from phone data ($R^2 = 0.002$ - 0.026).

4.2 Estimating Treatment Effects from Phone Data

We now test whether welfare *impacts* can be estimated from mobile phone data. We test this approach under two different regimes. In the first regime, we assume that the only opportunity for survey data collection occurs before treatment is administered. In the second regime,

[14] We cannot include results on financial inclusion, mental health, healthcare access, nor the seven-index composite from Table 1 Panel A, as the questions required to construct these indices were not included in the pre-treatment survey.

we instead assume that the only opportunity to collect survey data is after administering treatment. Results under both regimes are presented below.

4.2.1 Regime 1: Only Pre-Treatment Surveys Are Available

In the first regime, pre-treatment survey data (collected in September 2020) are matched to pre-treatment phone data (March - September 2020), and a machine learning algorithm is trained to to predict outcomes from phone data (as was shown in Table 3, Panel A). Then, after the program has been implemented, we conduct an impact evaluation by comparing *predicted endline outcomes* of treated and control individuals, where predicted endline outcomes are generated by passing post-treatment mobile phone data through the prediction model that was trained pre-treatment.[15]

Panel B of Table 1 shows the predicted average treatment effect of GD-Novissi, which is obtained by regressing the predicted outcome on the individual's treatment status. To estimate the variance of treatment effects derived from these phone-data-based predictions of welfare, we use a Bayesian bootstrap procedure (Rubin, 1981) to incorporate both the first-stage uncertainty in our ML models' predictions and the variation in predictions across treatment and control subscribers (see Appendix E). The treatment effect is not statistically significant for any of the welfare indices. For food security and perceived socioeconomic status — which both had significant positive treatment effects in the survey — the point estimates of phone-data-based treatment effects are -0.003 and -0.013, respectively.

4.2.2 Regime 2: Post-Treatment Surveys Are Available

In the second regime, we instead assume that the only opportunity to collect survey data is after administering treatment. In this approach, the machine learning algorithm is trained on post-treatment data (i.e., endline survey data from May 2021 that are matched to post-treatment phone data from November 2020 - April 2021), for a small sample of the actual beneficiary population. The trained model is then used to predict outcomes for all subscribers enrolled in the RCT, including those not surveyed.

Panels C and D of Table 1 present results from this second regime where only endline survey data are available to train the model. In Panel C, the models are trained using endline data (as evaluated in Panel B of Table 3); the models are then used to generate

[15] Specifically, one model corresponding to each welfare outcome is trained using the pre-treatment survey and phone data (from March - September 2020), with hyperparameters tuned through 5-fold cross-validation specific to that outcome. Each model is then used to generate predicted welfare outcomes for treated and control individuals, using phone data from the treatment period (November 2020 - April 2021).

predicted welfare outcomes for all subscribers using post-treatment mobile phone data.[16] We estimate the predicted average treatment effects as before by comparing predictions between the treatment and control groups, and estimates of variance are again produced with a Bayesian bootstrap. We do not estimate statistically significant treatment effects on food security, financial inclusion, perceived socioeconomic status, or the combined index; phone-data-based point estimates for these treatment effects range from 0.002 to 0.031 standard deviations. We estimate a positive and statistically significant treatment effect for the mental health index (0.030 SD, $p = 0.055$), which is smaller in magnitude than the corresponding survey-based treatment effect (0.072, $p < 0.01$). There is also a positive and statistically significant phone-data-based treatment effect on the index of healthcare access (0.031 SD, $p = 0.021$); the corresponding survey-based treatment effect is not significant. Thus, while we generally cannot reject equality between phone-data-based and survey-based treatment effects (the p-values associated with these Z-tests are reported in the last row of Panel C in Table 3), the magnitude and significance of the phone-data-based treatment effects rarely lines up with that of the survey-based treatment effects.

An alternative approach when endline survey data and mobile phone data are both available is to use both data sources for treatment effect estimation. The prediction-powered inference (PPI) method introduced by Angelopoulos et al. (2023) provides a framework for constructing statistically valid confidence intervals using both ground-truth observations (in our setting, endline survey data) and contemporaneous machine-learned predictions (in our setting, phone-data-based welfare estimates). In our setting, the PPI-estimated impacts (shown in Panel D of Table 1) are nearly identical in magnitude to those based on endline surveys alone (Table 1 Panel A). However, we do not see the same large improvements in precision observed by Angelopoulos et al. (2023).

4.3 Additional Tests of Robustness

In Section 5, we examine several reasons why the positive treatment effects estimated using survey data were, in general, not observed in treatment effects estimated exclusively with phone-data-based predictions. First, however, we present a few tests to ensure that the preceding results are robust to different variations of the machine learning methodology used to generate predicted treatment effects.

First, results are unchanged if we vary the duration of phone records used to generate

[16] These subscribers include the surveyed subscribers on which the model is trained; results are unchanged if we restrict the inference set to subscribers not surveyed (and thus not used in training).

welfare predictions. This experiment addresses the possibility that phone use may be most impacted by cash immediately following the transfer. For this test, we train and evaluate prediction models that use only two weeks of phone data (instead of the six months used in our main analysis). When matching to the pre-treatment survey, we use phone data from the two weeks during which the survey was conducted (September 17-30, 2020); for the post-treatment period, we use data from the two weeks immediately following the date on which each individual registered for GD-Novissi and received their first transfer. In Table S6, we do not observe improved predictive performance relative to Table 3 (R^2 = -0.002 - 0.112), and treatment effects are similar in magnitude and significance.

Second, we test whether using *changes* in mobile phone use between the pre-treatment period and the post-treatment period can improve predictive performance. Table S7 shows that using changes does not improve predictive performance for our machine learning models (R^2 = -0.004-0.036), and treatment effects are still not statistically significant.

Third, we test whether impact estimates based on mobile phone data are significant on subsets of the population where the survey-based treatment effects were largest. In particular, as discussed in Appendix C (Table S8 Panel C), the survey data indicate that treatment effects are larger for beneficiaries in the Savanes region in the north of Togo. However, when the machine learning model is trained using data from survey respondents in Savanes, and evaluated only in Savanes, predictive power remains low (Table S8 Panels A and B: R^2 = -0.012 - 0.120), and the treatment effects estimated from the phone data are still not statistically significant (Table S8 Panels D and E).

Finally, we run several tests of the machine learning models themselves. In addition to tuning the hyperparameters as described above, we take additional steps to ensure that data are not too sparse for the models being used (Bellman and Kalaba, 1959). Specifically, we introduce a feature selection step prior to model fitting, which eliminates all features that are not statistically significantly different between the treatment and control groups. Despite the substantial share of features that differ systematically between treatment and control subscribers (Section 4.1.1), this feature selection step does not improve predictive accuracy (R^2 = 0.000 - 0.139, full results available on request) or impact the significance of treatment effects.

5 Discussion

To summarize our main results, we find that (i) GD-Novissi cash transfers had positive and statistically significant impacts on food security, financial inclusion, mental health, perceived socioeconomic status, and an aggregate outcomes index in the endline survey; (ii) GD-Novissi transfers significantly impacted many dimensions of mobile phone use, particularly around calling patterns and volume; (iii) while a baseline proxy means test can be predicted reasonably well with mobile phone data and ML, other outcomes are not predicted accurately; and, likely as a result, (iv) estimates of the welfare impact of GD-Novissi are mostly not statistically significant when estimated using phone-data-based predictions of outcomes.

The first result is broadly consistent with several studies finding positive impacts of cash transfers on food security and mental health during the COVID-19 pandemic (Banerjee et al., 2020; Bottan et al., 2021; Londoño-Vélez and Querubin, 2022; Karlan et al., 2022). In comparison to other papers on COVID-19 cash transfers, the GD-Novissi transfer size is slightly smaller (monthly transfers USD 13-15.5 compared to USD 15-52 in other studies). However, effect sizes are of a similar magnitude to those observed in other studies.

The subsequent results are more nuanced and inform a rapidly evolving debate about if and how new digital data sources can be used to inform development research and policy. Where several recent studies have shown that phone data and machine learning can produce accurate estimates of consumption and asset-based wealth, we find that — at least in the rural Togolese context — a similar procedure does not produce reliable estimates of food security or self-perceived economic status.

5.1 Challenges to estimating non-economic outcomes from mobile phone data

Why can phone data and machine learning be used to accurately predict a PMT-based measure of wealth, but not food security or the other self-reported welfare outcomes? We explore four main hypotheses.

5.1.1 The ground truth measurements are noisier

Survey-based measures of food security and other vulnerability outcomes are noisier than survey-based measures of economic poverty (Hjelm et al., 2016; Tadesse et al., 2020), and prediction models trained on noisy outcomes generally perform poorly. While this is likely the case in our setting, it cannot by itself explain the absence of predicted treatment effects, since we observe statistically significant treatment effects when using the ground truth measurements of non-economic outcomes (Table 1 Panel A). To push this intuition further, we test whether the machine learning models can accurately predict an outcome that is measured very accurately: *treatment status*. The results in Table S9 indicate that mobile phone data cannot accurately predict GD-Novissi treatment status (AUC = 0.515 - 0.522), suggesting that measurement error in the survey is not the main reason that the ML-based estimates are not statistically significant.[17]

5.1.2 The difficulty of predicting outcomes of homogeneous populations

A second possible explanation for the low predictive power of non-PMT outcomes is that the study sample is so homogeneous. Whereas most prior work on predicting poverty from phone data has used nationally representative populations (Aiken et al., 2022b; Blumenstock et al., 2015; Blumenstock, 2018), our study focuses on a homogeneous subset of individuals identified to be living in poverty within Togo's poorest 100 cantons. In past work that has compared the accuracy of predicting poverty from mobile phone data in full-country evaluations vs. in rural areas only, predictive power is typically substantially lower when restricting to rural areas ($r^2 \approx 0.21$ vs. 0.10 in Togo for poverty prediction at the individual level (Aiken et al., 2022b) and $r^2 \approx 0.41$ vs. 0.25 in Rwanda for poverty prediction at the district level (Blumenstock et al., 2015)). However, while the homogeneity of the population in our study helps explain why predictive power for the PMT is lower than in previous papers that evaluate nationally representative samples, it does not explain why predictive accuracy for non-economic outcomes is substantially lower than for the PMT.

5.1.3 Relationship between phone use and vulnerability

A third hypothesis for why we are unable to predict any of our vulnerability indices from mobile phone data (when we are, to some extent, able to predict poverty) is that mobile phone use may be more closely related to long-term poverty outcomes than to short-term vulnerability metrics. For example, mobile money and mobile data usage are important

[17] To further confirm this result — and to test the validity of our pipelines for machine learning with mobile phone data — we replicate the experiment of predicting treatment status from six months of mobile phone data during the treatment period, this time including 'cheat code' features relating to mobile money use in the machine learning model. These features include information on the number and sizes of transactions placed and received by each subscriber, and thus directly reveal information about whether a subscriber has received a GD-Novissi cash transfer via mobile money. With the mobile money-related features included, the area under the curve score for predicting treatment status is 0.998. Table S11, which shows the feature importances for this machine learning model, further confirms that the key features used by the model relate to mobile money transactions.

predictors of wealth in Aiken et al. (2022b), and are related to long-term investments in smartphones and financial services technologies. Short-term changes in food security and perceived economic status may not result in the types of investments in phone capabilities (such as buying a new smartphone or investing in a large airtime bundle) that would be observable in mobile phone metadata. We observe that the correlation between the PMT and other outcome indices is modest in magnitude and sometimes negative (r = -0.08 - 0.1, Figure S2), suggesting that the PMT is measuring a different type of well-being than the other outcome indices. It could be that the types of poverty the PMT measures are more closely related to phone use than non-economic outcomes.[18]

5.1.4 Spatial structure in outcome indices

A fourth and final hypothesis is that poverty may have more geographic structure than the non-economic outcomes we examine. Spatial features related to the cell towers that subscribers use are important features in phone-data-based poverty prediction models (cf. Aiken et al., 2022b). It is possible that non-economic outcomes are less predictable from phone data because they are less geographically determined. To test whether spatial structure could explain the difference in predictive power between the PMT and other outcomes, Table S10 shows the within and between variance grouped by canton for both the pre-treatment and endline survey. We find that the ratio of between to within variance is substantially higher for the PMT (8.2 - 15.0 in the pre-treatment and endline surveys) than for any of the vulnerability indices (1.3 - 2.1). This result, combined with past documentation that spatial structure plays a key role in estimating poverty from mobile phone data (Aiken et al., 2022b; Hernandez et al., 2017), suggests that spatial structure in an index may be an important determinant of whether it can be accurately predicted from mobile phone data.

5.2 Additional challenges to estimating treatment effects from mobile phone data

Even if it were possible to accurately predict welfare outcomes from phone data, it might still prove difficult to use phone data to estimate the treatment effects of cash transfers on those same outcomes. Here, we provide suggestive evidence of two such issues: the modest size of the GD-Novissi cash transfers generates only small changes in phone use; and 'model

[18] On the other hand, prior work has shown that phone use changes in response to high-magnitude short-term shocks (Bagrow et al., 2011; Blumenstock et al., 2016); we might therefore expect that phone data would reflect short-term changes in welfare.

drift' in the relationship between phone use and vulnerability complicates the repeated use of machine learning models over time.

5.2.1 Magnitude of impacts

A challenge for detecting treatment effects from mobile phone data in the context of Novissi is the program's modest transfer sizes and welfare impacts. Our survey-based impact evaluation results detect treatment effects of 0.04-0.07 SD resulting from five monthly transfers of USD 13-15. Interventions of a larger magnitude would be expected to produce larger impacts (for example, Haushofer and Shapiro (2016) report a 0.26 SD increase in food security and mental health following a USD 404-1,525 PPP cash transfer in Kenya). The modest transfer sizes and impacts of the GD-Novissi program result in impacts on phone use (Table 2) of modest magnitude, which are difficult for an ML model to detect. The effects of larger transfers may be easier to recover from mobile phone data.

5.2.2 Model drift

A specific challenge to identifying treatment effects in the first regime we study — training a model prior to program roll-out and deploying it later on to monitor impacts — is *model drift* in the relationship between phone use and vulnerability over time. Particularly in the context of shocks like the COVID-19 pandemic, a model trained well before a program's implementation may no longer be accurate when cash transfers are distributed. Aiken et al. (2022b) empirically study model drift in Togo, and find a substantial drop in accuracy when a model is trained two years prior to its deployment (Spearman correlation of 0.42 at the time of training vs. 0.35 at the time of deployment). To test the extent to which the same issues of model drift are present in this work, we evaluate the accuracy of predictions from our poverty prediction model trained on the pre-treatment survey for generating predictions in the treatment period. In comparison to the R^2 score of 0.049 for the model trained on the endline survey, the predictions from the model trained on the pre-treatment survey achieve an R^2 of 0.030, providing suggestive evidence of model drift in the nine months that elapsed between the pre-treatment and endline surveys. However, given the differences in sampling approach between our baseline and impact evaluation surveys, we are wary to draw strong conclusions from this result.

6 Conclusion

The combination of non-traditional administrative data and machine learning has made it possible to estimate socioeconomic and demographic characteristics at a fraction of the cost of traditional surveys. We test whether machine learning predictions using phone data are sufficiently accurate to estimate the impact of a cash transfer program in Togo during COVID. While survey data produce modest and statistically significant treatment effect estimates (on food security, mental health, and perceived economic status), the phone data alone and accompanying machine learning predictive analytics generally do not (although mental health improvements are slightly predicted). Yet prior work does find phone data alone are predictive of poverty, specifically wealth. We infer that phone data and machine learning algorithms may be more useful in contexts where an intervention affects stocks such as wealth, when effects are larger, or when the treated population is more heterogeneous prior to treatment.

References

Aiken, E., Bedoya, G., Blumenstock, J., and Coville, A. (2022a). Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in Afghanistan. *arXiv preprint arXiv:2206.11400.*

Aiken, E., Bellue, S., Karlan, D., Udry, C., and Blumenstock, J. E. (2022b). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 603(7903):864–870.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American Statistical Association*, 103(484):1481–1495.

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023). Prediction-powered inference. *Science*, 382(6671):669–674. Publisher: American Association for the Advancement of Science.

Angrist, J. D., Hull, P. D., Pathak, P. A., and Walters, C. R. (2017). Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics*, 132(2):871–919.

Bagrow, J. P., Wang, D., and Barabási, A.-L. (2011). Collective Response of Human Populations to Large-Scale Emergencies. *PLoS ONE*, 6(3):e17680.

Banerjee, A., Faye, M., Krueger, A., Niehaus, P., and Suri, T. (2020). Effects of a universal basic income during the pandemic. *Innovations for Poverty Action Working Paper*.

Banerjee, A., Karlan, D., Trachtman, H., and Udry, C. R. (2022). Does poverty change labor supply? Evidence from multiple income effects and 115,579 bags. *National Bureau of Economic Research*, 27314.

Banerjee, A. V., Hanna, R., Kreindler, G. E., and Olken, B. A. (2017). Debunking the stereotype of the lazy welfare recipient: Evidence from cash transfer programs. *The World Bank Research Observer*, 32(2):155–184.

Barriga-Cabanillas, O., Blumenstock, J. E., Lybbert, T., and Putman, D. (2021). The potential and limitations of big data in development economics: The use of cell phone data for the targeting and impact evaluation of a cash transfer program in Haiti? *Presentation at 2021 Pacific Development Conference.*

Bastagli, F., Hagen-Zanker, J., Harman, L., Barca, V., Sturge, G., Schmidt, T., and Pellerano, L. (2016). Cash transfers: what does the evidence say. *A rigorous review of programme impact and the role of design and implementation features. London: ODI*, 1(7).

Bellman, R. and Kalaba, R. (1959). A mathematical theory of adaptive control processes. *Proceedings of the National Academy of Sciences*, 45(8):1288–1290.

Blumenstock, J., Cadamuro, G., and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076.

Blumenstock, J. E. (2018). Estimating economic characteristics with phone data. In *AEA Papers and Proceedings*, volume 108, pages 72–76.

Blumenstock, J. E., Eagle, N., and Fafchamps, M. (2016). Airtime transfers and mobile communications: Evidence in the aftermath of natural disasters. *Journal of Development Economics*, 120:157–181.

Bottan, N., Hoffmann, B., and Vera-Cossio, D. A. (2021). Stepping up during a crisis: The unintended effects of a noncontributory pension program during the COVID-19 pandemic. *Journal of Development Economics*, 150:102635.

Bryan, G., Choi, J. J., and Karlan, D. (2021). Randomizing religion: the impact of protestant evangelism on economic outcomes. *The Quarterly Journal of Economics*, 136(1):293–380.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632.

Chi, G., Fang, H., Chatterjee, S., and Blumenstock, J. E. (2022). Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3):e2113658119.

Crosta, T., Karlan, D., Ong, F., Rüschenpöhler, J., and Udry, C. (2024). Unconditional cash transfers: A bayesian meta-analysis of 50 randomized evaluations in 26 low and middle income countries. *Working Paper*.

Dobbie, W. and Song, J. (2020). Targeted debt relief and the origins of financial distress: Experimental evidence from distressed credit card borrowers. *American Economic Review*, 110(4):984–1018.

Egger, D., Miguel, E., Warren, S. S., Shenoy, A., Collins, E., Karlan, D., Parkerson, D., Mobarak, A. M., Fink, G., Udry, C., et al. (2021). Falling living standards during the COVID-19 crisis: Quantitative evidence from nine developing countries. *Science Advances*, 7(6):eabe0997.

Fatehkia, M., Tingzon, I., Orden, A., Sy, S., Sekara, V., Garcia-Herranz, M., and Weber, I. (2020). Mapping socioeconomic indicators using social media advertising data. *EPJ Data Science*, 9(1):22.

Gilraine, M., Gu, J., and McMillan, R. (2020). A new method for estimating teacher value-added. *National Bureau of Economic Research*, 27094.

Haushofer, J. and Shapiro, J. (2016). The short-term impact of unconditional cash transfers to the poor: Experimental evidence from Kenya. *The Quarterly Journal of Economics*, 131(4):1973–2042.

Hernandez, M., Hong, L., Frias-Martinez, V., Whitby, A., and Frias-Martinez, E. (2017). Estimating poverty using cell phone data: Evidence from Guatemala. *World Bank Policy Research Working Paper*, (7969).

Hjelm, L., Mathiassen, A., and Wadhwa, A. (2016). Measuring poverty for food security analysis: Consumption versus asset-based approaches. *Food and Nutrition Bulletin*, 37(3):275–289.

Huang, L. Y., Hsiang, S. M., and Gonzalez-Navarro, M. (2021). Using satellite imagery and deep learning to evaluate the impact of anti-poverty programs. *National Bureau of Economic Research*, 29105.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Kane, T. J. and Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. *National Bureau of Economic Research*, 14607.

Karlan, D., Lowe, M., Osei, R. D., Osei-Akoto, I., Roth, B. N., and Udry, C. R. (2022). Social protection and social distancing during the pandemic: Mobile money transfers in ghana. *National Bureau of Economic Research*, 30309.

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S.-L., Walters, E. E., and Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, 32(6):959–976.

Londoño-Vélez, J. and Querubin, P. (2022). The impact of emergency cash assistance in a pandemic: Experimental evidence from Colombia. *Review of Economics and Statistics*, 104(1):157–165.

Ratledge, N., Cadamuro, G., de la Cuesta, B., Stigler, M., and Burke, M. (2022). Using machine learning to assess the livelihood impact of electricity access. *Nature*, 611(7936):491–495.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, pages 130–134.

Sheehan, E., Meng, C., Tan, M., Uzkent, B., Jean, N., Burke, M., Lobell, D., and Ermon, S. (2019). Predicting economic development using geolocated Wikipedia articles. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2698–2706.

Tadesse, G., Abate, G. T., and Zewdie, T. (2020). Biases in self-reported food insecurity measurement: A list experiment approach. *Food Policy*, 92:101862.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., and Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11(1):1–11.

Table 1: Treatment Effects of GD-Novissi

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Food	Financial	Financial	Mental	Perceived	Health	Labor	**All seven**
	security	health	inclusion	health	status	care access	supply	**indices**
Panel A: Treatment effects from endline survey								
Treatment	0.064***	0.026	0.007	0.072***	0.040*	0.010	0.009	0.061***
	(0.022)	(0.024)	(0.021)	(0.019)	(0.022)	(0.023)	(0.025)	(0.023)
Obs.	9,511	9,511	9,511	9,511	9,511	9,511	9,511	9,511
Control Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FDR q-value	0.012	0.272	0.577	0.002	0.103	0.577	0.577	0.016
Panel B: Phone-data-based treatment effects, using ML model trained on pre-treatment survey								
Treatment	-0.003	-0.013	—	—	-0.013	—	0.000	—
	(0.016)	(0.014)	—	—	(0.013)	—	(0.012)	—
Obs.	48,759	48,759	—	—	48,759	—	48,759	—
Control Mean	0.000	0.000	—	—	0.000	—	0.000	—
Z-test p-value	0.016	0.159	—	—	0.483	—	0.002	—
Panel C: Phone-data-based treatment effects, using ML model trained on endline survey								
Treatment	0.006	0.005	0.005	0.030*	0.002	0.031**	0.015	0.021
	(0.014)	(0.019)	(0.022)	(0.019)	(0.018)	(0.015)	(0.013)	(0.017)
Obs.	39,252	39,252	39,252	39,252	39,252	39,252	39,252	39,252
Control Mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Z-test p-value	0.027	0.499	0.938	0.114	0.178	0.445	0.837	0.163
Panel D: Endline survey + ML model trained on endline survey, using prediction-powered inference								
Treatment	0.066***	0.048*	-0.006	0.095***	0.059**	0.030	0.036	0.087***
	(0.031)	(0.032)	(0.030)	(0.031)	(0.031)	(0.030)	(0.032)	(0.031)
Obs. (Survey)	9,511	9,511	9,511	9,511	9,511	9,511	9,511	9,511
Obs. (Predictions)	48,756	48,756	48,756	48,756	48,756	48,756	48,756	48,756
Control Mean (Survey)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Control Mean (Predictions)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Z-test p-value	0.937	0.582	0.767	0.509	0.581	0.653	0.460	0.452

Notes: *Panel A:* Treatment effects of GD-Novissi estimated using the endline survey. The dependent variable for each regression is indicated in the column title; see Appendix A for variable construction. All regressions control for the enumerator, week of the survey, and strata fixed effects. All observations are weighted by sampling and response probabilities. Robust standard errors are in parentheses. *Panels B-D*: Treatment effects derived using the phone-data-based machine learning model predictions from Table 3. In *Panel B*, the pre-treatment survey is used to train the machine learning model and predictions are used for inference from all individuals enrolled in the RCT (whether or not they were surveyed in the pre-treatment survey). In *Panel C*, the endline survey is used to train the model, and only predictions from individuals not surveyed at endline are used for inference. In both Panels A and B, treatment effects are estimated by regressing the phone-data-based estimate of the outcome variable on treatment status, with standard errors estimated using a Bayesian bootstrap. In *Panel D*, endline survey data are used together with endline predictions (for all individuals – surveyed and non-surveyed) using the prediction-powered inference methodology developed by Angelopoulos et al. (2023). The Z-test p-value in each of Panels B-D indicates the significance of the Z-test that the phone-data-based treatment effect (reported in the panel) and the survey-based treatment effects (reported in Panel A) are different. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table 2: Treatment Effects on Phone Use

	(1) Active days	(2) Calls	(3) Texts	(4) Contacts	(5) International Contacts	(6) % Initiated	(7) Regions	(8) Prefectures
Treatment	0.064***	0.021**	0.010	0.033***	0.015	-0.026***	0.049***	0.038***
	(0.009)	(0.009)	(0.010)	(0.009)	(0.013)	(0.010)	(0.009)	(0.009)
Control Mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Unstandardized Control Mean	98.892	625.654	70.979	77.098	3.164	0.902	4.236	9.867
Obs.	48,803	48,664	44,548	48664	22,410	44,548	48,803	48,803

Notes: Treatment effects on basic metrics of mobile phone use, selected from among the 823 metrics of mobile phone use used by our machine learning models. Metrics were selected by hand from the pool based on ease of interpretation: (1) active days of phone use, (2) total incoming and outgoing calls, (3) total incoming and outgoing texts, (4) unique contacts, (5) unique international contacts, (6) share of the individual's transactions initiated by them (rather than received from a contact), (7) unique regions visited (based on locations of mobile antennas), and (8) unique prefectures visited (based on locations of mobile antennas). All features are calculated over the entire six month treatment period (November 2020 - April 2022). All features are standardized to zero mean and unit variance in the control group (the unstandardized control mean is also provided for intuition). $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table 3: Predicting Welfare from Mobile Phone Data

	(0) PMT	(1) Food security	(2) Financial health	(3) Financial inclusion	(4) Mental health	(5) Perceived status	(6) Healthcare access	(7) Labor supply	(8) All seven indices
Panel A: Pre-treatment survey									
R^2	0.143	0.002	0.013	—	—	0.034	—	0.046	—
Obs (Training)	8,899	8,899	8,899	—	—	8,899	—	8,890	—
Panel B: Endline survey									
R^2	0.049	0.008	0.009	0.003	0.002	0.008	0.007	0.021	0.026
Obs (Training)	8,448	9,507	9,507	9,134	9,507	9,507	9,522	9,507	9,507

Notes: Performance of a gradient boosting model for predicting well being metrics from mobile phone data. In Panel A, the pre-treatment survey is used to train the machine learning model; in Panel B the endline survey is used to train the model. Sample weights are used in training and evaluation throughout.

A Details of the Impact Evaluation Survey

A.1 Sample Frame

The sample frame for the endline survey was drawn from subscribers who successfully enrolled in the GD-Novissi RCT between November 2020 and January 2021 (N=49,083). Thus, the sample was restricted to individuals who (a) had active mobile phone accounts; (b) were registered to vote in one of Togo's 100 poorest cantons; (c) completed the registration procedure for GD-Novissi; and (d) were predicted, based on their mobile phone data, to consume less than $1.25 per day.

The sample was stratified by treatment status and geography. The former was done to maximize statistical power in estimating treatment effects; the latter was done to account for the fact that one large region (Savanes) received payments unrelated to GD-Novissi during the period when GD-Novissi benefits were being delivered.[19]

A.2 Response Rate

The enumerators called all 24,294 phone numbers of our final sample in random order. We successfully surveyed 10,129 individuals (response rate of 42%). After removing low-quality surveys (see Appendix A.3), our final sample contained 9,511 observations (completion rate of 39%). This completion rate is similar to other phone surveying completed during COVID-19: for example Egger et al. (2021) analyzes random-digit dialing to conduct surveys on well-being in nine countries during COVID-19 and reports completion rates ranging from 17% to 59%. Table S1 shows that attrition rates do not differ statistically significantly between the treatment and control groups.

A.3 Data Collection Monitoring

We identified surveyors who performed poorly by comparing the data collected with the information contained in Novissi administrative data. We began our analysis by constructing "enumerator effects" (EE) estimates for every enumerator in our data. We predicted the EE on the basis of the correct answers to five questions for which we obtained the "truth" from the Novissi registry (prefecture, canton, age, gender, and Novissi status), and on the frequency of very short surveys (below 15 minutes) as well as surveys with no children reported (which avoids the roster part of the survey and simplifies the surveyor's work). We controlled for interviewee characteristics such as region and interview language to separate the enumerator's impact from observable interviewee selection.[20] Our approach to estimating

[19] See Appendix C.2 for details on the other Savanes program. Of the 36,090 subscribers registered in the Savanes region, we sampled 36%; of the 12,993 subscribers registered outside of the Savanes region, we sampled 88%.

[20] The phone number list was randomized and then distributed to the enumerators, so we believed that there is little room for sorting.

EE parallels the parametric empirical Bayes estimator of teacher effects (Kane and Staiger (2008); Chetty et al. (2014); Gilraine et al. (2020)).

We then normalized the EEs for each of the seven dimensions (prefecture, canton, age, gender, Novissi status, number of short surveys, number of surveys without kids), and took the sum of the coherently signed components for enumerators who conducted more than ten interviews. We classified the interviews of enumerators with an average EE lower than the sample mean minus two standard deviations as of "very poor quality" and remove them from the sample. 615 observations collected by five enumerators who were ranked "very poor quality" were removed from the dataset. In addition, on the second day of the survey, while monitoring data quality, we noticed an enumerator who was performing extremely poorly. After a warning from his supervisor, the quality of his data collection improved. We removed the data collected by this enumerator during the first two days (60 observations). Thus, we only use data from only 9,511 high-quality surveys in our main analysis.

A.4 Weights

We reweight observations in the endline survey by the inverse of the sampling probability and the inverse of the probability of response. Sampling probabilities are determined by four sampling strata: Savanes, treatment: 30.48%; Savanes, control: 41.26%; Outside Savanes, treatment: 76.25%; Outside Savanes, control: 100.00%.

To calculate response weights, we train a machine learning model to predict survey response from pre-survey covariates. In total, 9,511 phone numbers completed the survey out of 24,294 numbers sampled. We include the following pre-survey covariates as predictors:

- 824 features relating to phone use in the six months pre-survey (Nov 2022 - Apr 2022).
- 6 features from the Novissi registry: Age, gender, canton of registration (one hot encoded), number of payments received up until the survey date, profession (one hot encoded for the 20 most common professions), and registration week (one hot encoded).
- An indicator for treatment vs. control group.

Using a similar pipeline to the machine learning methods described in Section 4.1.2, we train a LightGBM classifier to predict response, and produce an out-of-sample predicted probability of response for each phone number using five-fold cross-validation. We tune hyperparameters using three-fold cross-validation separately on each of the five folds. With all predictions pooled together, our model achieves an AUC of 0.69.

The overall weight for each observation is the product of the inverse of the sampling probability and the inverse of the (predicted) probability of response.

A.5 Outcomes

The survey contained modules on household food security and consumption, health, access to social services, poverty, mental health, and experience with the Novissi program. Following our pre-analysis plan (American Economic Association Registry #7590), we constructed seven primary indexed outcomes using the index construction methodology described in Bryan et al. (2021). These seven indices are: food security, financial health, financial inclusion, mental health, perceived socioeconomic status, health care access, and labor supply.[21]

We standardize all our outcomes so that, within our control group of eligible active mobile subscribers, each outcome had zero mean and unit variance, with the exception of the mental health measure for which we use the Kessler K6 distress scale methodology (Kessler et al., 2002). Specifically, we first standardize each component — signed coherently beforehand — by subtracting its control group mean and dividing by its control group standard deviation. We then calculate the sum of the standardized components and standardize the sum again by the control group standard deviation.[22]

In addition to these seven primary welfare outcomes, we collected a proxy-means test (PMT) for each subscriber that proxies for consumption. We used the proxy-means test developed in Aiken et al. (2022b), which used machine learning methods to select twelve features that are most jointly predictive of consumption (training on data from a nationally representative household survey conducted in Togo in 2018).

A.6 Attrition and Balance Checks

We test for differential nonresponse between the treatment group and the control group in the impact evaluation survey by regressing a binary indicator for response on treatment status, among all 24,294 phone numbers called. In Table S1, we find that there is no statistically significant difference in response rates between the treatment and control groups.

We also test for covariate balance between the treatment and control groups in our impact evaluation survey sample in Table S2. We find that the treatment and control groups in the impact evaluation survey are balanced on self-reported age, gender, and occupation (Panel A), with results robust to substituting administrative data from the Novissi program for self-reported survey data (Panel B).

[21] Table S3 reports the specific wording of each component of each outcome index.

[22] We impute missing components using the other components in an index unless the missing components are children-related and the family had no children, in which case we compute the index omitting those components.

B Details of the Pre-Treatment Survey

Details of the sampling and design for the pre-treatment survey — conducted pre-program in September 2020 (see Figure S1) — are available in Aiken et al. (2022b). In short, the sample frame for the survey was all 240,000 mobile subscribers active in Togo between March 1 and September 30, 2020, and inferred based on their phone use to be living in cantons eligible for the GD-Novissi program. 30,244 of these phone numbers were randomly drawn for surveys; of these, 9,484 completed the survey and were included in the final sample.

The phone survey collected information on poverty (the PMT, see Appendix A) and the components of three of the four indices for which we observe significant treatment effects of GD-Novissi: food security, financial health and perceived socioeconomic status. The financial health and perceived socioeconomic status indices are constructed identically to the indices in the impact evaluation survey; however, only certain components of the food security index were collected in the pre-treatment survey, so we construct a "reduced food security index" for the pre-treatment survey. This index is less comprehensive than the food security index collected in the impact evaluation survey (Table S3) and does not include questions on food consumption. The components for the reduced food security index are listed in Table S4.

As in the impact evaluation survey, each index is constructed following Bryan et al. (2021), by standardizing each component across the surveyed population, summing components, and then standardizing the resulting index. Also as in the impact evaluation survey, each observation is weighted by the inverse of the sampling probability and the inverse of the probability of response (see Aiken et al. (2022b)). We use weights throughout our analysis involving the pre-treatment survey, except where otherwise noted.

C Treatment Effect Heterogeneity

We test for treatment effect heterogeneity on four pre-registered dimensions: gender, poverty, occupation, and region of residence (in or outside of the Togo's northernmost region, Savanes). For each dimension, we test for heterogeneous treatment effects on our seven outcomes and the aggregate welfare index in Table 1 Panel A.

C.1 Heterogeneity by gender, wealth, and occupation

We find little evidence that treatment effects were heterogeneous across the socioeconomic and demographic subgroups that we pre-specified in our pre-analysis plan. In particular, while GD-Novissi had an important gender component, whereby women received roughly 15% more money per month than men, the welfare impacts on women were not significantly larger than for men. These results can be seen in Panel A of Table S5: while women, in general, are worse off than men (the third row of coefficients), the coefficient on the interaction between treatment and female is never significantly different from zero.[23]

Panels B and C of Table S5 likewise indicate that treatment effects did not differ by pre-treatment wealth or occupation. In Panel B, we compare treatment effects for people with PMT scores above and below the sample median, and, with the exception of healthcare access, do not observe significant differences for any outcome. Panel C indicates that treatment effects were not different for farmers – the most common occupation in the rural areas of Togo where GD-Novissi was implemented, and the occupation reported by 60% of endline survey respondents.

C.2 Geographic heterogeneity

There is, however, one dimension where we find evidence of substantial heterogeneity in treatment effects, which is by the location of the beneficiary. In particular, Panel D of Table S5 highlights how treatment effects on food security and mental health were significantly larger for beneficiaries in the Savanes region in the far North of Togo. Indeed, with the exception of healthcare access, the treatment effects for beneficiaries outside Savanes are all close to zero and no longer statistically significant once we account for the differential effect of treatment in Savanes.

Savanes is unique in several respects: most GD-Novissi beneficiaries (70%) reside in the Savanes region, it is generally poorer than other regions eligible for GD-Novissi, it had higher rates of COVID-19 and related curfews than other regions, and the government provided an independent round of cash transfers called *Savanes-Novissi* (unconnected to GD-Novissi) to all residents in Savanes in February 2021 (two months before our endline surveys were conducted). While understanding the reasons for the substantial geographic heterogeneity between Savanes and non-Savanes beneficiaries is not the focus of this paper, we explore briefly below three possible hypotheses that could explain why the treatment effects of GD-Novissi are observed mainly in the Savanes region: (i) that differential registration for Savanes-Novissi between the GD-Novissi treatment and control groups resulted in additional cash impacts for the treatment group, (ii) that mobility reductions resulting from curfews in the Savanes region made cash transfers more impactful in Savanes than the rest of the country, and (iii) that price differences between Savanes and the rest of the country gave cash transfers more purchasing power in Savanes.

[23] According to administrative data from the Novissi program, women represent half of GD-Novissi beneficiaries, and 45% of our surveyed sample. However, the share of women among survey respondents is only 27% — that is, 40% of the phone numbers registered with female voter ID cards were answered by men. This could be indicative of a high degree of phone sharing at a household level and/or strategic behavior in which phones owned by men were used to register female voter IDs to maximize benefits.

C.2.1 Interaction Between GD-Novissi and Savanes-Novissi

In addition to the GD-Novissi program considered here, the Government of Togo implemented three other targeted cash transfer programs under the Novissi umbrella during the pandemic period. One of these, Savanes-Novissi, provided one-time cash transfers of USD 8-10 to all residents of Savanes who registered for Novissi in a two-week period beginning on February 22, 2021. Women received a one-time transfer of CFA 6,125 (USD 9.80), and men received a transfer of CFA 5,250 (USD 8.40). A total of 244,302 Savanes residents registered for and received Savanes-Novissi, of whom 114,311 (46.79%) were already registered for GD-Novissi.

We observe an approximately 20 percentage point difference in registration rates for Savanes-Novissi between the treatment and control groups in GD-Novissi, with the control group substantially more likely to register for the Savanes-Novissi program. 41% of the treatment group registered for Savanes-Novissi, while 63% of the control group registered.

There are two plausible explanations for the difference in enrollment: first, GD-Novissi provided enough assistance for the treatment group, so they were less in need of further cash transfers, and second, that confusion in communications around the two programs resulted in members of the treatment group believing they were ineligible for Savanes-Novissi. The second explanation is particularly plausible for two reasons. First, people located in Savanes who were registered for GD-Novissi were eligible for Savanes-Novissi, but were required to register separately for Savanes-Novissi, which could be confusing. Second, because the treatment group was receiving cash transfers from GD-Novissi in February 2021, the government of Togo initially excluded treated people of GD-Novissi from the Savanes-Novissi program. While the Savanes-Novissi amount was transferred at the time of registration for everyone else, people in the GD-Novissi treatment group received Savanes-Novissi cash transfers in the second week of the period of registration only.

We included specific questions in the impact evaluation survey to distinguish between the two explanations. We first asked people if they registered with Savanes-Novissi. If not, we asked them an open-ended question why not. The enumerators were told to classify the answers in one of the eight pre-defined categories, including "other" and "I don't know". Three of the possible categories are related to the confusion hypothesis ("I did not think I was eligible", "I did not think I needed to register, since I already registered with GD-Novissi", and "I heard about the program after the end of the registration period"). Three others are related to GD-Novissi impact hypothesis ("I receive GD-Novissi, I don't need extra money", "I have enough money, I don't need extra money", and "Other people are more in need than me, I prefer them to get the money").

Qualitatively, the treatment group is more likely to be confused about the eligibility criteria than the control group. The first main reason why people did not register is the lack of information, and there is a ten percentage points difference between the treatment and the control group: 36.5% of the control group versus 49.6% of the treatment group was confused about the eligibility criteria. Less than 3% of people in both treatment arms reported a lack of need for Savanes-Novissi as the main reason, supporting the second hypothesis for the enrollment differences (confusion in eligibility criteria).

However, in comparing the welfare outcomes of people who did and did not register with Savanes-Novissi by treatment arm (Table S12), we do observe that people from the treatment group who did not register with Savanes-Novissi have a higher food security and financial health index than those who did register. There are no such differences between registrants and non-registrants in the control group. The fact that the treatment group self-selected in Savanes-Novissi supports the first hypothesis, suggesting that GD-Novissi contributed to the low enrollment rates for Savanes-Novissi in the treatment group.

We conclude, based on this evidence, that both hypotheses (the welfare impact of GD-Novissi and confusion around eligibility criteria) likely contributed to lower registration rates for GD-Novissi in the treatment group in Savanes. Importantly, however, the difference in registration rates does not explain the larger welfare impacts of GD-Novissi in Savanes in comparison to the rest of the country: if anything, we would expect the lower registration rates for Savanes-Novissi in the treatment group to attenuate treatment effects relative to other regions of the country.

C.2.2 Price Differences

A final testable explanation for the GD-Novissi treatment effects in Savanes (in comparison to the rest of the country) is that price differences between the Savanes region and the rest of the country give GD-Novissi transfers more purchasing power in Savanes. We collected price information for staple goods in the consumption module of the impact evaluation survey; in our analysis, we restrict to goods for which at least 50% of the respondents provided a price. Among these seven goods, we observe statistically significant differences in prices between Savanes and the rest of the country for only three goods: palm oil and milk are more expensive in Savanes, while Niebe is cheaper (table available upon request). Given that there are no systematic price differences in a consistent direction between Savanes and the rest of the country, we conclude that price differences are not a major driver of the GD-Novissi treatment effects in Savanes.

C.2.3 Discussion

We explored three channels to account for the program's differential impact in the Savanes region and could not conclude that one drove the differential impacts. However, we cannot completely rule out two underlying explanations: differences in regional poverty levels and the effects of the lockdown. Despite not finding evidence that treatment effects vary by wealth or occupation nor identifying systematic price differences across regions, poverty levels might still drive the differential impact. In Savanes, the country's poorest region with unique norms and culture, a given transfer may be more meaningful than elsewhere.

Similarly, we could not link higher mobility levels with higher treatment effects. If mobility does not accurately proxy for the lockdown effects, this evidence does not eliminate the possibility that the lockdown in Savanes negatively affected the residents who benefited the most from the GD-Novissi transfers.

D Related Work on COVID-19 Cash Transfers

Since the COVID-19 pandemic, a growing body of research has emerged to document the welfare impacts of cash transfers distributed in response to the pandemic. Many of these studies are reviewed in Karlan et al. (2022). Broadly, this literature shows modest, positive, and statistically significant impacts of cash transfers on a wide range of welfare metrics, including food security and mental health.

Specifically, Banerjee et al. (2020) use phone surveys and an RCT design to show that universal basic income transfers of USD 22.5 nominal per month to households under lockdown in Kenya reduced the probability of households experiencing hunger (by 5-11 percentage points, relative to a control mean of 68%), and had modest positive impacts on mental health. Similarly, Londoño-Vélez and Querubin (2022) use an RCT and phone surveys to measure impacts of a monthly VAT refund of USD 19 in Colombia, finding a 4.4 percentage point increase in the probability of treated households purchasing food in the week preceding the survey (relative to a control mean of 72%), but no statistically significant impacts on food security. The paper also reports positive and statistically significant impacts on mental health indices (1.2-2.1 percentage points) and a financial health index (0.055 standard deviations). Karlan et al. (2022) follow a similar experimental design, using an RCT and several rounds of phone surveys to evaluate the impact of eight monthly cash transfers of $15, recording an 8% increase in food consumption among treated households. In a non-randomized approach, Bottan et al. (2021) use online surveys and a regression discontinuity design to show that pension payments of USD 43-50 per month in Bolivia decreased the probability of households going hungry by 8-12 percentage points, relative to a comparison mean of 22%.

The first portion of our analysis contributes to this literature by documenting the impacts of pandemic cash transfers in Togo, using an extensive cash transfer program where treatment was randomly assigned at the individual level. While the cash transfers we study are smaller ($13-15.50 per month) than most of the other programs studied ($15-50 per month), we document comparable effect sizes (0.04-0.07 standard deviations).

Our results on heterogeneous treatment effects (Table S5) are also broadly consistent with the other papers studying the impacts of COVID-19 cash transfers on well-being, which for the most part do not find significant heterogeneity across dimensions studied (Londoño-Vélez and Querubin, 2022; Karlan et al., 2022). However, two results stand in contrast: Londoño-Vélez and Querubin (2022) finds that treatment effects are driven primarily by households in urban areas, while we find that treatment effects are driven primarily by households in Savanes, which is the most rural region of Togo; and Karlan et al. (2022) finds that treatment effects on food security are larger for female-headed households than male-headed households, whereas we find no heterogeneous treatment effects by gender of the recipient.

E Predicting welfare using mobile phone features

For each outcome index captured in the pre-treatment survey, we calculate the five-fold cross-validated R^2 as follows. The full dataset that matches completed surveys to phone records (N=8,899) is divided randomly into five partitions ("folds"). A machine learning model is trained on four of the five folds and predictions are produced for observations in the remaining fold; the process is repeated for each of the remaining four folds. The percentage of variation explained by the predictions (R^2) is then calculated, pooling predictions across all the folds. Survey weights and response weights are used in both training and calculating R^2 scores. We use a gradient boosting model; results are similar or worse for other ML methods, including linear models and random forests. We tune several hyperparameters using nested cross-validation: (1) Winsorization of features (selected from {no winsorization, 1% winsorization}, (2) minimum data in each leaf of the forest (selected from {10, 20, 50}), (3) number of leaves for each tree (selected from {5, 10, 20}), (4) learning rate (selected from {0.05, 0.075, 0.1}), and (5) number of trees (selected from {50, 100, 200}).

The Bayesian bootstrap procedure (Rubin, 1981) incorporates both the first-stage uncertainty in our ML models' predictions and the variation in predictions across treatment and control subscribers. Following Angrist et al. (2017) and Dobbie and Song (2020), we assign each observation that appears in the training and/or inference sets a "bootstrap weight" drawn from a Dirichlet distribution Dirichlet(1, ..., 1). These weights are used (in combination with the survey and response weights) in training the ML model, and in calculating treatment effects. We repeat this procedure with 100 random draws, and report the mean and standard deviation across these 100 bootstrap estimates. We select 100 bootstraps due to the computational expense of the procedure.

Following is a list of all mobile phone features used in our machine learning models. These features are calculated with open-source python library cider.[24]

The following features are calculated from metadata on calls and SMS messages. For quantities that are distributions (marked below with "(distribution)"), multiple moments of the distribution are used as features: the mean, standard deviation, median, skewness, kurtosis, minimum, and maximum. For all quantities, the feature is calculated separately for daytime, nighttime, weekdays, and weekends, as well as weekday daytime, weekday nighttime, weekend daytime, and weekend nighttime. Where applicable, features are also calculated separately for incoming and outgoing transactions.

[24] https://github.com/Global-Policy-Lab/cider

- Number of active days
- Number of unique contacts
- Call duration (distribution)
- Percent nocturnal
- Percent initiated conversations
- Percent initiated interactions
- Response delay (distribution, texts only)
- Response rate (texts only)
- Entropy of contacts
- Share of transactions with each contact that are outgoing (distribution)
- Interactions per contact (distribution)
- Time between transactions (distribution)
- Percent pareto interactions (share of contacts that account for 80% of transactions)
- Number of interactions
- Number of antennas used
- Entropy of antennas
- Radius of gyration
- Frequent antennas (number of antennas that account for 80% of transactions)
- Percent at home
- Number of international transactions
- Number of unique international contacts
- Number of days with international transactions

The following additional features are calculated about specific locations using metadata on calls and SMS messages:

- Number of transactions in each region of Togo
- Share of transactions in each region of Togo
- Number of transactions in each prefecture of Togo
- Share of transactions in each prefecture of Togo

The following features are calculated using information about mobile data usage:

- Total mobile data usage
- Mean, minimum, maximum, and standard deviation of daily mobile data usage
- Number of days with mobile data usage

F Additional Tests for Estimating Treatment Effects from Mobile Phone Data

In this section, we use alternative specifications to test whether it is possible to recover treatment effects from GD-Novissi mobile phone records. In table S6 we test using a two-week period to derive features from mobile phone data rather than a six-month feature period. In Table S7, we try using changes in features between the pre-treatment and during-treatment periods to predict each of our outcomes (using the endline survey as ground truth). In Table S8 we try predicting outcomes and inferring treatment effects in the Savanes region only, since the survey-based treatment effects were only observable in Savanes. In results available on request, we try the same specifications using only features that are statistically significantly different between the treatment and control groups (22% of all features). Finally, to test for whether noise in survey data is the cause of low predictive power, in Table S9 we train and evaluate a model to predict treatment status from the mobile phone feature set. The poor performance of each of these models suggests that it is the inability of phone data to identify differences between the treatment and control groups — rather than an issue of noisy survey data — that drives the low predictive power of the phone-based models and thus the null effects in downstream inference tasks.

Figure S1: GD-Novissi timeline

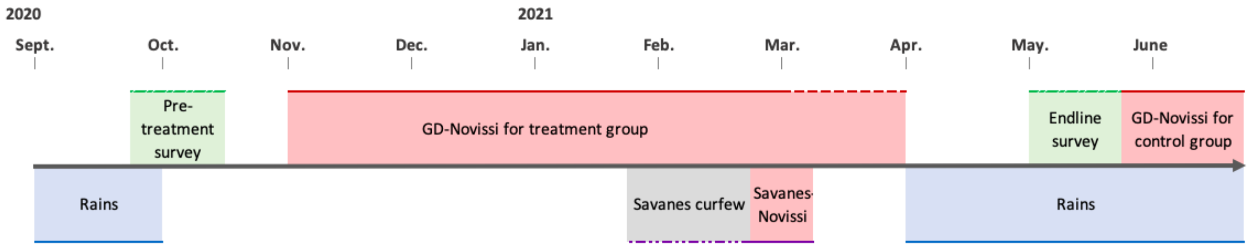
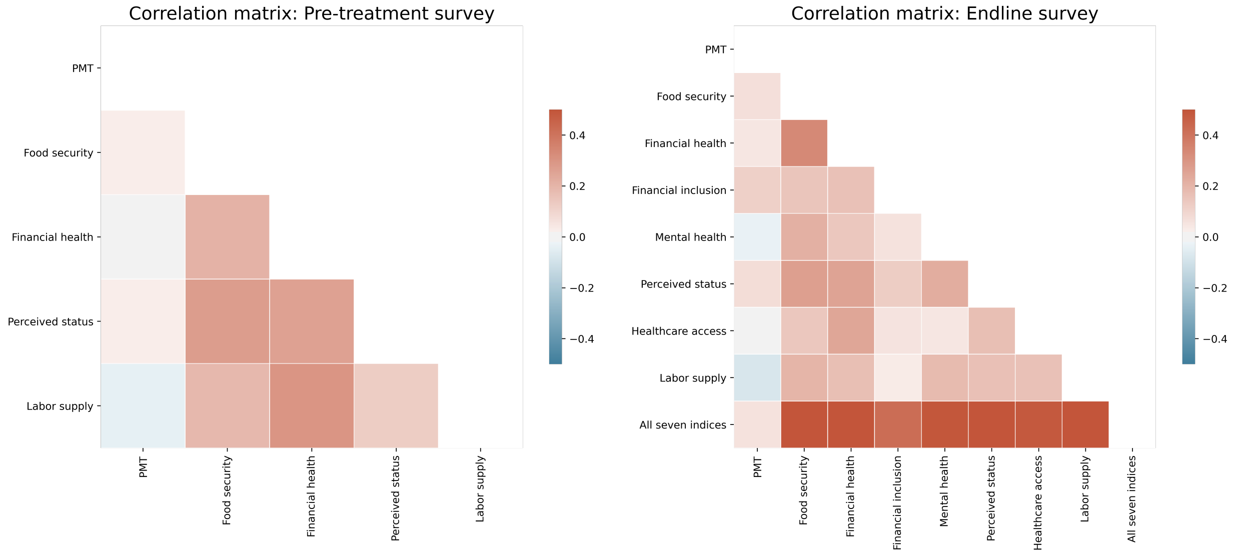


Figure S2: Correlation matrix for survey-based outcomes



Supplementary Figures and Tables

Table S1: Differential attrition

	Probability of non-response
Treatment	-0.01 (0.01)
N	24,294
Control Mean	0.61

Notes: Effect of treatment on attrition is estimated by regressing non-response on treatment, without fixed effects.

Table S2: Summary Statistics and Balance Checks

	Baseline Sample		*Endline Sample*		
	N	Mean	N	Mean	Diff. T-C
Panel A. Survey data					
PMT	8,821	$1.49	8,452	$1.31	$0.00
		(0.74)		(0.49)	(0.01)
Female	8,821	0.23	9,511	0.31	0.03**
		(0.42)		(0.46)	(0.01)
Age	8,716	33.37	9,310	36.03	-0.30
		(11.98)		(11.44)	(0.30)
Farmers	8,819	0.41	9,511	0.59	-0.02
		(0.49)		(0.49)	(0.01)
Savanes	8,821	0.51	9,443	0.72	-0.01
		(0.50)		(0.45)	(0.01)
Panel B. Novissi registry data					
Female	5,493	0.50	9,511	0.49	0.02*
		(0.50)		(0.50)	(0.01)
Age	5,493	36.02	9,429	37.63	0.09
		(13.96)		(12.70)	(0.33)
Farmers	5,402	0.23	9,375	0.38	-0.02*
		(0.42)		(0.49)	(0.01)
Savanes	5,493	0.52	9,511	0.74	-0.00
		(0.50)		(0.44)	(0.01)

Notes: Standard deviation in parenthesis. Column "Diff. T-C" contains the balance checks that are conducted by regressing the demographic variable of interest on treatment status (balance checks are conducted for the endline survey only). All observations are weighted by sampling probabilities. All regressions control for the enumerator, week of the survey, and strata fixed effects. Robust standard errors are in parenthesis. $^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01. See Appendix C.1 for a discussion of differences in demographic records between the survey and Novissi registry, particularly with respect to gender.

Table S3: Components for impact evaluation outcomes

Question	**Possible Answers**
Panel A. Food security	
Yesterday, how many meals did you eat?	0-3
In the past 7 days, how often were you unable to eat preferred foods because of a lack of money or other resources?	0 Never - 4 Every day
In the past 7 days, how often have you had to limit portion size at meal times?	0 Never - 4 Every day
In the past 7 days, how often did you have to reduce the number of meals eaten in a day?	0 Never - 4 Every day
In the past 7 days, how often have the children over 3 in your household had to reduce the number of meals eaten in a day?	0 Never - 4 Every day
Yesterday, how many meals did the children over 3 in your household eat?	0-3
When was the last time your household had each of the following items: Powdered milk, sugar, smoked anchovy, fresh onion, dried fish, sesame, red palm oil, traditional bread, orange, cowpea/dried beans.	0 Never - 5 Less than a week
How much did you spend on purchasing each of the items above, the last time?	Integer
Panel B. Financial health	
Were you able to save money last month? If so, how much?	Integer
God forbid, if your household stopped getting income from any source, how long could your household easily continue to meet your basic needs for food and housing? (Winsorized 95th percentile.)	Integer
God forbid, if there was a major emergency and your household needed money, how much money could you easily obtain within the next seven days? (Winsorized 95th percentile.)	Integer
Panel C. Financial inclusion	
Fraction of the adults in the household with a bank account.	Float
Fraction of the adults in the household (excluding the respondent) with a mobile money account.	Float
Panel D. Mental health (Kessler K6 nonspecific distress scale)	
During the past 7 days, about how often did you feel nervous?	Integer
During the past 7 days, about how often did you feel hopeless?	Integer
During the past 7 days, about how often did you feel restless or fidgety?	Integer
During the past 7 days, about how often did you feel that everything was an effort?	Integer
During the past 7 days, about how often did you feel so sad that nothing could cheer you up?	Integer
During the past 7 days, about how often did you feel worthless?	Integer
Panel E. Self perception of socioeconomic status	
In general, relative to other people in Togo, would you say that you are...	1 very poor - 5 very well off
How do you think other communities perceive the wealth of your household?	1 very poor - 5 very well off
Panel F. Labor supply	
Hours worked last week (winsorized 99th percentile)	Integer
During the past 7 days, how much income/pay did you receive?	Integer
Panel G. Healthcare access	
The last time you or someone else in your household needed healthcare, did you get healthcare?	Yes/no
When you last needed health care, did you get it at the hospital?	Yes/no
God forbid, if a child in your household needed to go the hospital, would you be able to bring him or her?	Yes/no

Notes: Components for each of the outcomes in the endline survey. All indices are produced using the index construction methadology from Bryan et al. (2021) except for the mental health index, which is based on simple addition of the components.

Table S4: Reduced food security index

Question	**Possible Answers**
Yesterday, how many meals did you eat?	0-3
In the past 7 days, how often were you unable to eat preferred foods because of lack of money or other resources?	0 Never - 4 Every day
In the past 7 days, how often have you had to limit portion size at meal times?	0 Never - 4 Every day
In the past 7 days, how often have you had to reduce the number of meals eaten in a day?	0 Never - 4 Every day
In the past 7 days, how often have the children in your household over age three had to reduce the number of meals eaten in a day?	0 Never - 4 Every day
Yesterday, how many meals did the children in your household over age three eat?	0-3
In the past 7 days, were you able to buy the amount of food you usually buy?	Yes/no

Notes: Components for the reduced food security index in the pre-treatment survey.

Table S5: Survey-based treatment effect heterogeneity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Food	Financial	Financial	Mental	Perceived	Health	Labor	All seven
	security	health	inclusion	health	status	care access	supply	indices
Panel A: Gender								
Treatment * Female	-0.037	-0.015	-0.070	0.006	0.003	-0.075	-0.011	-0.053
	(0.049)	(0.051)	(0.047)	(0.041)	(0.048)	(0.054)	(0.050)	(0.049)
Treatment	0.077***	0.034	0.025	0.072***	0.042	0.036	0.018	0.081***
	(0.025)	(0.029)	(0.025)	(0.023)	(0.027)	(0.026)	(0.033)	(0.027)
Female	-0.050	-0.121***	0.201***	-0.074**	-0.116***	-0.079**	-0.221***	-0.122***
	(0.035)	(0.039)	(0.034)	(0.030)	(0.037)	(0.037)	(0.037)	(0.036)
Panel B: Poverty								
Treatment * Poor	0.039	-0.021	-0.068	-0.049	-0.021	0.098**	-0.049	-0.019
	(0.045)	(0.052)	(0.046)	(0.039)	(0.047)	(0.048)	(0.054)	(0.048)
Treatment	0.053	0.048	0.040	0.105***	0.051	-0.048	0.024	0.072**
	(0.033)	(0.036)	(0.034)	(0.028)	(0.034)	(0.035)	(0.034)	(0.035)
Poor	-0.120***	-0.024	-0.111***	0.012	-0.080**	0.062*	0.058	-0.054**
	(0.030)	(0.037)	(0.033)	(0.029)	(0.034)	(0.033)	(0.040)	(0.035)
Panel C: Occupation								
Treatment * Farmer	0.016	0.032	0.024	-0.032	-0.027	0.049	0.028	0.024
	(0.046)	(0.050)	(0.044)	(0.039)	(0.046)	(0.048)	(0.052)	(0.048)
Treatment	0.052	0.006	-0.011	0.090***	0.053	-0.018	-0.010	0.043
	(0.037)	(0.039)	(0.035)	(0.031)	(0.036)	(0.039)	(0.043)	(0.039)
Farmers	-0.152***	-0.114***	-0.277***	-0.031	-0.172***	-0.002	-0.135***	-0.234***
	(0.032)	(0.037)	(0.032)	(0.028)	(0.035)	(0.033)	(0.040)	(0.035)
Panel D: Region								
Treatment * Savanes	0.087**	0.041	0.055	0.064*	0.065	-0.076*	0.015	0.066
	(0.042)	(0.044)	(0.041)	(0.037)	(0.044)	(0.045)	(0.043)	(0.043)
Treatment	0.000	-0.003	-0.033	0.025	-0.007	0.066*	-0.002	0.012
	(0.032)	(0.031)	(0.031)	(0.029)	(0.034)	(0.034)	(0.029)	(0.032)
Savanes	-0.076**	0.024	-0.037	0.014	-0.010	0.144***	0.017	0.020
	(0.032)	(0.033)	(0.030)	(0.029)	(0.035)	(0.032)	(0.033)	(0.033)
Obs	9,511	9,511	9,511	9,511	9,511	9,511	9,511	9,511

Notes: Heterogeneous treatment effects for outcomes for which we detect a statistically significant survey-based treatment effect in Table 1 Panel A. The dependent variable for each regression is indicated in the column title; see Appendix A for variable construction. In Panels A, C, and D, gender, occupation, and region of residence are determined by information provided by the respondents in the survey. In Panel B, poverty is determined by having a below-median PMT score. All regressions control for the enumerator, week of the survey, and strata fixed effects. All observations are weighted by sampling probabilities and response probabilities, and observations are restricted to subscribers who were active prior to the program's launch. Robust standard errors are in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table S6: Estimating Treatment Effects from Two Weeks of Mobile Phone Data

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	PMT	Food	Financial	Financial	Mental	Perceived	Healthcare	Labor	All seven
		security	health	inclusion	health	status	access	supply	indices
Panel A: Predicting welfare outcomes using ML trained on pre-treatment survey									
R^2	0.112	0.003	0.014	—	—	0.017	—	0.028	—
Obs.	8,593	8,593	8,593	—	—	8,593	—	8,584	—
Panel B: Predicting welfare outcomes using ML trained on endline survey									
R^2	0.031	0.004	0.006	-0.002	0.004	0.001	0.007	0.010	0.011
Obs.	8,238	9,261	9,261	8,898	9,261	9,261	9,276	9,261	9,261
Panel C: Phone-data-based treatment effects trained on the pre-treatment survey									
Treatment		0.001	0.014	—	—	-0.003	—	0.007	—
		(0.014)	(0.014)	—	—	(0.013)	—	(0.012)	—
Obs.		46,327	46,327	—	—	46,327	—	46,327	—
Control Mean		0.000	0.000	—	—	0.000	—	0.000	—
Z-test p-value		0.018	0.673	—	—	0.745	—	0.005	—
Panel D: Phone-data-based treatment effects trained on the endline survey									
Treatment		0.014	0.008	-0.012	-0.001	0.012	0.004	0.011	0.010
		(0.012)	(0.013)	(0.015)	(0.014)	(0.012)	(0.012)	(0.012)	(0.012)
Obs.		46,327	46,327	46,327	46,327	46,327	46,327	46,327	46,327
Control Mean		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Z-test p-value		0.049	0.524	0.464	0.002	0.267	0.819	0.943	0.049

Notes: Replication of phone-data-based prediction and treatment effect estimation pipeline using two weeks of phone data to derive features (rather than six months). In the first regime — in which the ML models are trained using data from the impact evaluation survey — the two weeks of phone data for model training are obtained from the two weeks during which the pre-treatment survey took place in September 2021. The mobile phone data used to train the ML model in second regime — in which the ML models are trained using data from the endline survey — is taken from the two weeks immediately after each subscriber registered for GD-Novissi. The immediate post-treatment two weeks are used to generate well-being predictions in both regimes. Standard errors from Bayesian bootstrap procedure in parentheses. $^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01.

Table S7: Estimating Treatment Effects from Mobile Phone Data using Changes in Features

	(1) PMT	(2) Food security	(3) Financial health	(4) Financial inclusion	(5) Mental health	(6) Perceived status	(7) Healthcare access	(8) Labor supply	(9) All seven indices
Panel A: Predicting welfare outcomes									
R^2	0.036	0.005	0.004	0.004	-0.004	0.006	0.000	0.015	0.020
Obs.	8,446	9,504	9,504	9,131	9,504	9,504	9,519	9,504	9,504
Panel B: Phone-data-based treatment effects									
Treatment		0.000	-0.015	-0.007	0.025*	-0.011	0.005	-0.004	-0.005
		(0.018)	(0.019)	(0.016)	(0.017)	(0.018)	(0.019)	(0.015)	(0.014)
Obs.		48,726	48,726	48,726	48,726	48,726	48,726	48,726	48,726
Control Mean		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Z-test p-value		0.025	0.177	0.591	0.062	0.068	0.861	0.657	0.015

Notes: Replication of phone-data-based prediction and treatment effect estimation pipeline using changes in phone-derived features between the pre-treatment and during-treatment periods as inputs to the model (once for the six month time period, and once for the two week time period). Ground truth measures come from the endline survey. Standard errors from Bayesian bootstrap procedure in parentheses. $^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01.

Table S8: Estimating Treatment Effects from Mobile Phone Data in Savanes

	(1) PMT	(2) Food security	(3) Financial health	(4) Financial inclusion	(5) Mental health	(6) Perceived status	(7) Healthcare access	(8) Labor supply	(9) All seven indices
Panel A: Predicting welfare outcomes using ML trained on pre-treatment survey									
R^2	0.120	-0.012	0.009	—	—	0.042	—	0.038	—
Obs.	3,701	3,701	3,701	—	—	3,701	—	3,698	—
Panel B: Predicting welfare outcomes using ML trained on endline survey									
R^2	0.034	0.007	-0.003	-0.006	0.001	-0.002	0.001	0.016	0.023
Obs.	3,089	3,478	3,478	3,368	3,478	3,478	3,481	3,478	3,478
Panel C: Treatment effects from endline survey									
Treatment		0.089***	0.038	0.018	0.090***	0.056**	-0.010	0.012	0.078***
		(0.027)	(0.030)	(0.027)	(0.023)	(0.028)	(0.029)	(0.032)	(0.029)
Obs.		4902	4902	4902	4902	4902	4902	4902	4902
Control Mean		-0.01	0.02	-0.01	0.04	0.00	0.04	0.01	0.03
Panel D: Phone-based treatment effects trained on the pre-treatment survey									
Treatment		-0.002	-0.009	—	—	-0.011	—	-0.005	—
		(0.020)	(0.014)	—	—	(0.015)	—	(0.016)	—
Obs.		35,889	35,889	—	—	35,889	—	35,889	—
Control Mean		0.000	0.000	—	—	0.000	—	0.000	—
Z-test p-value		0.028	0.208	—	—	0.548	—	0.002	—
Panel E: Phone-based treatment effects trained on the endline survey									
Treatment		0.003	0.004	0.010	0.022	-0.008	0.000	0.013	0.008
		(0.018)	(0.018)	(0.024)	(0.020)	(0.022)	(0.017)	(0.017)	(0.016)
Obs.		35,889	35,889	35,889	35,889	35,889	35,889	35,889	35,889
Control Mean		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Z-test p-value		0.034	0.477	0.929	0.068	0.127	0.738	0.900	0.059

Notes: Replication of phone-data-based prediction and treatment effect estimation pipeline with only subscribers located in Savanes. Standard errors from Bayesian bootstrap procedure in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table S9: Estimating treatment status from mobile phone data

Phone Data Period	**AUC**	*N*
Panel A: All subscribers in RCT		
Six months	0.5151	49,079
Two weeks	0.5219	46,370
Panel B: Savanes only		
Six months	0.5254	36,086
Two weeks	0.5153	34,049

Notes: Predictive performance of a gradient boosting model for predicting treatment status from mobile phone records from the treatment period. Predictions are obtained over 5 fold cross validation, and the pooled area under the curve (AUC) score is reported.

Table S10: Variation Between and Within Canton

	(1)	(2)	(3)
Outcome	Between Variance	Within Variance	Ratio
Panel A: Pre-treatment survey			
PMT	4.30	0.29	15.00
Food Security	1.438	0.96	1.44
Financial Health	1.25	0.98	1.27
Perceived Socioeconomic Status	1.80	0.95	1.89
Labor Supply	1.58	0.94	1.67
Panel B: Endline survey			
PMT	2.08	0.25	8.21
Food Security	1.85	0.99	1.87
Financial Health	1.65	0.99	1.87
Financial Inclusion	1.59	0.92	1.72
Mental Health	1.89	0.98	1.92
Perceived Socioeconomic Status	1.26	0.99	1.28
Healthcare Access	1.46	1.00	1.45
Labor Supply	1.57	0.96	1.63
Aggregate Welfare Index	2.03	0.98	2.08

Notes: Between vs. within variance, with groups defined by canton (self-reported in the baseline survey, determined by location of Novissi registration in the endline survey). Only individuals in cantons with at least 10 individuals surveyed are included in the analysis. All outcomes except for the PMT are standardized to 0 mean and unit variance in the control group.

Table S11: Feature importances in a machine learning model including mobile money data

Feature	**Importance**
Maximum amount of transactions in category “other”	281
Maximum balance before outgoing transactions	246
Mean balance before outgoing transactions	211
Maximum balance before outgoing transactions in category “other”	172
Mean amount of transactions in category “other”	145
Number of outgoing transactions	140
Number of outgoing transactions in category “other”	123
Mean balance before outgoing transactions in category “other”	114
Mean balance after outoging transactions in category “other”	105
Maximum balance before outgoing transactions in category “other”	102

Notes: Feature importances for machine learning model predicting treatment status from mobile phone data *including data on mobile money transactions* using six months of phone data from during the treatment period (see Section 5.1). Feature importances are derived from the gradient boosting model as the total number of times a feature is split upon in the entire ensemble of regression trees. Only the top 10 most important features are shown.

Table S12: Registration with Savanes-Novissi

	(1) Food security	(2) Financial health	(3) Financial inclusion	(4) Mental health	(5) Perceived status	(6) Health care access	(7) Labor supply	(8) **All seven** **indices**
T, non-SN	0.148***	0.063	0.041	0.091***	0.110***	0.023	0.017	0.131***
	(0.037)	(0.042)	(0.034)	(0.030)	(0.037)	(0.038)	(0.044)	(0.039)
T, SN	0.039	-0.036	0.053	0.070*	0.074*	-0.010	-0.049	0.037
	(0.040)	(0.045)	(0.039)	(0.036)	(0.040)	(0.045)	(0.051)	(0.044)
C, non-SN	0.042	-0.023	0.038	-0.017	0.085**	0.052	-0.036	0.038
	(0.038)	(0.044)	(0.039)	(0.033)	(0.040)	(0.038)	(0.046)	(0.041)
C, SN Mean	-0.01	0.02	-0.01	0.04	0.00	0.04	0.01	0.03
F-test 1-2	7.43***	5.55**	0.10	0.39	0.85	0.57	2.39	5.37**
Obs.	4,755	4,755	4,755	4,755	4,755	4,755	4,755	4,755

Notes: Results for regressing the main survey outcomes on the interaction of GD-Novissi treatment status and Savanes-Novissi registration status. *T* indicates treatment, *C* indicates control, and *SN* and *non-SN* indicates beneficiaries and non-beneficiaries Savanes-Novissi, respectively. *F-test 1-2* row provides the p-value of the statistical comparison of the coefficients for “Treatment, not Savanes-Novissi” and “Treatment, Savanes-Novissi”. All regressions control for the enumerator, week of the survey, and strata fixed effects. All observations are weighted by sampling probabilities. Robust standard errors are in parentheses. *p <0.1; **p <0.05; ***p <0.01.