A Comment on "Resisting Social Pressure in the Household Using Mobile Money: Experimental Evidence on Microenterprise Investment in Uganda"*

Lenka Fiala, Jack Fitzgerald, Essi Kujansuu, and David Valenta

November 18, 2024

Abstract

In a pre-registered experiment, Riley (2024) finds that providing microcredit loans onto mobile money accounts yields significantly more profit and capital for women's businesses than providing loans in cash, as this disbursement technique permits women to resist family pressure to share loans. We uncover two credibility issues. First, we find evidence suggesting that most of the experiment's participants are not assigned to treatment using the pre-registered stratified randomization protocol described in the paper. Second, the reported variables and empirical methods contradict commitments in the paper's pre-registration; these contradictions are unacknowledged and meaningfully impact the paper's main findings.

Keywords: microcredit, entrepreneurship, gender, field experiment, stratified randomization, pre-registration, pre-analysis plan
JEL codes: C93, D13, G51, J16, O12, O16

*Authors: Fiala: University of Bergen, lenka.fiala@uib.no, corresponding author. Fitzgerald: Vrije Universiteit Amsterdam, j.f.fitzgerald@vu.nl. Kujansuu: University of Innsbruck, essi.kujansuu@uibk.ac.at. Valenta: University of Ottawa, dvalenta@uottawa.ca. We have no conflict of interest to report. We are grateful for the invaluable work of several researchers who choose to remain anonymous. We also thank Emma Riley for responding to data concerns, as well as Abel Brodeur and Derek Mikola for organizational support and facilitating communication. This replication project has ethical approval from Vrije Universiteit Amsterdam's School of Business and Economics. Replication code and instructions for downloading replication data can be found at this link.

1 Introduction

Microcredit is a critical source of financing for small-scale entrepreneurs throughout the developing world, with microcredit institutions reaching more than 140 million low-income clients in 2019 (Fassin & Valette 2020; Meager 2022). Empirical evaluations of microcredit typically find that microcredit access has strong impacts on poverty alleviation, women's empowerment, and firm investment into assets (see Khandker 2005; Pronyk et al. 2006; Kim et al. 2007; Banerjee et al. 2015a; Crépon et al. 2015). However, microcredit access has (at best) modest impacts on firm profitability (e.g., see Morduch 1999; Cull et al. 2007; Banerjee et al. 2015b; Meager 2019; Meager 2022). This latter finding has important implications for the financial sustainability of microcredit institutions, and for the priority which governments and nonprofits should assign to funding microcredit initiatives if such funding bodies seek to stimulate economic development.

Riley (2024b) postulates that a key reason for the modesty of microcredit's success in improving firm profits is that microcredit loan recipients face pressure to share loans with family, and estimates the impacts of treatments that can potentially mitigate this pressure. Specifically, Riley (2024b) reports the results of a pre-registered field experiment conducted with several thousand entrepreneurial women in Uganda who are approved to receive loans from BRAC, a microcredit institution. Whereas the control group receives their loans in cash without any additional intervention, two treatment groups are given additional interventions related to mobile money accounts. The "mobile account" treatment gives women a SIM card that provides women with access to a mobile money account. The "mobile disbursement" treatment both provides women with a mobile money account and directly disburses their loan amount onto that account.

Riley's (2024b) pre-registration commits to assigning these treatments using a stratified randomization strategy (see Riley 2018). Within a given sampling block, the pre-registered randomization protocol commits to assign women to treatment within strata that match on five baseline characteristics. The published version of Riley (2024b) implies that the experimental protocol follows through on these pre-registered commitments.

Riley (2024b) concludes that the mobile disbursement treatment is particularly successful at improving outcomes for treated women's businesses, and attributes this higher success to women's increased capacity to resist family pressure to share loan funds. In Table 1, Riley (2024b) finds that compared to the control group, the businesses of women in the mobile disbursement treatment group have significantly higher levels of self-reported profit and capital investment. In contrast, the differences in these outcomes between the businesses of women in the mobile account treatment group and the control group are not statistically significantly different from zero. Table 2 shows estimates of interaction effects between the experimental treatments and an index of family pressure. Women who face high family pressure exhibit significantly higher treatment effects on business profits and capital for the mobile disbursement treatment, but these interaction effect estimates for the mobile account treatment are not statistically significantly different from zero.

This comment critically re-examines this paper's replication data (Riley 2024a), and reveals several problems with the credibility of the published paper (Riley 2024b). First, we find evidence suggesting that much of the data is not assigned to treatment using the stratified randomization strategy described in the paper. The majority of strata exhibit mismatch on pre-registered stratification variables, and the majority of women in the sample are assigned to these mismatched strata. It is unclear how treatment is assigned for women assigned to the mismatched strata.

Further, we find evidence that for a substantial proportion of participants, at least one of the claimed stratification variables could not have been collected as described in the paper. In particular, Riley (2024b) claims that participants are stratified based on whether they hide money from their spouse more than the median participant in an incentivized game. However, more than a third of participants are unmarried at baseline. Of course, it is impossible to honestly incentivize a 'hiding money from the spouse' game for women with no spouse. Accordingly, all participants who are unmarried at baseline have missing values in a variable that captures results from the money-hiding game, implying that these participants never play the money-hiding game. Women who are unmarried at baseline appear to be assigned to treatment without being stratified by high money-hiding status, as claimed in the paper and pre-registration. For other issues with data collection for stratification variables, see Section 2.2.

Finally, we find that Riley's (2024b) analytical choices deviate from pre-registered protocols in undisclosed ways that affect the paper's main conclusions. Most notably, Riley (2024b) lumps inventory values together with business assets – both of which are pre-registered outcome variables – to create a variable called 'capital', which is not pre-registered. The tables in the main text of Riley (2024b) exclusively report treatment effects on capital, and do not report treatment effects on inventory values or business assets. However, robust support for Riley's (2024b) key mechanism is eliminated after splitting capital into its original pre-registered components.

Section 2 provides evidence that the stratified randomization process used to assign treatment meaningfully differs from the protocol described in the pre-analysis plan and the paper, and discusses data collection issues with several stratification variables in depth. Section 3 documents important deviations from the pre-analysis plan in the published paper's reported variables and analyses. Section 4 concludes.

2 Stratified Randomization and Stratification Variables

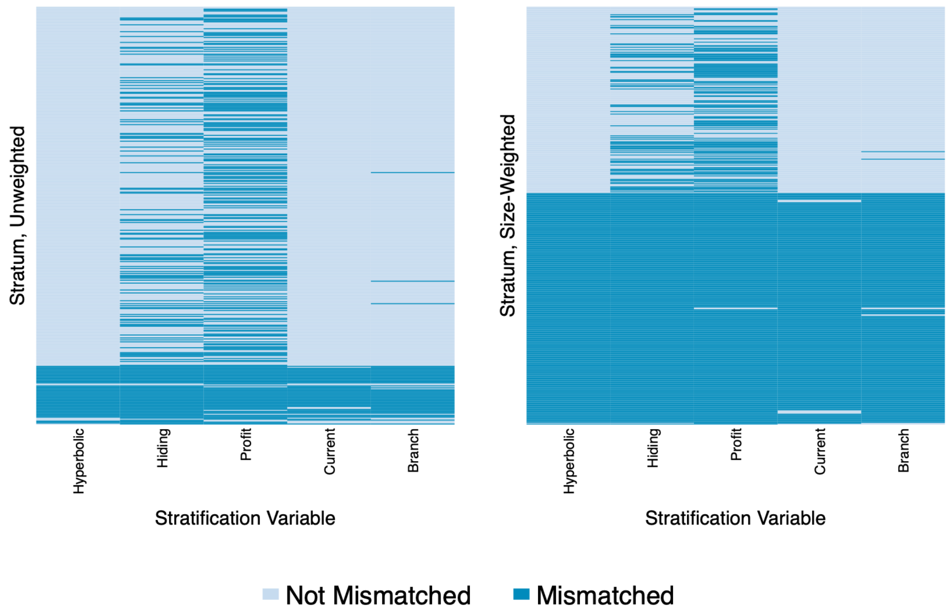
We find evidence that the data was not entirely collected using stratified randomization, as claimed in the paper. From pg. 1422:

> "The study included 3,000 female microentrepreneurs, assigned using an individual stratified randomization as follows: 1,000 to the control group, 1,000 to the Mobile Account arm, and 1,000 to the Mobile Disbursement arm. Randomization took place weekly in batches of 250 women determined by the timing of requesting a new loan. Within each batch, all women accepted for a loan were individually randomized into the treatment or control groups. Randomization continued for approximately 5 months until the sample size of 3,000 was achieved. Lists of treatment assignments from the randomization were sent to the BRAC branches weekly, and only women who had completed the baseline survey and had been assigned a treatment could have a loan disbursed to them.... The randomization was stratified by five variables at baseline: a dummy variable capturing present bias from a multiple price list incentivized game... a dummy variable capturing if the woman switched above the median in a willingness to pay to hide money from the spouse game... a dummy variable capturing if the client is a first-time

> borrower with BRAC, the microcredit branch, and a dummy variable for above-median business profit."

This randomization protocol is a commitment in Riley's pre-registration in the AEA RCT Registry, specifically documented using the repository file `PAP+mobile+money+BRAC+RILEY.pdf` (Riley 2018). To our knowledge, the stratification variables are `hyperbolic_base`, `above_m_median_base`, `current_loan_base`, `branch_name`, and `high_profits_base` in dataset `survey_data.dta`. These five variables are all discrete – in fact, all are binary except `branch_name`, which takes on one of six values depending on the BRAC branch from which the loan was disbursed.

Though each of the five stratification variables should take on the same value within each stratum if treatment was uniformly assigned using stratified randomization, strata are frequently mismatched on these variables. Figure 1 displays the extent of the stratification failure. Of the 427 strata containing individuals who consented to the endline survey (and therefore are used to compute the main estimates), 291 are mismatched on at least one stratification variable. This includes all but one of the 60 strata numbered 384 or higher. More than 58% of the participants who consented at endline belong to one of these strata numbered 384 or higher; their higher weight can be observed in the dark blue region of the right-hand graph in Figure 1. In total, over 87% of participants are part of a stratum that is mismatched on at least one claimed stratification variable. This raises



Note: In the left-hand graph, each row of the plot represents a single stratum. In the right-hand graph, each row of the plot represents a single observation, ordered by stratum number in the same fashion as the left-hand graph. Dark blue shading indicates that the variable specified by the column does not take the same value for all observations in the stratum represented by the row. "Hyperbolic" refers to a hyperbolic preferences dummy. "Hiding" indicates whether, in the game where money was to be hidden from spouses, the woman chooses to hide money from the spouse more than four times (i.e., whether she is 'above median' in hiding money from the spouse). "Profit" indicates whether the woman makes $85 or more in self-reported business profit at baseline (i.e., whether she is 'above median' in baseline self-reported business profit). "Current" denotes whether the woman is a pre-existing BRAC customer at baseline. "Branch" indicates the BRAC branch from which the woman receives her loan disbursement. The graphs are constructed using the `panelview` command in Stata (Mou, Liu, & Xu 2023).

Figure 1: Stratification Failure

concerns about the data-generating process for treatment assignment. In the remainder of this section, we discuss the unique features of this stratification failure, which provide important details about data collection that are not discussed in Riley (2024b).

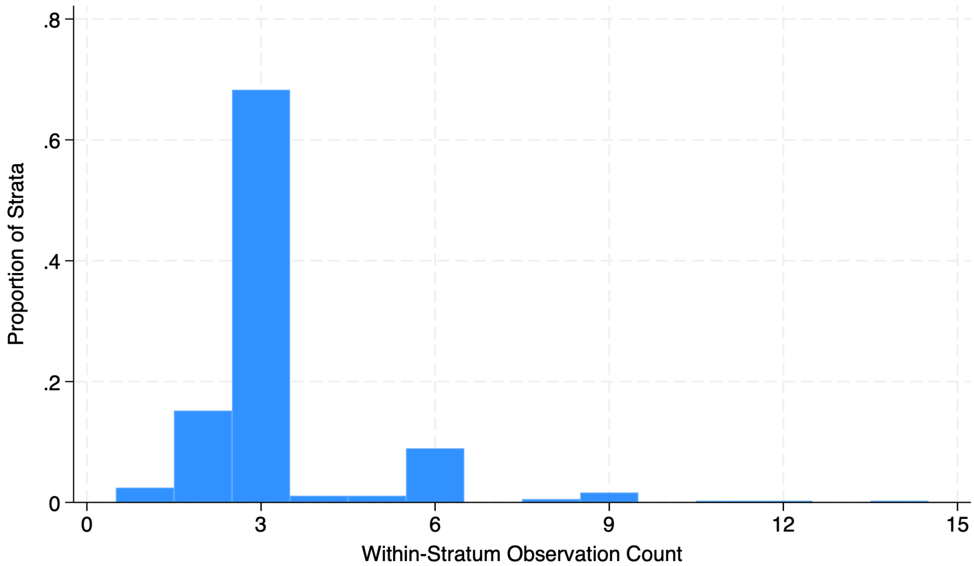
2.1 Fallback Strata

Each row in Figure 1's left-hand graph represents a stratum, which are sorted by the stored stratum number in `survey_data.dta`. The graph shows that all strata are perfectly balanced on `hyperbolic_base` up until stratum 384, which is numerically the first stratum wherein there is imbalance on `hyperbolic_base`. The dark blue region runs from stratum 384 onwards, where the vast majority of strata exhibit imbalance on at least some, and often all, stratification variables.

These graphs raise natural questions about the strata at and beyond stratum 384. Why are these strata so much more imbalanced on all stratification variables than the prior strata? And looking to the corresponding size-weighted data in the right-hand graph, why are so many more women assigned to these 60 problematic strata than are assigned to the other 367 more well-behaved strata?

The data is most consistent with the following explanation. A genuine effort appears to have been made to stratify on (at least some of) the pre-registered stratification variables. This stratified randomization takes place in 'triads' of three women who match on (most) stratification variables at baseline and could be conveniently split into each respective treatment assignment. For each triad, one woman in the triad is placed into the control group, one is placed into the mobile account treatment, and one is placed into the mobile disbursement treatment. Strata prior to stratum 384 constitute (groups of) successfully-matched triads. However, women who do not conveniently fit into a matched triad when treatment is assigned are instead assigned to a 'fallback stratum'. Treatment is not assigned using stratified randomization in these fallback strata. Strata at and above stratum 384 are fallback strata.

Though such a procedure is not described in Riley (2024b), this explanation of the treatment assignment process is consistent with the following empirical facts. First, strata prior to stratum 384 have observation counts that are multiples of three far more frequently than would be expected if strata are not sampled with any systematic divisibility target. Figure 2 shows the distribution of observation counts by stratum for strata prior to stratum 384. The histogram shows clear spikes in observation count density at multiples of three. The median and modal stratum size in strata prior to stratum 384 is three, with 68% of such strata having exactly three observations. Another spike appears at a stratum size of six, comprising around 9% of such strata. Around 79% of strata prior to stratum 384 have observation counts that are multiples of three, far more than would be expected by chance. In contrast, only around 37% of strata at and after stratum 384 have observation counts that are multiples of three, much closer to the 33% that would be expected by chance. These patterns are consistent with non-fallback strata representing groups of triads, with the

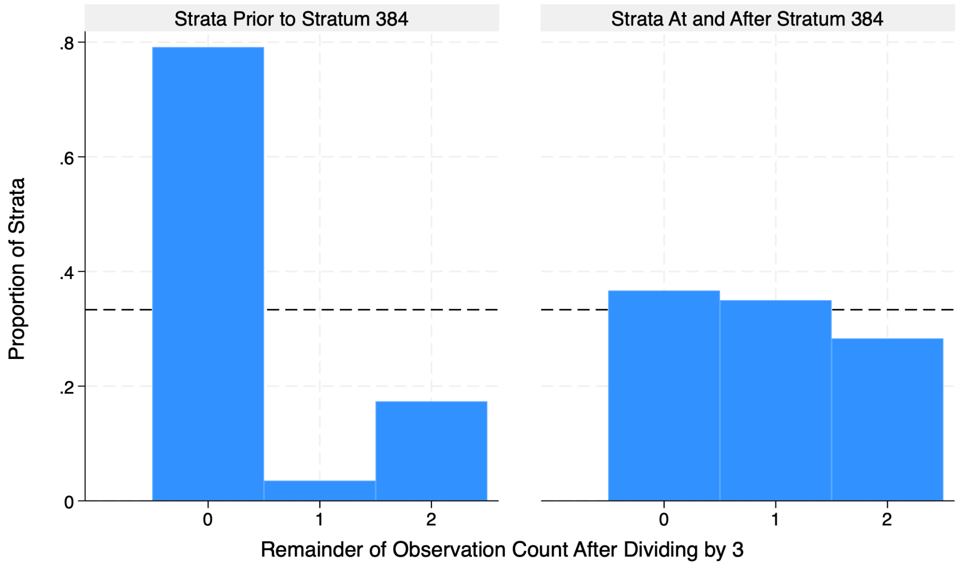


Note: The histogram displays the density of strata by number of observations for strata prior to stratum 384.

Figure 2: Observation Counts in Strata Prior to Stratum 384

vast majority being groups of one or two triads, and with fallback strata being sampled without any clear divisibility target.

Strata at and after stratum 384 do not exhibit similar jumps in triad-consistent observation counts. Figure 3 shows the frequency of strata by observation count modulo with respect to three. Strata for whom $\texttt{mod}(N, 3) = 0$ have observation counts that are multiples of three, whereas nonzero modulo values represent the remainder of the stratum's observation count if divided by three. As is visible in Figure 3's left-hand histogram, for strata prior to stratum 384, stratum-level observation counts that are multiples of three are over 22 times more common than such counts where $\texttt{mod}(N, 3) = 1$, and over 4.5 times more common than such counts where $\texttt{mod}(N, 3) = 2$. The right-hand graph makes clear that no similar jumps



Note: The histogram denotes the proportion of strata by levels of `mod`$(N, 3)$, where N is the within-stratum observation count. The dashed horizontal line is located at $\frac{1}{3}$.

Figure 3: Modulo Analysis

occur in strata at and after stratum 384.

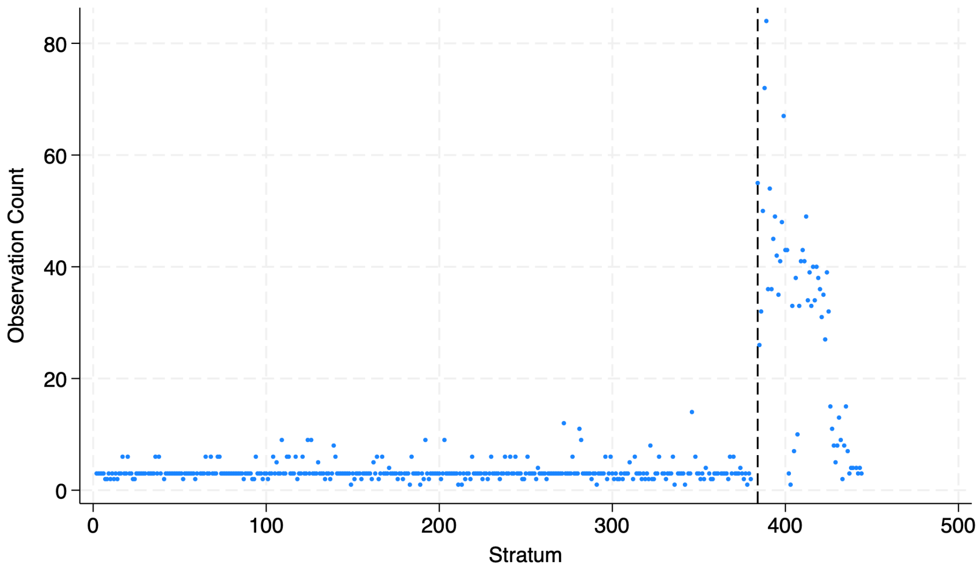
Triads are also split perfectly evenly between treatment conditions in the vast majority of strata prior to stratum 384, but this is not the case for strata at and after stratum 384. In nearly 79% of the strata prior to stratum 384, each of the three treatment conditions has exactly the same number of observations. This is consistent with a stratified sampling strategy that attempts to form strata from one or more triads of women, where one woman in each triad is assigned to each of the three treatments. In contrast, fewer than 7% of strata at and after stratum 384 have the same number of observations in all three treatment conditions, which suggests that these strata are formed using a different sampling strategy.

Further, the sizes of the strata are consistent with expected failures of the experiment's ambitious stratification strategy. As aforementioned, Riley (2024b) commits to stratifying participants by four binary baseline variables and by BRAC branch. Some values of the baseline variables are also quite rare. Specifically, only about 20% of the women in Riley's (2024b) sample exhibit hyperbolic discounting, and fewer than 18% are new BRAC customers at the time that they receive their loan. At a given BRAC branch in a given sampling period, it is unsurprising that many participants do not fit perfectly into triads who completely match on all four baseline stratification variables. If these 'leftover' participants are to continue participating in the experiment, then their treatments must be assigned using a different strategy.

In Riley (2024b), strata prior to stratum 384 are much larger than strata at and after stratum 384. Figure 4 shows observation counts by stratum. The dashed vertical line marks stratum 384, where there is a clear upward jump in stratum size. The average stratum at and after stratum 384 has over 25 more women than the average stratum prior to stratum 384 ($t = 10.16$), constituting an average within-stratum size jump of 779%.

Finally, the timing of data collection is also consistent with strata at and after stratum 384 being sampled using an alternative sampling strategy. Figure 5 shows the start date of each participant's baseline survey and the stratum to which they are assigned.[1] The dashed horizontal line marks

[1]This reflects the sampling timeframe relevant to the stratification strategy because nearly all women receive their loan disbursement within two weeks of completing the

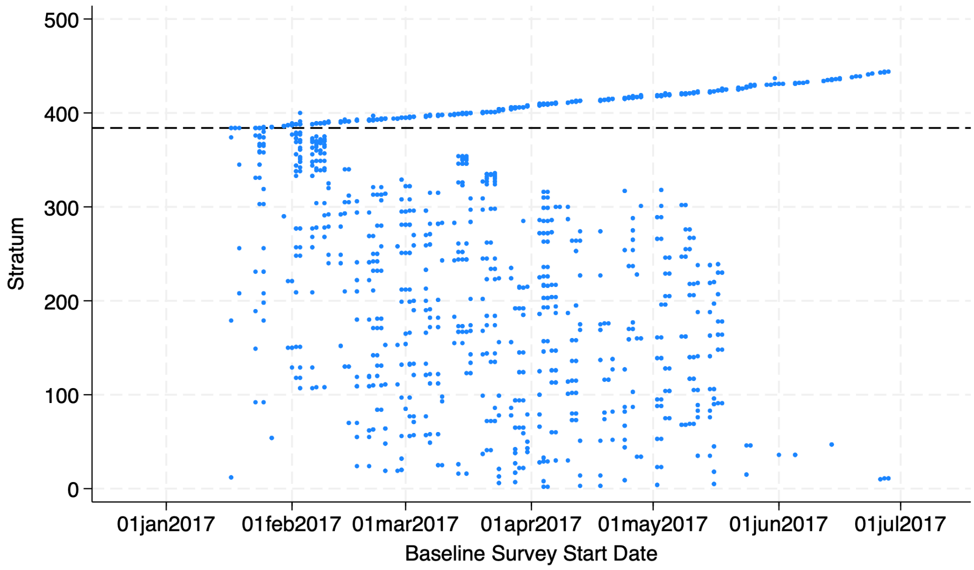


Note: Within-stratum observation counts are presented for each stratum. The dashed vertical line denotes stratum 384.

Figure 4: Stratum Sizes

stratum 384. The strata at and after stratum 384 form a distinct contour on this graph. Whereas the strata prior to stratum 384 are constructed quickly and sporadically with no clear chronological order, the strata at and after stratum 384 are sampled over longer periods of time, are consistently constructed throughout the duration of baseline data collection, and are relatively well-ordered. The upward slope of this contour is consistent with fallback strata absorbing leftover participants who could not be successfully matched on all stratification variables, and continuing to absorb participants until a sampling period ends and/or the fallback stratum becomes too overcrowded, at which point new (higher-numbered) fallback strata are opened to continue absorbing leftover participants. If this is in-

baseline survey (see also Section 2.2).



Note: Start dates of the baseline survey are presented at the observation level, along with the stratum to which each observation belongs. The dashed horizontal line denotes stratum 384. Respondents with start dates prior to 2017 are dismissed as device errors and dropped; this drop affects only 0.61% of the sample.

Figure 5: Timing of Treatment Assignment by Stratum

deed how participants in strata at and after stratum 384 are sampled, then the majority of this experiment's participants are sampled using a strategy that contradicts pre-registration commitments and is not disclosed in the published paper.

2.2 Stratification on Baseline Profit Medians

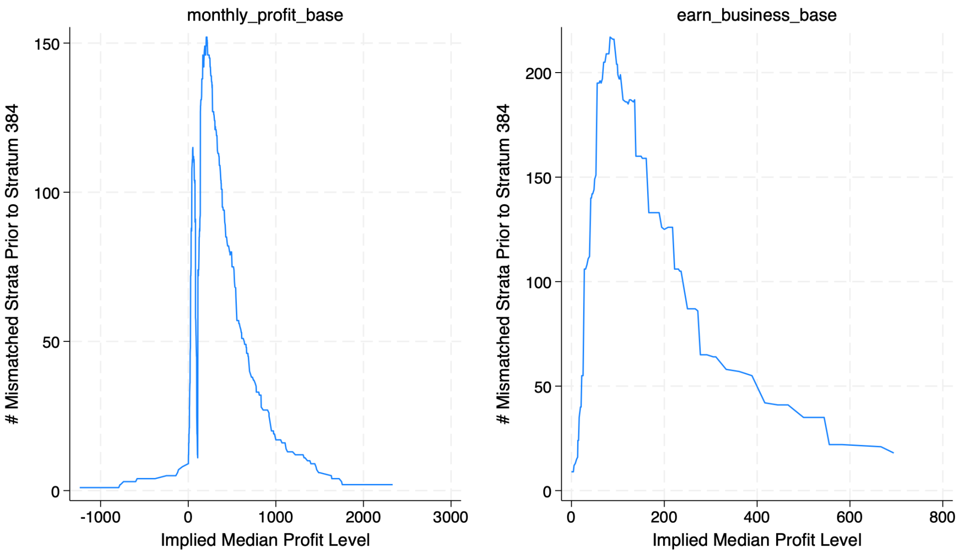
Returning to Figure 1, it is clear that even strata prior to stratum 384 are better-matched on some stratification variables than others. Over 51% of all strata prior to stratum 384 contain some women that have 'above-median' baseline business profits and some women that have 'below-median' baseline business profits. Over 27% of strata prior to stratum 384 similarly contain both above-median and below-median women in terms of hiding money from spouses (in an incentivized game; see Section 2.3 for details). In contrast, this proportion does not exceed 1% for any other stratification variable. Why are some variables so much better-behaved than others?

One factor that may be driving these mismatches is that given Riley's (2024b) sampling procedure, it is not clear how the median baseline business profit is computed prior to all (or even most) baseline survey data being collected. There is no mention of piloting in the published paper, nor in any documentation underlying this experiment (e.g., in pre-registrations, amendments thereto, etc). The baseline sampling period takes place from January-June 2017, and around 97% of respondents receive their loan disbursement within two weeks of starting the baseline survey. This implies that it is not possible to know the median baseline business profits of all women in the sample at the time that most women in the sample are randomized into treatment.

We begin by examining the cutoff that Riley (2024b) sets in the data. Variable `high_profits_base` perfectly splits baseline self-reported profit variable `earn_business_base`; all observations with `high_profits_base` = 1 (= 0) exhibit `earn_business_base` above (below) 85. `high_profits_base` also perfectly divides `earn_business_base` at its median of 83.33333.

However, given that this division yields mismatches in over half of strata prior to stratum 384, we examine whether there are other divisions in the baseline profit data that would better match women within each stratum. Riley (2024b) has many different variables for baseline profits, with differing degrees of winsorization, functional form, temporal range (i.e., weekly vs. monthly profits), and elicitation strategy (i.e., self-reported vs. computed/inferred). Within each variable, we search for 'implied medians' that would yield better strata matching. Specifically, for each variable j, we loop over all values k of variable j and examine how many mismatched strata prior to stratum 384 arise if we assume that stratification is conducted by matching observations i based on whether observation i's value of variable j is greater than k. Naturally, for the purposes of this exercise, fewer mismatched strata is better.

This exercise reveals that the effective threshold used for stratifying women into above-median and below-median baseline business profit groupings is likely incorrectly coded in Riley's (2024b) data. Figure 6 shows the results of the exercise for two baseline business profit variables, one of which sticks out as a strong candidate for the true stratification variable: `monthly_profit_base`. This is because `monthly_profit_base` exhibits a pronounced dip in mismatched strata, which is minimized when the implied median is assigned as 108.3333. This value is the largest observed value that is less than the actual median of `monthly_profit_base`, which is 111.1111. `earn_business_base` does not exhibit a similar dip. In Appendix Figure A1, we verify that the same is true of a wide range of other



Note: Number of strata prior to stratum 384 that are mismatched on an 'above median baseline business profits' indicator, where the 'implied median' is plotted on each x-axis in units of the baseline business profits variable listed in each graph's title.

Figure 6: Mismatched Strata by Implied Median Across 'Profit' Variables

alternative baseline business profit variables that could have been used for stratification. This implies that stratification likely occurs by balancing observations based on whether they are above-median or below-median on `monthly_profit_base`. However, Riley's (2024b) indicator for above-median baseline business profits (`high_profits_base`) does not meaningfully divide `monthly_profit_base`, implying that this stratification variable is incorrectly coded.

Resolving this coding error still does not eliminate within-strata mismatches on the 'high baseline profits' indicator. Even when using the best available 'implied median' for `monthly_profit_base`, 11 strata prior to stratum 384 are still mismatched on the 'high baseline profits' indicator. It remains unclear why these stratification failures still occur.

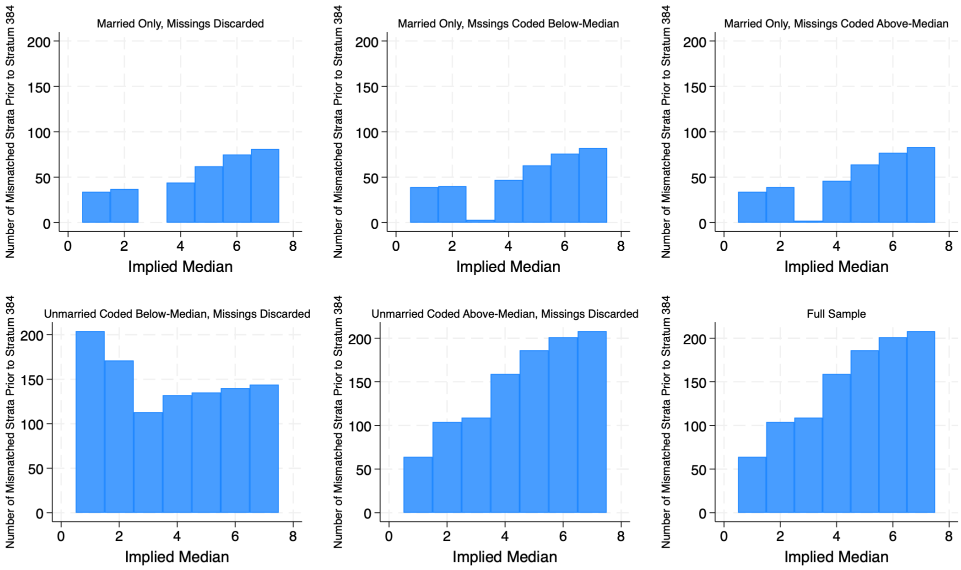
2.3 Hiding Money from the Spouse and Unmarried Participants

As discussed in Section 2.2, over 27% of strata prior to stratum 384 contain both some women who are above-median and some who are below-median in terms of hiding money from their spouses. This measure is based on an incentivized multiple price list game wherein women are asked to make eight choices, either receiving 8000 Ugandan shillings themselves or sending another payment from 7200 to 36,000 shillings to their spouse (see Riley 2024b, Appendix Table A1). Seven of the eight choices in this game yield more money for the spousal pair if the money is sent to the spouse, so each woman's desire to hide money from her spouse is measured by how many times the woman decides to receive the money herself rather than send money to her spouse.

The above-median money-hiding indicator does not accurately split the money-hiding variable at its median. Riley (2024b) appears to store the number of choices made to send money to the spouse under variable `switch_m_base` and the above-median indicator as `above_m_median_base`. `switch_m_base` does completely split `above_m_median_base`: observations with `above_m_median_base` = 0 exhibit values of `switch_m_base` between one and three, and those with `above_m_median_base` = 1 exhibit `switch_m_base` values between four and eight. However, the median value of `switch_m_base` is five, in the middle of the range of `switch_m_base` values for observations with `above_m_median_base` = 1. This suggests the potential that an arbitrary value of `switch_m_base` is selected to divide the sample. This is plausible, considering that the number of choices to give money to the spouse takes on nine levels (zero through eight), and four is thus the midpoint, meaning that Riley's (2024b) current division is effectively indicating whether an observation is at or above the midpoint. It is thus possible that Riley (2024b) is using the term 'median' loosely.

An even larger problem emerges for this stratification strategy – which depends on a 'hiding money from the spouse' game – because 34% of Riley's (2024b) sample is unmarried at baseline. Naturally, it is not possible to incentivize a 'hiding money from the spouse' game for women who have no spouse. It appears that unmarried women never play the money-hiding game; all women who are unmarried at baseline have missing values for `switch_m_base`. However, no unmarried woman has a missing value for `above_m_median_base`; the above-median indicator is uniformly imputed to zero for all women who are unmarried at baseline. There are also 20 women who are married at baseline, yet have missing values for `switch_m_base`. It is unclear if these women play the money-hiding game, as none of these 20 women have missing values for `above_m_median_base`, and `above_m_median_base` is not equivalent for these 20 women; six of these women are assigned `above_m_median_base` = 1 and 14 are assigned above_m_median_base $= 0$.

Within-stratum mismatches on the above_m_median_base indicator are entirely driven by unmarried women, as well as by married women with missing values for switch_m_base, with the former driving the bulk this of mismatching. To show this, we entertain the notion that the within-stratum mismatches on above_m_median_base are caused by an incorrectly-coded threshold and repeat the exercise from Section 2.2 where we search for 'implied medians' that would yield less within-stratum mismatching across several subsamples of Riley's (2024b) data. Figure 7 displays the results of this exercise. The upper left-hand graph restricts the sample exclusively to married women with defined values of switch_m_base, and reveals that in this subsample, an implied median of three – the same switch point that perfectly splits switch_m_base by values of above_m_median_base – completely eliminates within-stratum mismatches on an above-median money-hiding indicator. The two upper graphs to the center and right show that adding married women with missing values of switch_m_base yields slightly more mismatching at this switch point, but only for a few strata. The bottom three graphs in Figure 7 show that regardless of how unmarried women are coded, and regardless of how married women with missing values of switch_m_base are coded, adding unmarried women back to the sample dramatically increases the extent of within-stratum mismatches on any above-median money-hiding indicator. This confirms that the within-



Note: Number of strata prior to stratum 384 that are mismatched on an 'above-median money-hiding' indicator, where the 'implied median' is plotted on each x-axis. Each of the six graphs displays the results for a different subsample of Riley's (2024b) data.

Figure 7: Mismatched Strata by Implied Median of Money-Hiding for Different Subsamples

stratum mismatches on `above_m_median_base` shown in the second column of each graph in Figure 1 are primarily driven by unmarried women, who (as aforementioned) likely do not play the money-hiding game.

This suggests the possibility that during stratification, unmarried women are treated as 'wild cards' who can be freely sorted either into high-money-hiding or low-money-hiding strata. As aforementioned, the stratification variable `above_m_median_base` = 0 for all unmarried women. If all unmarried women are assigned to low-money-hiding strata, then there should be no strata prior to stratum 384 mismatched on `above_m_median_base`. The within-stratum mismatches on `above_m_median_base` that we observe arise because many unmarried women are assigned to high-money-hiding strata. This problem is not fixed by supposing that unmarried women 'should' be assigned to high-money-hiding strata. The bottom left and bottom center graphs in Figure 7 show that coding unmarried women as high-money-hiding women barely reduces the degree of within-stratum mismatch on a dummy variable indicating whether `switch_m_base` > 3. This implies that many unmarried women are indiscriminately assigned to both low-money-hiding and high-money-hiding strata, and are only assigned `above_m_median_base` $= 0$ *ex post*. Given the fact that married women with missing values for `switch_m_base` are found both in high-money-hiding and low-money-hiding strata, it is possible that these women are also treated as 'wild cards' that can be assigned to either high-money-hiding or low-money-hiding strata without consequence.

3 Analytical Deviations from the Pre-registration

Riley's (2024b) pre-analysis plan exists in two parts, both of which are included in the repository for the paper's AEA RCT Registry pre-registration (Riley 2018). `PAP+mobile+money+BRAC+RILEY.pdf` contains the first part of the pre-analysis plan. This portion of the pre-analysis plan was added to the pre-registration repository in December 2017, after the end of baseline data collection. This first part of the pre-analysis plan concerns analyses of primary outcomes (such as business performance, savings, and investment), as well as additional secondary outcomes for which analyses are supposed to be exploratory. The second part of the pre-analysis plan is contained in `mobile-money-pap-amendment-riley.pdf`. This part of the pre-analysis plan is published in July 2018 after data is obtained from the mobile money account provider MTN, and concerns 'intermediate' outcomes related to account usage.

We do not comment on deviations from the pre-analysis plan that are acknowledged in the paper. We also refrain from commenting on exploratory analyses that are confined to the Online Appendix. Additionally, we omit any further discussion of the ways in which the stratified randomization procedure employed in practice does not conform with that in the pre-registration; for a detailed discussion, see Section 2.

3.1 Investment Behavior: Capital, Inventory, and Business Assets

The most consequential pre-registration deviation is that one of the paper's primary business performance outcomes does not match the pre-registration. Tables 1 and 2 in Riley (2024b) display estimated (heterogeneous) treatment effects on `capital`. This variable is the sum of `inventory_value` and `ent_asset_value`, which represent business inventory values and business asset values (respectively). However, though `inventory_value` and `ent_asset_value` both appear in the list of pre-registered outcomes,[2] `capital` does not appear. The amendment to the pre-analysis

[2]"Value of business assets" is pre-registered as the summary outcome for the third main outcome family. In contrast, inventory is pre-registered as one sub-outcome of the

plan also does not discuss this change.

The `capital` outcome used in the published paper produces much stronger main treatment effect estimates than the pre-registered `inventory_value` outcome. Table 1 shows Riley's (2024b) main treatment effect estimates with `capital`, `inventory_value`, and `ent_asset_value` as dependent variables. Following Riley (2024b), Benjamini, Krieger, & Yekutieli (2006) q-values are displayed in curled brackets. For each outcome, q-values are computed under a 'what-if' scenario where we observe what the q-values would have looked like had Riley (2024b) used the outcome variable specified by the column instead of `capital`. Model 1 in our Table 1 directly replicates the q-values (and other results) from Model 3 of Table 1 in Riley (2024b), whereas the q-values in Models 2-3 in our Table 1 are isolated from alternate full replications of Table 1 in Riley (2024b) where `capital` is replaced with `inventory_value` and `ent_asset_value` (respectively).

The pre-registered `inventory_value` outcome produces estimates for the mobile disbursement treatment that are considerably smaller than those for `capital` and for `ent_asset_value`. This is not simply a matter of scale; compared to the control mean, the mobile disbursement treatment effect on inventory is 6.4%, over one third smaller than that for capital (10.2%) and over two thirds smaller than that for business assets (20.6%). This

first primary outcome family, belonging under "business performance."

Table 1: Main Treatment Effects on Investment Variables

	(1) Capital	(2) Inventory Value	(3) Business Assets
Mobile account	13.477	-0.424	13.820
	(24.180)	(19.991)	(12.478)
	{1.000}	{1.000}	{1.000}
Mobile disburse	69.207	32.338	36.869
	(23.868)	(19.617)	(12.082)
	{0.010}	{0.273}	{0.006}
Observations	2,639	2,638	2,610
R-squared	0.512	0.468	0.415
Control mean	678.28	501.71	178.8

Note: Intent-to-treat estimates. All outcomes are winsorized at the 99% level in USD. All regressions include strata dummies and include the baseline value of the outcome. 'Mobile account' is the treatment where only a mobile money account was provided and the loan was disbursed as cash. 'Mobile disburse' is the treatment where a mobile money account was provided and the loan also disbursed onto this account. Capital is the value of all assets the woman uses in her business plus the value of inventory held for her business. Benjamini, Krieger, & Yekutieli (2006) q-values are displayed in curled brackets, computed under the 'what-if' scenario where in Riley's (2024b) Table 1, capital was replaced with the outcome variable specified by the column. Robust standard errors in parentheses. (Text copied and modified from Riley 2024b).

treatment effect estimate for `inventory_value` is not robustly statistically significant ($q = 0.273$).

Though one could look at the results from our Table 1 as evidence confirming that the mobile disbursement treatment successfully promotes investment behavior into business assets by decreasing family sharing pressure, this mechanism is not robust once the new `capital` outcome is split into its original pre-registered components. Our Table 2 shows the heterogeneous treatment effect estimates in Riley's (2024b) Table 2 with `capital`, `inventory_value`, and `ent_asset_value` as dependent variables. As in our Table 1, our Table 2 computes q-values under the 'what-if' scenario where we observe what q-values would look like if `capital` in Riley's (2024b) Table 2 were replaced by the column dependent variable; Model 1 in our Table 2 is a direct replication of Model 3 in Riley's (2024b) Table 2.

The interaction effect between the above-median indicator for the family pressure index and the mobile disbursement treatment appears to be imprecisely estimated when business assets are the dependent variable; this estimate is not statistically significant after multiple hypothesis corrections ($q = 0.083$). This q-value is computed quite generously. Our 'what-if' scenario computational method for q-values implicitly presumes that `capital`, `inventory_value`, and `ent_asset_value` have not all already been tested for statistical significance, yet even this generous computation yields a q-value for this key interaction effect that is not statistically significant at nominal levels. The only pre-registered investment outcome for which there are statistically significant interaction effects between the mobile disbursement treatment and the family pressure index is `inventory_value`, but this outcome is not itself significantly impacted by the mobile disbursement treatment (see our Table 1). It appears that Riley's (2024b) significant heterogeneous treatment effect estimate for `capital` depends on the higher precision of that heterogeneous treatment effect for `inventory_value`, despite the fact that the main treatment effect of the mobile disbursement treatment on `inventory_value` is not statistically significant.

This pre-registration deviation thus drives some of the paper's key find-

Table 2: Heterogeneous Treatment Effects on Investment Variables

	(1) Capital	(2) Inventory Value	(3) Business Assets
Mobile account	15.37	13.86	1.60
	(37.32)	(30.74)	(19.02)
	{0.99}	{0.99}	{0.99}
Mobile disburse	-20.96	-33.79	16.18
	(36.38)	(30.48)	(17.73)
	{0.99}	{0.99}	{0.99}
MA*self control	19.42	15.44	-2.62
	(51.95)	(42.23)	(27.17)
	{0.99}	{0.99}	{0.99}
MD*self control	38.06	41.70	-9.91
	(50.22)	(41.02)	(26.09)
	{0.99}	{0.851}	{0.99}
MA*family pressure	-24.81	-48.68	29.50
	(53.18)	(43.12)	(27.75)
	{0.99}	{0.712}	{0.792}
MD*family pressure	183.55	122.69	57.67
	(51.33)	(41.41)	(26.57)
	{0.001}	{0.008}	{0.083}
Family pressure	8.46	11.39	-0.92
	(40.17)	(31.83)	(21.67)
Observations	2,639	2,638	2,610
R-squared	0.52	0.47	0.42

Note: Intent-to-treat estimates. Monetary outcomes are winsorized at the 99% level and in USD. All regressions include strata dummies. 'Mobile account (MA)' is the treatment where only a mobile money account was provided and the loan was disbursed as cash. 'Mobile disburse (MD)' is the treatment where a mobile money account was provided and the loan was also disbursed onto this account. Self control and family pressure indices are defined as in Riley (2024b). The heterogeneous treatment effect estimates interact the MA and MD treatments with dummies indicating women who are above-median in the specified index. Benjamini, Krieger, & Yekutieli (2006) q-values are displayed in curled brackets, computed under the 'what-if' scenario where in Riley's (2024b) Table 1, capital was replaced with the outcome variable specified by the column. Robust standard errors in parentheses. (Text copied and modified from Riley (2024b).)

ings on mechanisms. Principally, immediately after discussing the results in Table 1, Riley (2024b) makes the explicit claim that the mobile disbursement treatment increases women's investment into both business assets *and* inventory. From pgs. 1428-1429:

> "The Mobile Disbursement treatment allows women to accumulate more assets and inventory in their businesses: women in the Mobile Disbursement treatment have 0.6 additional assets and have increased the total value of their business assets by US$37 and of their inventory value by US$33 (online Appendix Table A8). Hence, businesses that received the Mobile Disbursement treatment seem to be buying both more inventory and one moderate-value asset of a different variety to their existing assets."

As additional justification for this claim, Riley (2024b) points to Appendix Table A8, which makes no multiple hypothesis testing adjustments, in contrast to pre-registration commitments to adjust for multiple hypothesis testing within primary outcome families. It is only absent these pre-registered corrections that the mobile disbursement treatment appears to have precisely-estimated impacts on inventory value.

This claim concerning inventory value is important for making the investment mechanism behind Riley's (2024b) treatment effects on business profits plausible. Riley (2024b) notes that for the women in this experiment, around 80% of business value is in inventory, rather than in business assets. The observed mobile disbursement treatment effect on profits – a 16% increase over the control mean – is not plausibly primarily explained by a 20.6% increase in investment into a capital class that only makes up around 20% of business value in Riley's (2024b) sample.

Further, splitting the non-pre-registered `capital` outcome into its pre-registered components eliminates robust support for a family pressure mechanism. The story that Riley (2024b) tells about the mechanism behind the mobile disbursement treatment's positive effect on business profits depends on two separate findings being simultaneously true. First, it must be true that the mobile disbursement treatment significantly increases business investment, and second, it must be true that the mobile disbursement treatment has significantly higher impacts on investment for women facing high family pressure to share loan funds. However, when looking at the pre-registered business investment variables, statistically significant evidence for the former finding only exists for business assets, and statistically significant evidence for the latter finding only exists for inventory values. The key story of the paper – that the mobile disbursement treatment increases business investment, particularly among women who are subjected to strong family pressure – is not cleanly defended by statistically significant results for either of the pre-registered inventory or business asset outcomes alone. It is only when these outcomes are combined into the non-pre-registered `capital` variable that this story is cleanly defended by statistically significant results.

3.2 Missing and New Outcomes

All results in Tables 1-4 and Table 6 of Riley (2024b) are analyzed using outcomes winsorized at the 1% level, and the paper does not acknowledge that this analytical choice deviates from pre-registration commitments. Riley (2018) commits to repeat all analyses using outcomes winsorized at the 0.5%, 1%, and 2% levels *as robustness checks* (see Table 3). However, every outcome examined in all but one of the tables in Riley (2024b) is winsorized at the 1% level. It is unclear why winsorized outcomes are chosen for the paper's main analyses over unwinsorized outcomes without an acknowledgment that this deviates from the pre-analysis plan.

Several secondary outcome families described in the pre-analysis plan contain one or several (sub)outcomes that are not examined either in the paper or in the Online Appendix. The majority of the variables in outcome family ten (household consumption) are missing, including outcomes on temptation spending, healthcare expenditure, and all four clothing variables. Outcome family seven (female well-being) is missing analysis of an overall well-being index that summarizes all of the individual components of the outcome family. Finally, outcome family nine (household wealth) is missing two of its three pre-registered outcomes, including a poverty score measure and a first principal component of personal asset and housing characteristics.

3.3 Missing and New Analyses

Several types of analyses that are pre-registered are missing from the published paper and Online Appendix. These missing analyses are detailed in Table 3. Additionally, in Table 5 of the published paper, Riley (2024b) examines treatment effects on stated preferences of borrowers. This outcome is not discussed either in the pre-registration nor in its amendment, and is not disclosed to be exploratory.

4 Conclusion

Randomized controlled trials in the field are costly and often organized with independent field partners. These factors make replicating such studies difficult, and thus place extra burdens on researchers to accurately describe the designs of such experiments. We find discrepancies between how Riley (2024b) describes the processes of data collection and treatment assignment, and what the paper's replication files imply about these processes. Detailed documentation is necessary to clarify how this experiment's treatments were assigned, as it seems that the pre-registered protocol for stratified randomization was not successfully carried out.

Table 3: Missing Analyses

Analysis	Location in Pre-Analysis Plan	Notes
Local average treatment effect	Page 7, Section 4.1	Commitment: "To estimate the local average treatment effect, [Equation 1] will be estimated where assignment to treatment is replaced with actual take-up, which is instrument [*sic*] by assignment, giving the two-stage least squares estimator."
Different winsorization levels	Page 10, Section 5.1, Point 2	Commitment: "I will confirm my results are robust to winsorizing at the 0.5, 1 and 2% level." These robustness checks are performed for the results in Table 1 (see Online Appendix Table A11), but are missing for results in other tables.
Kling, Liebman, & Katz (2007) indices	Page 10, Section 5.1, Point 3	Commitment: "I will perform the following robustness checks: (...) 3. I will construct all index variables using the method described in Kling et al. (2007)." Used in Online Appendix Table A27, but missing for remaining indices. E.g., the method is not used for the family pressure or self control indices.
Lee (2009) bounds	Page 10, Section 5.1 and 5.2	Commitment: "For any specific outcome where responses are missing for more than 10% of the sample, I will use the following analysis to assess the sensitivity of my results to missing data: 1. I will construct bounds on parameters using the trimming procedure described in (Lee, 2009)." Baseline attrition is 10.7%.
Correlates of treatment assignment in attrition sample	Page 10, Section 5.2, Point 3	Commitment: "If any one outcome is missing for more than 10% of the sample, I will implement three analyses to characterize the missing data: (...) 3. I will regress a vector of baseline covariates on treatment indicators for the sample of individuals who attrited [*sic*]." Baseline attrition is 10.7%.

Note: Analyses committed to in `PAP+mobile+money+BRAC+RILEY.pdf` (Riley 2018) that are not performed in Riley (2024b).

Furthermore, when results are reported as arising from a pre-registered experiment, it is important that readers can trust that the analyses in the paper accurately reflect the analyses committed to in the pre-registration. The analyses in the published version of Riley (2024b) deviate in undisclosed ways from pre-registration commitments. One undisclosed deviation from the plan, which replaces one of the paper's main outcomes with an alternative variable, has material consequences for some of the paper's main results on mechanisms.

References

Banerjee, Abhijit et al. (2015a). “A multifaceted program causes lasting progress for the very poor: Evidence from six countries”. In: *Science* 348.6236. DOI: 10.1126/science.1260799.

— (2015b). “The miracle of microfinance? Evidence from a randomized evaluation”. In: *American Economic Journal: Applied Economics* 7.1, pp. 22–53. DOI: 10.1257/app.20130533.

Benjamini, Yoav, Abba M Krieger, and Daniel Yekutieli (2006). “Adaptive linear step-up procedures that control the false discovery rate”. In: *Biometrika* 93.3, pp. 491–507.

Crépon, Bruno et al. (2015). “Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in Morocco”. In: *American Economic Journal: Applied Economics* 7.1, pp. 123–150. DOI: 10.1257/app.20130535.

Cull, Robert, Asli Demirgüç-Kunt, and Jonathan Morduch (2007). “Financial performance and outreach: A global analysis of leading microbanks”. In: *The Economic Journal* 117.517, F107–F133. DOI: 10.1111/j.1468-0297.2007.02017.x.

Fassin, Baptiste and Carine Valette (2020). *Microfinance Barometer 2019*. Report. Convergences. URL: https://www.convergences.org/wpcontent/uploads/2019/09/Microfinance-Barometer-2019_web-1.pdf.

Khandker, S. R. (2005). "Microfinance and poverty: Evidence using panel data from Bangladesh". In: *The World Bank Economic Review* 19.2, pp. 263–286. DOI: 10.1093/wber/lhi008.

Kim, Julia C. et al. (2007). "Understanding the impact of a microfinance-based intervention on women's empowerment and the reduction of intimate partner violence in South Africa". In: *American Journal of Public Health* 97.10, pp. 1794–1802. DOI: 10.2105/ajph.2006.095521.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz (2007). "Experimental analysis of neighborhood effects". In: *Econometrica* 75.1, pp. 83–119. DOI: 10.1111/j.1468-0262.2007.00733.x.

Lee, David S. (2009). "Training, wages, and sample selection: Estimating sharp bounds on treatment effects". In: *Review of Economic Studies* 76.3, pp. 1071–1102. DOI: 10.1111/j.1467-937x.2009.00536.x.

Meager, Rachael (2019). "Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments". In: *American Economic Journal: Applied Economics* 11.1, pp. 57–91. DOI: 10.1257/app.20170299.

— (2022). "Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature". In: *American Economic Review* 112.6, pp. 1818–1847. DOI: 10.1257/aer.20181811.

Morduch, Jonathan (1999). “The microfinance promise”. In: *Journal of Economic Literature* 37.4, pp. 1569–1614. DOI: 10.1257/jel.37.4.1569.

Mou, Hongyu, Licheng Liu, and Yiqing Xu (2023). “Panel data visualization in R (panelView) and Stata (panelview)”. In: *Journal of Statistical Software* 107.7. DOI: 10.18637/jss.v107.i07.

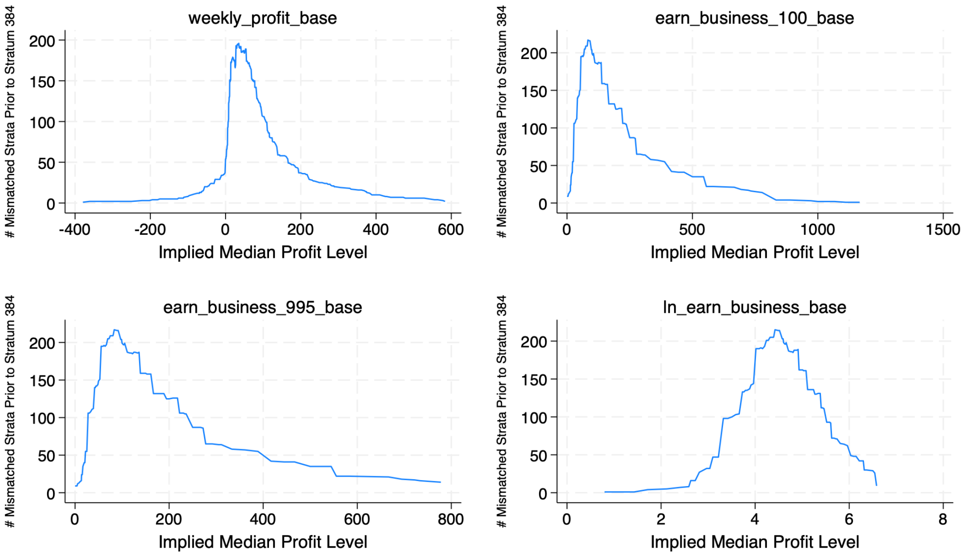
Pronyk, Paul M et al. (2006). “Effect of a structural intervention for the prevention of intimate-partner violence and HIV in rural South Africa: A cluster randomised trial”. In: *The Lancet* 368.9551, pp. 1973–1983. DOI: 10.1016/s0140-6736(06)69744-4.

Riley, Emma (2018). *Impact of disbursing microfinance loans on mobile money accounts*. AEA RCT Registry V5. DOI: 10.1257/rct.1836-5.0.

— (2024a). *Data and code for: Resisting social pressure in the household using mobile money: Experimental evidence on microenterprise investment in Uganda*. Dataset V2. Ann Arbor, MI, U.S.A.: Inter-university Consortium for Political Science Research. DOI: 10.3886/E194886V2.

— (2024b). “Resisting social pressure in the household using mobile money: Experimental evidence on microenterprise investment in Uganda”. In: *American Economic Review* 114.5, pp. 1415–1447. DOI: 10.1257/aer.20220717.

Appendix



Note: Number of strata prior to stratum 384 that are mismatched on an ‘above median baseline business profits’ indicator, where the ‘implied median’ is plotted on each x-axis in units of the baseline business profits variable listed in each graph’s title.

Figure A1: Mismatched Strata by Implied Median Across Alternative ‘Profit’ Variables