RATIONALITY
& COMPETITION
CRC TRR 190

Financial Education Affects Financial Knowledge and Downstream Behaviors

Tim Kaiser (University of Koblenz-Landau & DIW Berlin)
Annamaria Lusardi (George Washington University)
Lukas Menkhoff (HU & DIW Berlin)
Carly Urban (Montana State University)

Discussion Paper No. 240

April 29, 2020

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Dr. Klaus M. Schmidt, University of Munich, 80539 Munich, Germany
+49 (89) 2180 3405 | info@rationality-and-competition.de

Financial education affects financial knowledge and downstream behaviors

Tim Kaiser, Annamaria Lusardi, Lukas Menkhoff, and Carly Urban

April, 2020

Abstract

We study the rapidly growing literature on the causal effects of financial education programs in a meta-analysis of 76 randomized experiments with a total sample size of over 160,000 individuals. The evidence shows that financial education programs have, on average, positive causal treatment effects on financial knowledge and downstream financial behaviors. Treatment effects are economically meaningful in size, similar to those realized by educational interventions in other domains and are at least three times as large as the average effect documented in earlier work. These results are robust to the method used, restricting the sample to papers published in top economics journals, including only studies with adequate power, and accounting for publication selection bias in the literature. We conclude with a discussion of the cost-effectiveness of financial education interventions.

JEL-Classification: D14 (personal finance), G53 (financial literacy), I21 (analysis of education)
Keywords: financial education, financial literacy, financial behavior, RCT, meta-analysis

We thank participants of the 5th Cherry Blossom Financial Education Institute in Washington, D.C., and Michael Collins, Andrea Hasler, Rachael Meager, Olivia Mitchell, and Pierre-Carl Michaud for many helpful comments. We thank Shawn Cole, Daniel Fernandes, Xavier Gine, John Lynch, Richard Netemeyer, and Bilal Zia for providing details about their studies.
Financial support by DFG through CRC TRR 190 is gratefully acknowledged.

Tim Kaiser, University of Koblenz-Landau and German Institute for Economic Research (DIW Berlin), 76829 Landau, Germany; kaiser@uni-landau.de

Annamaria Lusardi (corresponding author), The George Washington University School of Business and NBER, Washington, DC 20052, USA; alusardi@gwu.edu

Lukas Menkhoff, Humboldt-University of Berlin and German Institute for Economic Research (DIW Berlin), 10108 Berlin, Germany; lmenkhoff@diw.de

Carly Urban, Montana State University and Institute for Labor Studies (IZA), Bozeman, MT 59717, USA; carly.urban@montana.edu

1 Introduction

The economic importance of financial literacy is documented in a large and growing empirical literature (Hastings et al. 2013; Lusardi and Mitchell 2014; Lusardi et al. 2017; Lührmann et al. 2018). Consequently, the implementation of national strategies promoting financial literacy and the design of financial education policies and school mandates have become a high priority for policymakers around the world. Many of the largest economies, including most OECD member countries, as well as India and China, have implemented policies enhancing financial education in order to promote financial inclusion and financial stability (OECD 2015). Together, these financial education policies seek to reach more than five billion people in sixty countries, and the number of countries joining this effort continues to grow.

Despite the many initiatives to foster financial literacy, the effectiveness of financial education is debated in quite fundamental ways. Much of the debate stems from the fact that the limited number of early rigorous experimental impact evaluations sometimes showed muted effects, and these early findings have contributed to the perception of mixed evidence on the effectiveness of financial education (see, for example, Fernandes et al. 2014). However, empirical studies on financial education have grown rapidly in the past few years. To account for the large increase in research in this field, we take stock of the recent empirical evidence documented in randomized experiments and provide an updated and more sophisticated analysis of the existing work.

Our main finding is clear-cut: financial education in 76 randomized experiments with a total sample size of more than 160,000 individuals has positive causal treatment effects on financial knowledge and financial behaviors. The treatment effects on financial knowledge are similar in magnitude to the average effect sizes realized by educational interventions in other domains, such as math and reading (see Hill et al. 2008; Cheung and Slavin 2016; Fryer 2016;

Kraft 2019). The effect sizes of financial education on financial behaviors are comparable to those realized in behavior-change interventions in the health domain (e.g., Rooney and Murray 1996; Portnoy et al. 2008; Noar et al. 2017) or behavior-change interventions aimed at fostering energy conserving behavior (e.g., Karlin et al. 2015).

Specifically, the estimated (weighted average) treatment effect is *at least three times as large* as the weighted average effect documented in Fernandes et al. (2014), which examined 13 Randomized Controlled Trials (RCTs). The analysis from our more sophisticated meta-analysis, which accounts for the possibility of cross study heterogeneity, results in an estimated effect of financial education interventions that is more than *five times as large* as the effect reported in Fernandes et al. (2014).

Additionally, we calculate the effect sizes resulting from these interventions and show that they are of economic significance. Our results are robust, irrespective of the model used, when restricting the sample to only those RCTs that have been published in top economics journals, when restricting the sample to only those studies with adequate power to identify small treatment effects, and when employing an econometric method to account for the possibility of publication selection bias favoring the publication of statistically significant results.

In contrast to earlier studies, we do not find differences in treatment effects for low-income individuals and the general population. We also do not find strong evidence to support a rapid decay in the realized treatment effects, though we do not find support for the sustainability of long-run effects either.

For completeness and to asses the external validity of the findings, we also discuss the findings from recent evaluations of financial education mandates and school financial education programs operated at scale.

With this work, we make four main contributions. First, we provide the most comprehensive analysis of the burgeoning work on financial education by using the most rigorous studies: randomized control trials. Second, we focus on a critical feature of empirical analyses on micro data: the heterogeneity in the programs and the many differences that normally one finds in the programs; for example, differences in target groups, quality and the intensity of interventions. Third, we discuss the magnitudes of the effects in terms of economic significance and consider the per participant costs of programs. Fourth, we provide a thorough discussion of topics raised in previous work, i.e., how to assess the impact of financial education and whether education decays with time. We believe that this work can provide useful guidance for those evaluating future financial education programs.

The paper has seven sections: section 2 serves as a primer on statistical meta-analyses; section 3 describes our method; section 4 presents descriptive statistics of our data; section 5 presents the results of our analyses; section 6 discusses the economic significance of our effect sizes and the cost-effectiveness associated with these effects; section 7 concludes.

2 Background

As the amount of evidence from rigorous empirical studies in a given field grows over time, there is an increased need to synthesize and integrate the existing findings to reach a consistent conclusion. Traditionally, economists have relied on narrative reviews, where experts on a given literature select and discuss the most relevant findings. The advantage of such an approach is that the experts are expected to have a good understanding of the existing studies and can add value by summarizing, interpreting, and linking together the most convincing (internally valid) studies in a narrative review. Examples of widely cited narrative reviews in the financial education literature are Fox et al. (2005), Collins and O’Rourke (2010), Xu and Zia (2012), Hastings et al. (2013), and Lusardi and Mitchell (2014).

As empirical literatures grow larger, however, narrative literature reviews can become difficult, since it is hard to describe a large number of empirical estimates and discuss all of the possible sources of heterogeneity in reported findings. Meta-analyses have thus become more common in economics when aggregating findings from many studies. Some examples of recent meta-analyses in economics include Meager (2019), which studies microcredit expansions, and Beuermann and Jackson (2018), which examines the effect of going to parent-preferred schools. Meta-analyses can serve as a complement to narrative reviews when there is a sufficiently large number of well-identified studies on the same empirical research question. A meta-analysis—a systematic, quantitative literature review—is well suited to obtain an estimate of the average effects of a given program and to study the heterogeneity in reported findings (Stanley 2001).

As noted earlier, Fernandes et al. (2014) was the first meta-analysis performed in the field of financial education. We differ from this initial and well-cited study in three major ways. First, we update the dataset to incorporate the many papers that have been written since the meta-analysis by Fernandes et al. was published. As Figure 1 shows, the field grew exponentially after 2014, so previous reviews cover only a small part of the work that currently exists. Second, we attempt to replicate the findings in Fernandes et al. (2014), and we provide estimates more common in meta-analysis literature, which account for heterogeneity in effect sizes across studies. This takes into consideration, for example, the intensity of the program. Third, we have chosen to focus solely on what are considered the most rigorous sources of evidence, i.e., randomized experiments. RCTs provide more consistent internal validity than observational and quasi-experimental studies, especially since there are no universally accepted instruments for financial literacy, and one can debate whether existing non-randomized trials have made use of convincing empirical strategies addressing endogeneity of selection into treatment. Judging the quality of quasi-experimental studies and determining which to include or exclude from the meta-analysis gives researchers an additional degree of freedom that we wish to remove. Importantly, the number of RCTs has grown from just 13 in the Fernandes et al. (2014) review to 76 as of 2019. In those 13 studies, the authors found the weakest effects of financial education interventions reviewed in their work. Fernandes et al. (2014) assert that these studies provide the strongest evidence against financial education.

< Figure 1 about here >

In addition to Fernandes et al. (2014), there have been three follow-up meta-analyses on financial education programs: Miller et al. (2015); Kaiser and Menkhoff (2017); and Kaiser and Menkhoff (2019). These meta-analyses present a more nuanced view of financial education interventions than the original paper by Fernandes et al (2014) by including additional studies and accounting for differences in program design and outcomes studied. This study will build upon those, but it expands the contribution by focusing solely on RCTs, including additional years of data, deepening the methodological discussion (including new robustness checks), providing a thorough discussion of economic significance, and incorporating information on program costs. By contrast, Miller et al. (2015) focus on less than 20 studies and put emphasis on examining impact differences across outcomes. Kaiser and Menkhoff (2017) concentrate on the determinants of effective financial education interventions, while Kaiser and Menkhoff (2019) focus on financial education interventions in schools.

3 Methods

This section describes our inclusion criteria for the papers on financial education (Section 3.1), the details we use in constructing our database of effect sizes (Section 3.2), and the specifics of the empirical model we employ (Section 3.3).

3.1 Inclusion criteria

In order to draw general conclusions about a given literature, one has to conduct a systematic search of the literature and apply inclusion criteria that are defined ex-ante. We conducted a search of all relevant databases for journal articles and working papers (see

Appendix A for the list of the studies we considered and a summary of the data we extracted from those studies), and apply three inclusion criteria to the universe of records return in this set. Criteria of inclusion: (i) Studies reporting the causal effects of educational interventions designed to strengthen the participants' financial literacy and/or leading to behavior change in the area of personal finance; (ii) studies using random assignment into treatment and control conditions; (iii) studies providing a quantitative assessment of intervention impact that allows researchers to code an effect size estimate and its standard error. Where necessary information is partially missing, we consulted additional online resources related to the article or contacted the authors of the studies. We only consider the main results discussed in the text, and we do not code redundant effect sizes (e.g., effect sizes arising from other specifications of a given statistical model in the robustness section). Table A1 provides a list of all the studies considered in our analysis.

3.2 Constructing the database

Our analysis aggregates treatment effects of financial education interventions into two main categories. First, we code the effect of financial education on *financial literacy* (i.e., a measure of performance on a financial knowledge test) since improvement in knowledge is usually the primary goal of financial education programs (Hastings et al. 2013; Lusardi and Mitchell 2014) and is expected to be one of the channels via which financial behavior is influenced. We do not include self-assessments of changes in financial knowledge as an outcome.

Second, we code the effect of financial education on *financial behaviors.* These behaviors can be further disaggregated into the following categories: Borrowing, (retirement) saving, budgeting and planning, insurance, and remittances. It is useful to know, for example, which behavior is more easily impacted by financial education. Table A3 provides an overview of the categories and definitions of outcome types.

We code all available effect sizes per study on financial knowledge and behavioral outcomes. We include multiple estimates per study if multiple outcomes, survey-rounds, or treatments are reported. We only extract main treatment effects reported in the papers. Thus, we do not consider estimates reported in the "heterogeneity-of-treatment-effects-section" within papers, such as sample splits or interaction-effects of binary indicators (e.g., gender, income, ability, etc.), with the treatment indicators. We aim to only consider intention-to-treat effects (ITT), unless these are not reported. If only local average treatment effects (LATE) or the treatment effect on the treated (TOT) are reported, we included these in our analysis and check for statistical differences, as described in Appendix B.[1]

This process leads to the inclusion of 76 independent randomized experiments described further in Section 4.

3.3 Empirical model

A major challenge in every meta-analysis lies in the heterogeneity of the underlying primary studies and how to account for it. In the financial education literature, heterogeneity arises from several sources; in our sample, randomized experiments on financial education programs have been conducted in 32 countries with varying target groups (see Table A1 in Appendix A). Moreover, the underlying educational interventions are very diverse, ranging from provision of an informational brochure to offering high-intensity classroom instruction; outcomes are also measured at different points in time and with different types of data. Accommodating this heterogeneity is important in order to draw general conclusions about the findings.

[1] We also show results for the sample of studies reporting the ITT in Appendix B.

When there is such heterogeneity in the studies under consideration, meta-analyses require certain assumptions about the sources of variance in the observed treatment effect estimates. Consider a set of j randomized experiments, each of them reporting an estimate of a causal (intention to treat) treatment effect relative to a control group.[2] Assuming no heterogeneity in true effects implies that the observed estimates of a treatment effect are sampled from a distribution with a single true effect β_0 and variance σ^2, as in the following meta-analysis model:

$$y_j = \beta_0 + \epsilon_j \qquad (1)$$

where y_j is an estimate of a treatment effect in the jth study, β_0 defines the common true effect, and ϵ_j is the study level residual with $\epsilon_j \sim N(0, \sigma_j^2)$. Thus, the estimate of the common true effect is given by estimating the above model with weighted least squares using inverse variance weights ($w_j = \frac{1}{\sigma_j^2}$). While this may be a reasonable assumption for some empirical literatures, such as medical trials with identical treatment, dosage, and procedures for measuring outcomes, this is clearly not a reasonable assumption in the context of educational interventions, which tend to be quite diverse.

A more reasonable approach in an educational setting would be to assume heterogeneity between studies, hence assuming a distribution of possible true effects, allowing true effects to vary across studies with identical within-study measurement error. The weighted average effect

[2] Because each study j may report its treatment effect estimate in a different unit (i.e., a different currency or on different scales), we convert each estimate to a (bias corrected) standardized mean difference (Hedges' *g*), such that the treatment effect estimate y_j is standardized as $g_j = \frac{\mathrm{M}_T - \mathrm{M}_\mathrm{C}}{\mathrm{SD}_\mathrm{p}}$ with $SD_p = \sqrt{\frac{(\mathrm{n_T}-1)\, SD_T{}^2 + (\mathrm{n_C}-1)\, SD_C{}^2}{n_T{}^2 + n_C{}^2 - 2}}$, i.e., the mean difference in outcomes between treatment (M_T) and control ($\mathrm{M_C}$) as a proportion of the pooled standard deviation (SD_p) of the dependent variable. $\mathrm{n_T}$ and SD_T are the sample size and standard deviation of the treatment group, and n_C and SD_C are for the control group. Additionally, the standard error of each standardized mean difference is defined as: $SE_{g_j} = \sqrt{\frac{\mathrm{n_T}+\mathrm{n_C}}{\mathrm{n_T}\mathrm{n_C}} + \frac{g_j^2}{2(\mathrm{n_T}+\mathrm{n_C})}}$.

then does not represent a single true effect but instead the mean of the distribution of true effects. Thus, the model can be written as:

$$y_j = \beta_0 + \upsilon_j + \epsilon_j \qquad (2)$$

with $\upsilon_j \sim N(0, \tau^2)$ and $\epsilon_j \sim N(0, \sigma_j^2)$. τ^2 is the between-study variance in true effects that is unknown and has to be estimated from the data,[3] and σ_j is the within-study standard error of the treatment effect estimate y_j that is observed for each study j. Subsequently, weighted least squares is used to estimate β_0 with inverse variance weights defined as $w_j = (\tau^2 + \sigma_j^2)^{-1}$. Thus, instead of estimating one common effect, the goal is to estimate the mean of the distribution of true effects.

While the illustration so far has considered cases in which each study contributes one independent treatment-effect estimate, this is generally not the case in the financial education literature. Instead, studies may report treatment effect estimates from multiple treatments and a common control group within studies, at multiple time-points and for multiple outcomes. Therefore, we extend the model above to incorporate multiple (and potentially correlated) treatment effect estimates within studies:

$$y_{ij} = \beta_0 + \upsilon_j + \epsilon_{ij} \qquad (3)$$

y_{ij} is the ith treatment effect estimate within each study j. β_0 is the mean of the distribution of true effects, $\upsilon_j$is the study-level random effect with $\upsilon_j \sim N(0, \tau^2)$, τ^2 is the between study variance in true effects, and $\epsilon_{ij} \sim N(0, \sigma_{ij}^2)$ is the residual of the ith treatment effect estimate within each study j. This model allows between-study heterogeneity in true effects but assumes that treatment effect estimates within studies relate to the same study-specific true effect. This

[3] There are several possible algorithms to estimate the between-study variance τ^2. Our approach uses the method of moments estimator (see Harbord and Higgins 2008), but iterative approaches, such as (restricted) maximum likelihood or empirical Bayes estimation, are also frequently used in meta-analyses.

means the common within-study correlation of treatment effect estimates is induced by random sampling error.

While the estimator proposed in Hedges et al. (2010) does not require an exact model of the within-study dependencies in true effects, Tanner-Smith and Tipton (2014) and Tanner-Smith et al. (2016) suggest that the following inverse variance weights (w_{ij}) are approximately efficient in case of a correlated effects model:

$w_{ij} = \left\{\left(\tau^2 + \frac{1}{k_j}\sum_{k_j=1}^{k_i}\sigma_{ij}^2\right)\left[1 + \left(k_j - 1\right)\rho\right]\right\}^{-1}$, where τ^2 is the estimated between-study variance in true effects, $(\frac{1}{k_j}\sum_{k_j=1}^{k_i}\sigma_{ij}^2)$ is the arithmetic mean of the within-study sampling variances (σ_{ij}^2) with k_j being the number of i effect size estimates within each study j, and ρ is the assumed common within-study correlation of treatment effect estimates.

We estimate the model with these weights and choose ρ = 0.8 as the default within-study correlation of estimates (see Tanner-Smith and Tipton 2014). However, sensitivity analyses of such an assumption are easily implemented, and we show results for ρ = [0, 0.9] in increments of 0.1 in Appendix B.

Our method addresses several shortcomings of the analysis presented in Fernandes et al. (2014). First, we are able to formally investigate the importance of modeling between-study heterogeneity in treatment effects and to compare the results to a model with the common-effect assumption used in Fernandes et al. (2014). This is important because, as mentioned before, financial education programs can be very different from each other. Second, we make use of the all of the statistical information reported in primary studies, since the method used in this paper is able to accommodate multiple estimates within studies, and thus is not dependent on creating highly aggregated measures, such as the within-study average effect sizes reported in Fernandes et al. (2014). To probe the robustness of our results, we estimate five alternative models (see Appendix B), including a correction for potential publication selection bias and a consideration of the power of the underlying primary studies. We are also careful to replicate the methods of Fernandes et al. (2014), as reported in Appendix D.

4 Data

To arrive at an unbiased estimate of the mean of the distribution of true effects of financial education programs, we collect a complete list of randomized experiments in the financial education literature. We build on an existing database and update it using the search strategy described earlier, which is also used in Kaiser and Menkhoff (2017). We augment the earlier dataset used in previous work with published randomized experiments on financial education through January 2019 (end of collection period for this paper).[4] Appendix A contains a detailed description of the papers included in our meta-analysis and the types of outcomes coded. Applying our inclusion criteria, we arrive at a dataset of as many as 68 papers reporting the effects of 76 independent-sample experiments. This is a much bigger sample of RCTs than any previous meta-analyses.

An important part of our meta-analysis is the inclusion of many recent papers in our dataset, which enables us to provide a comprehensive and updated review of the large and rapidly growing amount of research done on this topic. The review by Fernandes et al. (2014) is the first paper in the literature, and it covers only 13 RCTs from which they code 15 observations. The meta-analysis in Miller et al. (2015) covers a total of seven RCTs. Of our 76 independent-sample experiments, one-third have not been included in the most recent meta-analysis, by Kaiser and Menkhoff, (2017).[5] Thus, we expand greatly on those previous studies. Table C1 in Appendix C contains a comparison of our dataset of RCTs to these earlier accounts of the literature.

[4] This paper has gone through revisions and the end of the collection period refers to when we started extracting and analyzing the data.

[5] We are also careful to update all of the papers to the latest version and include, for example, the estimates in the published version of the papers.

From our sample of 76 independent randomized experiments, we extract a total of 673 estimates of the effects of the program (the treatment effects). Out of these, 64 studies report a total of 458 treatment effects on financial behaviors (see Table A4 in Appendix A). Thus, we are able to work on a large number of estimates. The studies vary in their choice of dependent variables, ranging from a number of financial behaviors to financial knowledge. To illustrate some simple differences in studies, we note that 23 studies report 115 treatment effect estimates on *credit behaviors*, and 23 studies report 55 treatment effect estimates on *budgeting behavior.* The largest number of estimates is on *saving behavior,* with 54 studies reporting a total of 253 treatment effect estimates. Six studies report 18 treatment effect estimates on *insurance behavior*, and six studies report 17 estimates on remittance behavior. Fifty studies report 215 treatment effect estimates on *financial knowledge* and 38 studies report treatment effects on both knowledge and behaviors. We have a sizeable number of estimates for each outcome.

We start our analysis by showing that the descriptive statistics alone suggest that financial education is, on average, effective in improving both knowledge and behavior.

< Table 1 about here >

The average effect size across all types of outcomes, reported in Table 1, is 0.123 standard deviation (SD) units (SD=0.183), and the median effect size is 0.098 SD units.[6] The minimum effect size is -0.413, and the maximum effect size is 1.374. The average standard error of the treatment effect is 0.085 (SD=0.049) and the median standard error is at 0.072.[7]

We first note that there is substantial variation in instruction time in the programs, where the average estimate is associated with a mean of 11.71 hours of instruction (SD=16.27), and the median is associated with 7 hours of instruction (Table 1). Treatment effects are estimated 30.4 weeks (7 months) after treatment, on average, with a standard deviation of 31.65 weeks

[6] Note that all effect sizes are scaled such that desirable outcomes have a positive sign (i.e., we are coding a negative coefficient on “loan default” as a positive treatment effect (i.e., reduction in loan default) and vice versa.

[7] The average sample size across the 76 randomized experiments is 2,136 and the median sample size is 840.

(7.3 months). The median study does not focus on immediate effects: the median time passed between financial education treatment and measurement of outcomes is 25.8 weeks (5.9 months). This is useful information for assessing the impact of programs, in particular if one hypothesizes a decay of effectiveness with time, as emphasized by Fernandes et al. (2014). Further, we note that nearly three quarters (72.4 percent) of the treatment effect estimates target low-income individuals (income below the median), and 60.8 percent of the estimates are from programs studied in developing economies; 30.8 percent of all estimates reported in randomized experiments appear in top economics journals, which reflects the high quality of this sample of studies. The average age across all reported estimates is 33.5 years, where 7.5 percent of estimates are focused on children (<14 years old), 20 percent are focused on youth (14-25 years old), and 72.4 percent are focused on adults (>25 years old).

When assessing the effectiveness of financial education, interventions may not necessarily lead to changes in behavior if people have resource constraints or are in the early part of the life cycle, as highlighted in Lusardi et al. (2017). In some cases, people may already be acting optimally and in other cases, even after exposure to financial education, it may be optimal to not change behavior. Determining which behaviors should optimally change requires a theoretical framework sometimes lacking in this literature.

5 Results

We present the results in three steps. Section 5.1 shows the main results of our meta-analysis of the universe of randomized experiments (up to 2019) and compares the results to the first meta-analysis of the literature by Fernandes et al. (2014). Section 5.2 summarizes the results of comprehensive robustness exercises that are reported in full in Appendix B. Section 5.3 examines our main effects further by discussing the results by outcomes, such as financial knowledge and a variety of financial behaviors. Section 5.4 presents our main results once we disaggregate the data into various sub-samples of interest.

5.1 A meta-analysis of randomized experiments

We describe our findings by first plotting the universe of 673 raw effects extracted from the 76 studies against their inverse standard error (precision) in Figure 2. We disaggregate the data and distinguish between estimated treatment effects on *financial behaviors* (n=458) and *financial knowledge* (n=215).[8] The unweighted average effect on financial behaviors is 0.0898 SD units, and the unweighted average effect on financial knowledge is 0.187 SD units. With this simple analysis of the raw data, we find that financial education improves both financial knowledge and behaviors.

A visual inspection of the plot in Figure 2 shows that both samples of effect sizes resemble a roughly symmetric funnel until effect sizes of 0.5 SD units and above. We investigate the possibility of publication selection bias[9] in the financial education literature in Appendix B (see Figure B1 and Table B1) and find that accounting for this potential publication bias does not qualitatively change the result of positive average effects of financial education.

< Figure 2 about here >

Next, we provide a comparison of the data in our study with the results in Fernandes et al. (2014). Specifically, we estimate the weighted average effect on financial behaviors using 'Robust Variance Estimation in Meta-Regression with Dependent Effect Size Estimates' (RVE) under the common true effect assumption[10] made in Fernandes et al. (2014) and compare our

[8] We refer to n as the number of estimates and not the number of participants in the studies.

[9] Publication selection bias refers to the potential behavior of researchers to be more likely to report and journal editors being more likely to publish statistically significant results.

[10] Thus, we assume $\tau^2 = 0$, i.e., the weights are defined as $w_{ij} = \left\{\left(\frac{1}{k_j}\sum_{k_j=1}^{k_i}\sigma_{ij}^2\right)\left[1 + \left(k_j - 1\right)\rho\right]\right\}^{-1}$. Note, that Fernandes et al. (2014) use only one observation per study by creating within-study average effect sizes, i.e., the weights in their study are defined as $w_j = \frac{1}{\sigma_j^2}$. We show results with this approach in Table B3 of Appendix B.

result in the larger sample of 64 RCTs to their earlier result based on 15 observations from 13 RCTs.[11] These results are reported in Figure 3.

< Figure 3 about here >

A few important clarifications are in order: Fernandes et al. (2014)'s estimate and standard error in Figure 3 is from the analysis of 15 observations of RCTs in their paper, not from our analysis of their data. We were not able to exactly replicate this result, and in the process, we uncovered four data errors in the direct coding and classification of RCT effect sizes. In Appendix D, we describe our attempt to replicate the original result by Fernandes et al. (2014) and thoroughly document each coding discrepancy.

Taking their estimates at face value, Figure 3 shows that simply updating the dataset to incorporate the burgeoning recent work increases the effect by more than three times. Compared to the estimate reported in Fernandes et al. (2014) of 0.018 SD units (with a 95% confidence interval (CI_{95}) from -0.004 to 0.022), the weighted average effect in this larger sample of recent RCTs is about 3.6 times higher. The new estimate of the effect size, even with the identical assumption of a common true effect, clearly rules out a null effect of financial education (0.065 SD units with CI_{95} from 0.043 to 0.089). Thus, one of the main findings of Fernandes et al. (2014) is not confirmed in this larger sample of RCTs.

Because the common true effect assumption is potentially problematic in the context of heterogeneous financial education interventions, we estimate the mean of a distribution of true effects using the model specified in equation 3. In addition to the mentioned theoretical reasons

[11] We convert the correlations used as an effect size metric by Fernandes et al. (2014), *(r)* to a standardized mean difference (Cohens' *d*) $d = \frac{2r}{\sqrt{1-r^2}}$ and we convert the standard error using $SE_d = \sqrt{\frac{4SE_r^2}{(1-r^2)^3}}$ (cf. Lipsey and Wilson 2001). This is true under the assumption that the outcome measures in each group are continuous and normally distributed and that the treatment variable is a binary variable indicating treatment and control groups, i.e., a valid assumption in the context of RCTs. To arrive at the "bias corrected standardized mean difference" (Hedges' *g*) one may apply the following bias correction factor ex post $g = d\left(1 - \frac{3}{4(n_1+n_2-2)-1}\right)$ (cf. Borenstein et al. 2009) but these metrics are near identical in the context of the financial education literature where the average sample size is 2,136 and the median sample size is 840.

to assume a distribution of true effects rather than a single true effect, we note that formal tests of heterogeneity show that at least 86.4 percent of the observed between-study variance can be attributed to heterogeneity in true effects and only 13.6 percent of the observed variance would have been expected to occur by within-study sampling error alone (see Table B3 in Appendix B).[12]

Figure 3 shows the result of the random-effects RVE model. In our view, this estimated mean of the distribution of financial education treatment effects is the most appropriate aggregate effect size to consider; the estimate results in a mean of 0.1003 SD units [CI_{95} from 0.071 to 0.129], and thus, is significantly different from the estimate using the common true effect assumption. The effect of financial education is now approximately 5.5 times larger than the estimate reported in Fernandes et al. (2014). This effect is very similar in magnitude to statistical effect sizes reported in meta-analyses of behavior-change interventions in other domains such as health (e.g., Rooney and Murray 1996; Portnoy et al. 2008; Noar et al. 2007) or energy conservation behavior (e.g., Karlin et al. 2015).

To summarize, evidence that incorporates the updated set of papers shows that financial education is effective, on average. Hence, we do not confirm the estimates from early studies, which are based on a small number of interventions.

5.2 Model sensitivity

We probe the robustness of our findings about the average effect of financial education programs with various sensitivity checks that are reported in full in Appendix B. These tests include (i) estimating three alternative meta-analyses including models with a common-effect assumption, (ii) investigating and correcting for potential publication selection bias, (iii) restricting the sample to only those studies with adequate power to identify small treatment

[12] A Cochrans Q-test of homogeneity (with one synthetic effect size per study) results in a Q-statistic of 464.71 ($p<0.000$).

effects, (iv) choosing different assumed within-study correlations of treatment effect estimates for the random-effects RVE approach, and (v) creating one synthetic effect size per study (inverse-variance weighted within-study average) and estimating both fixed-effect and random-effects models with one observation per study. All of these robustness checks confirm the main conclusions of our paper.[13]

5.3 Outcome domains

In addition to the effects on financial behaviors aggregated above (Figure 3), i.e., all behaviors, we also include estimates on financial knowledge (Figure 4). Treatment effects on *financial knowledge* are larger than the effect sizes on *financial behaviors*.

< Figure 4 about here >

Specifically, we find that the mean of the distribution of true effects in our sample is estimated to be 0.204 [CI_{95} from 0.152 to 0.255]. Hence, here as well, we cannot confirm the finding by Fernandes et al. (2014) based on 12 papers (average effect of about 0.133 SD units).[14] Instead, our average effect on financial knowledge is very similar to the average effects of educational interventions in math or reading (see Hill et al. 2008; Cheung and Slavin 2016; Fryer 2016; Kraft 2019).

Effect sizes on *financial behaviors* are mostly not statistically different from each other, suggesting the adequacy of pooling across these outcomes. However, additional analyses shown in Table B2 in Appendix B suggest that the results on *saving behavior* and *budgeting behavior* are the most robust, while the effects on other categories of financial behaviors are less certain due to either fewer studies including these outcomes (*insurance* and *remittances)* or high heterogeneity in the estimated treatment effects (*credit behaviors)*. This result is

[13] We also check the robustness of results when excluding any papers of the authors of this meta-analysis.

[14]See Fernandes et al. (2014), p. 1867: "In 12 papers reporting effects of interventions on both measured literacy (knowledge) and some downstream financial behavior, the interventions explained only 0.44% of the variance in financial knowledge," i.e., $\sqrt{r^2} = 0.066$ or d=0.133.

generally in line with earlier accounts of the literature, such as Fernandes et al. (2014), Miller et al. (2015), and Kaiser and Menkhoff (2017), and extends to the larger set of RCTs.

5.4 Subgroup analyses

In order to better understand the sources of heterogeneity in this literature, we further disaggregate our data into various subgroups and investigate the mean effect of financial education interventions.

5.4.1 Sample population

We disaggregate the sample of RCTs by characteristics of the sample population. First, we split the sample by country-level income, distinguishing between high income economies and developing economies, to account for differences in resources.[15] We find that the treatment effects of interventions in developing economies on financial behaviors are about 9.56 percent smaller than those in richer countries; however this difference is not statistically significant (see Panel A(a) of Table 2). Previous meta-analyses have found slightly smaller effect sizes for interventions in developing economies when controlling for additional features of the programs, such as intensity (cf. Kaiser and Menkhoff 2017). Treatment effects on financial knowledge are about 46 percent smaller in developing economies than in high income economies (see Panel B(a) of Table 2); this difference is statistically significant, and this is also in line with earlier evidence based on a smaller sample of RCTs (cf. Kaiser and Menkhoff 2017).

< Table 2 about here >

We next look at differences between low-income individuals and people with average or above average individual income (relative to the average within-country income). While

[15] Country groups are based on the World Bank Atlas method and refer to 2015 data on Gross National Income (GNI) per capita. Low-income economies are defined as those with a GNI per capita of $1,025 or less in 2015, lower-middle income economies are defined by a GNI per capita between $1,026 and $4,035, upper-middle income economies are those with a GNI per capita between $4,036 and $12,475, and high-income economies are defined by a GNI per capita greater than $12,475.

interventions with low-income individuals show smaller treatment effects, on average, which is in line with earlier accounts of the literature (Fernandes et al. 2014; Kaiser and Menkhoff 2017), we—in contrast to these earlier studies—do not find any significant differences between these two samples (see Panel A(b) and Panel B(b)); this indicates that recent RCTs added to the sample show smaller differences in treatment effects between groups than those interventions studied in the earlier literature.

Additionally, we disaggregate our sample by the age of the participants (see Panel A(c) and Panel B(c) of Table 2). Treatment effects on financial behaviors are smallest for children (below age 14) (0.064 SD units) relative to youth (ages 14 to 25) (0.1203 SD units) and adults (above age 25) (0.1068 SD units), while the latter difference is only marginally significant. Treatment effects on financial knowledge, on the other hand, are estimated to be largest among children (0.2763 SD units) relative to youth (0.1859 SD units) and adults (0.2001 SD units). These differences, however, are not statistically significant due to large uncertainty around the estimate for children, which is based on 15 observations in seven studies (CI_{95} from 0.0076 to 0.545).

5.4.2 Journal quality

To address possible concerns regarding the internal validity and general rigor of the included experiments and to focus on what editors and reviewers have judged to be the highest quality evidence, we restrict the sample to studies published in top general interest or top field economics journals only.[16] We compare the estimated treatment effects on financial behaviors of the 15 studies published in these journals to the estimated treatment effects of the other 49 studies published in other journals or as working papers. While treatment effects are estimated

[16] These journals are: (1) *Quarterly Journal of Economics*, (2) *Journal of Political Economy*, (3) *American Economic Journal: Applied Economics*, (4) *American Economic Journal: Economic Policy*, (5) *Journal of the European Economic Association*, (6) *Economic Journal*, (7) *Journal of Finance*, (8) *Review of Financial Studies*, (9*) Management Science*, (10) *Journal of Development Economics*. There were no publications in other top journals, such as the *American Economic Review, Econometrica, and the Review of Economic Studies*.

to be slightly smaller in these types of publications, there are no statistically significant differences between these types of publications (see Panel A(d) and Panel B(d) of Table 2). The same is true for effect sizes on financial knowledge where eight experiments published in top general interest or top field economics journals report smaller, albeit not statistically different, effect sizes than 42 experiments published in other journals or as working papers.

5.4.3 Time horizon

Finally, we tackle the important topic of potential decay of effectiveness of financial education over time. We disaggregate the sample of treatment effects within studies, considering the time span between financial education treatment and measurement of outcomes (see Panel A(e) and Panel B(e) of Table 2). We start by looking at treatment effect estimates that measure outcomes in the very short run (i.e., a time span of less than six months). The average effect of financial education on financial behaviors within this sample of 34 RCTs (180 effect sizes) is 0.0991. Looking at treatment effects on financial behaviors that are measured at a time span of six months or more (28 experiments and 260 estimates), we find that the estimates reduced to 0.071 SD units [CI_{95} from 0.0425 to 0.0995], which is a marginally significant difference relative to the set of studies with the shorter time horizon.

We next restrict the sample further to 18 studies that measure treatment effects on financial behaviors after at least one year. The estimate is statistically not different to the studies with shorter time horizon after treatment (0.0878 SD units). Restricting the sample to even longer time spans, i.e., ten RCTs that measure effects on financial behaviors at least 1.5 years after treatment or longer, results in an estimated average of 0.0653 SD units. These effects are slightly reduced but are still not statistically different from the other estimates. Restricting the set of RCTs further to those seven studies that measure treatment effects on financial behaviors at least two years after treatment or longer, results in an estimate of 0.0574 SD units, which is again not statistically different from the other estimates and does not include the possibility of zero effects (within the limits of the 95% CI). Overall, there is some decay in effectiveness when measurement is delayed by six months or more; however, beyond this threshold we do not observe any further significant decline.

Regarding the decay in financial knowledge, we find significantly larger effects (0.2305 SD units) in 36 RCTs measuring effects on financial knowledge in the very short run (i.e., at a time span shorter than six months) relative to those with time horizons above six months (0.1408 SD units), but no statistically significant differences at longer time horizons (more than 6 months or more than 12 months). However, only five studies measure treatment effects on financial knowledge considering time horizons between 12 and 18 months, and no longer-term studies exist in our sample.

Overall, these examinations of the possible decay in outcomes highlighted by Fernandes et al. (2014) do not find conclusive evidence. This indicates one can neither rule out sustained and relatively large effects nor close to zero effects of financial education at longer time spans due to a very limited number of studies that measure very long-run outcomes. We attribute the previous finding of a relatively rapid decay to the fact that Fernandes et al. (2014) chose to model this relationship in a meta-regression model with four covariate variables based on a sample of only 29 observations.[17] Thus, the evidence suggesting insignificant effects after time spans of more than 18 months is based on a very limited number of observations and should be viewed with caution in light of the large uncertainty around this estimated effect.

6 Discussion of the economic significance of financial education

[17] We also rerun their type of model (a regression of the estimated effect size on "linear effects of mean-centered number of hours of instructions, linear and quadratic effects of number of months between intervention and measurement of behavior, and the inter action of their linear effects" (Fernandes et al. 2014, p. 1867) with our updated data (419 observations within 52 studies) and find coefficient estimates with large standard errors (i.e., insignificant coefficients) throughout (see Table B6 in Appendix B).

As is true with any analysis of interventions, it is important to understand not just the statistical effect size but also the economic significance of the effects of financial education. A growing literature in education is concerned with interpreting effect sizes across studies, samples, interventions, and outcomes. This section discusses the choice in Fernandes et al. (2014) to focus on the "variance explained" as a measure of the effect size (Section 6.1). We couch our effect sizes into the recent literature on explaining and comparing the effects of education interventions (Section 6.2), provide a back of the envelope analysis of the cost-effectiveness of financial education interventions based on our findings (Section 6.3), and discuss the external validity of the RCT estimates by taking into account recent quasi-experimental studies (Section 6.4).

6.1 Statistical effect sizes

A main argument in Fernandes et al. (2014) is that even though the statistical effects of financial education on financial outcomes are positive in the overall sample, the magnitudes are small. However, Fernandes et al. (2014) create the illusion of miniscule effects (when, in fact, they can be economically significant) by using "variance explained," i.e., a squared correlation coefficient, as their effect size metric.

The fact that this metric creates the illusion of miniscule effects can be illustrated with a simple example. Consider the median effect of education (and specifically, structured pedagogy) interventions in developing countries, which is roughly 0.13 SD units (see Evans and Yuan 2019). Translating this to the (partial) correlation in Fernandes et al. (2014) results in a correlation coefficient of 0.06, which explains only 0.36 percent of the variance in learning outcomes. Thus, according to this criterion, this education intervention would be interpreted to be ineffective, as it "explains little of the variance." However, Evans and Yuan (2019) report that this is actually equivalent to a sizeable effect, approximately 0.6-0.9 years of "business as usual schooling," depending on their choice of specification. In further analysis, they estimate the returns to education (and specifically literacy) in Kenya, and estimate the net present value of this intervention to be 1,338 USD at an average annual income of 1,079 USD in 2015 PPP. Reported in this way, rather than the metric chosen by Fernandes et al (2014), these effects are unlikely to be considered economically miniscule. Thus, it can be problematic to rely upon the "variance explained" in determining the economic interpretation of statistical effect sizes.

6.2 Interpreting treatment effects in the education literature

Recent work in education interventions aims to compare effect sizes across heterogeneous treatments, populations, and outcomes—as we are doing in our analysis—and we turn to that work to get some guidance on interpreting effects. Kraft (2019) suggests five key considerations in determining whether or not programs are effective. First, one should make sure only studies with a causal interpretation (e.g., RCTs) are included in "effect sizes." Second, one should expect effects to be larger when the outcome is easier to change; this is particularly relevant if the intervention is *designed* to change the specific outcome. Third, one should take into account heterogeneous effects on different populations. Fourth, one should always consider costs per participant. A small effect size can have a large return on investment if the per participant cost is low. Fifth, one should consider whether the program is easily scalable. We have followed these recommendations.

With these five points in mind, Kraft (2019) further points to a scheme for assessing the effect of education interventions with academic outcomes (i.e., test scores) as the main outcome of interest. He suggests that effects larger than 0.20 standard deviations are "large," effects between 0.05 and 0.20 standard deviations are "medium," and effects under 0.05 standard deviations are "small." This classification is roughly consistent with the What Works Clearinghouse (2014), Hedges and Hedberg (2007) and Bloom et al. (2008). Our effects on financial knowledge in Figure 4 show an effect size of roughly 0.203, consistent with a "large" effect of an education intervention on test scores.

Kraft (2019) also notes that it is more difficult to affect long-run outcomes that are not directly addressed in the intervention. It is, thus, not surprising that effects on financial behavior are more modest than effects on financial knowledge. Even so, these effects are classified as "medium" in magnitude in his interpretation of effect sizes realized in RCTs.

6.3 Cost-effectiveness

While understanding effect sizes in standard deviation units is more consistent across educational interventions and more intuitive than "variance explained," a discussion of effect sizes is incomplete without quantifying costs, as also noted in Kraft (2019). Unfortunately, only 20 papers within the 76 studied include a discussion of cost. If we conduct a meta-analysis with only these papers, we find that the estimated treatment effects are smaller in the set of studies reporting costs than in the fully aggregated sample. In Appendix B Figure B6, we regress a binary indictor of reporting costs on sample and experiment characteristics to examine which are the studies that do report costs. The only notable difference is that studies reporting costs are more likely to involve low-income samples. Since we see no difference in effect sizes based on whether or not the intervention was targeted to low-income populations, we cannot precisely say what is driving the difference in effect sizes with respect to studies reporting costs.

To give readers a visual assessment of costs and effect sizes, we report the average costs by study in Appendix A Table A1 in 2019 U.S. dollars. Averaging across all studies reporting costs, the mean and median per participant costs are $60.40 and $22.90, respectively. Using the Kraft (2019) scheme with respect to effect sizes, an average cost of $60 per participant would be classified as a "low cost" educational intervention. It could be that studies reporting costs have, on average, lower costs than those that do not report costs. If that is the case, costs are understated, as are benefits since effect sizes are smaller in the reporting sample. Several studies mention their interventions had "minimal costs" but do not report a number; we do not include these studies in the cost estimates. Some programs may have costs that are difficult to quantify. Other programs may be difficult to scale. For example, Calderone et al. (2018) report a $25 per person cost and $39 per person benefit for a financial education program in India. However, they state the program is still too costly for a large company to implement at scale. While some studies pass a cost-benefit analysis on the surface, there may be other barriers prohibiting implementation.

Overall, our cost-effectiveness ratio is $60.40 per person for one-fifth of a standard deviation improvement in outcomes. Figure 5 displays the cost and effect size by outcome domain for each study. There are two direct takeaways from the figure. First, most effect sizes lie above the zero line but below 0.5 standard deviations. The effects below the zero line largely reflect papers that study the impact of financial education on remittances (e.g., switching to a cheaper financial product when transferring money across countries). Second, there does not appear to be a linear relationship between costs and effect sizes. Figure B7 in Appendix B displays the effect sizes and costs for each outcome domain separately, where we also include 95% confidence intervals for each estimate.

< Figure 5 about here >

To make the discussion more salient, we use one paper that clearly spells out the costs, from a large-scale randomized control trial in Peruvian schools (Frisancho 2018). That paper reports a cost per pupil of $4.80 USD and that a $1 increase in spending on the program yields a 3.3 point improvement in the PISA financial literacy assessment. Since this study represents financial education within a year-long class and average and median interventions in the sample are only 12 and 7 hours, respectively, it is likely that the average effect across studies corresponds to lower costs. Frisancho (2018) also shows that the course does not detract from performance in other courses, limiting opportunity costs.

Our back of the envelope estimate is conservative in that it does not consider positive externalities of the program. For example, Frisancho (2018) documents that in addition to improving student outcomes, teachers' financial literacy and credit scores also increase. Further, Bruhn et al. (2016) document positive "trickle up" effects for parents. Thus, financial education programs may have externalities beyond the target group, such as affecting behaviors of teachers, parents, and possibly peers (Haliassos et al. 2019).

6.4 External validity

While a benefit of only including RCTs is that there is little debate regarding their internal validity, it is more common to study long-term effects in quasi-experimental settings. There exists mounting quasi-experimental evidence that requiring U.S. high school students to complete financial education prior to graduating improves long-term financial behaviors. This body of literature uses a difference-in-difference strategy comparing students who would have graduated just before and just after the requirement was in place within a state with a requirement, as well as across states with and without requirements over the same time period.[18]

High school personal finance graduation requirements, which include standalone courses and personal finance standards incorporated into another required class or curriculum, show that financial education reduces non-student debt (Brown et al. 2016), increases credit scores (Brown et al. 2016; Urban et al. 2018), reduces default rates (Brown et al. 2016; Urban et al. 2018), shifts student loan borrowing from high-interest to low-interest methods (Stoddard

[18] Cole, Paulson, and Shastry (2016) used this method but studied "personal finance mandates" between 1957-1982, which often did not comprise course requirements but instead brought a representative from a bank to give a one-off lecture. The authors documented no effects of the education on investment or credit management behaviors. This was in contrast to Bernheim, Garret, and Maki (2001), who found that these same mandates improved investment behaviors, though they did not include state-level fixed effects in their analysis.

and Urban 2019), increases student loan repayment rates (Mangrum 2019), reduces payday loan borrowing for young adults (Harvey, 2019), and increases bank account ownership for those with only high school education (Harvey 2020). This recent literature as well confirms the findings in the meta-analysis.

7 Conclusions

Our analysis of the existing research on financial education using the most rigorous evaluation methods has three main findings.

First, financial education treatment effects from RCTs have, on average, positive effects on financial knowledge and behaviors. This result is very robust: it holds up to accounting for publication bias, including only adequately powered studies, looking only at studies published in top economics journals, and accounting for heterogeneity across studies. Financial education interventions have sizable effects on both financial knowledge (+0.2 SD units) and financial behaviors (+0.1 SD units). Thus, the treatment effects on financial knowledge are quite similar to or even larger in magnitude than the average effect sizes realized by educational interventions in other domains such as math and reading (see Hill et al. 2008; Cheung and Slavin 2016; Fryer 2016; Kraft 2018) and the effect sizes on financial behaviors are comparable to those realized in behavior-change interventions in the health domain (e.g., Rooney and Murray 1996; Portnoy et al. 2008; Noar et al. 2007) or behavior-change interventions aimed at fostering energy conserving behavior (e.g., Karlin et al. 2015). Our findings are in stark contrast to the findings presented in the first meta-analysis of the financial education literature (Fernandes et al. 2014). How can we interpret these differences in findings? While we are unable to replicate the original result on RCTs presented in Fernandes et al. (2014) (see Appendix D), we observe that the number of recent RCTs added to the database is driving the more positive result of financial education treatment effects on financial knowledge and behaviors. Additionally, we show that explicitly accounting for heterogeneity in studies and programs is crucial in assessing the average impact of financial education.

Second, there is no evidence to support or refute decay of financial education treatment effects six months or more after the intervention. Since only six studies in our sample look at impacts 24 months beyond the intervention, we cannot rule out that this effect is statistically different from short-run effects. Because the present literature is characterized by very few longer-term impact assessments, the evidence on the sustainability of effects is inconclusive. What we can say, however, is that we do not find evidence for dramatic decay up to six months after the intervention.

Third, we document that the estimates of statistical effect sizes are economically significant. We further document that many of the financial education interventions studied in randomized experiments are cost-effective. This finding is crucial, since the discussion of the effectiveness of financial education has focused on statistical effect sizes without considering their economic interpretation.

The evidence in this meta-analysis summarizes financial education interventions from 33 countries and six continents, across the lifespan of individuals. The analysis carefully accounts for heterogeneity across interventions. However, there are still some limitations. Since few RCTs study long-run effects, it is hard to determine the long-run impacts of these interventions. The same is true for the quality of the data used to study changes in financial behaviors: Few studies are able to link their experiments to administrative data, so the usual caveats of having to rely on self-reported survey data also apply to this literature. Future research should aim to collect longer-run administrative data or follow up with original participants from earlier field experiments. Finally, we encourage more studies to report on the costs of their programs, in order to provide policymakers with an estimate of cost-effectiveness.

References

Beuermann, D. W. and C. K. Jackson (2018). The short and long-run effects of attending the schools that parents prefer. Working paper. https://works.bepress.com/c_kirabo_jackson/37/

Bloom, H. S., Hill, C. J., Black, A. R., and Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, 1(4): 289–328.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley. http://dx.doi.org/ 10.1002/9780470743386

Bruhn, M., de Souza Leao, L., Legovini, A., Marchetti, R., and Zia, B. (2016). The impact of high school financial education: Evidence from a large-scale evaluation in Brazil. *American Economic Journal: Applied Economics*, 8(4): 256–295.

Brown, M., Grigsby, J., van der Klaauw, W., Wen, J., and Zafar, B. (2016). Financial education and the debt behavior of the young. *Review of Financial Studies*, 29(9): 2490–2522.

Cheung, A. and Slavin, R. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, 45(5): 283–292

Collins, J. M. and O'Rourke, C. M. (2010). Financial education and counseling - still holding promise. *Journal of Consumer Affairs*, 44 (3): 483–98.

Evans, D. and Yuan, F. (2019). Equivalent years of schooling: A metric to communicate learning gains in concrete terms. *World Bank Policy Research Working Paper No. 8752.*

Fernandes, D., Lynch Jr., J.G., and Netemeyer, R.G. (2014). Financial literacy, financial education, and downstream financial behaviors. *Management Science*, 60(8): 1861–1883.

Fox, J., S. Bartholomae, and Lee, J. (2005). Building the case for financial education. *Journal of Consumer Affairs*, 39 (1): 195–214.

Fryer, R. G. (2016). The production of human capital in developed countries: evidence from 196 randomized field experiments. *NBER Working Paper No. 22130.*

Haliassos, M., Jansson, T., and Karabulut, Y. (2019). Financial literacy externalities. *Review of Financial Studies* 33 (2): 950–989.

Harbord, R. M., Higgins, J. P., et al. (2008). Meta-regression in Stata. *Stata Journal*, 8(4):493519.

Harvey, M. (2019). Impact of financial education mandates on young consumers' use of alternative financial services. *Journal of Consumer Affairs,* forthcoming.

Harvey, M. (2020). Does state-mandated high school financial education affect savings by low-income households? Working paper. https://static1.squarespace.com/static/5c4d314bb27e3999d515a9e4/t/5e0a1b2841180e2960023175/1577720633380/Harvey_FinEd_Savings_Working+Paper_v20191230.pdf

Hastings, J. S., Madrian, B. C., and Skimmyhorn, W. L. (2013). Financial literacy, financial education, and economic outcomes. *Annual Review of Economics*, 5: 347–373.

Hedges, L. V., Tipton, E., and Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1(1): 39–65.

Hedges, L. V. and E. C. Hedberg (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1): 60–87.

Hill, C. J., Bloom, H. S., Black, A. R., and Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3): 172–177.

Kaiser, T. and Menkhoff, L. (2017). Does financial education impact financial behavior, and if so, when? *World Bank Economic Review*, 31(3): 611–630.

Kaiser, T. and Menkhoff, L. (2019). Financial education in schools: A meta-analysis of experimental studies. *Economics of Education Review.* https://doi.org/10.1016/j.econedurev.2019.101930

Karlin, B., Zinger, J. F., and Ford, R. (2015). The effects of feedback on energy conservation: a meta-analysis. *Psychological Bulletin* 141(6): 1205–1227.

Kraft, M. A. (2019). Interpreting effect sizes of education interventions. *Educational Researcher*, forthcoming.

Lipsey, M. W. and Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks: Sage.

Lührmann, M., Serra-Garcia, M., and Winter, J. (2018). The impact of financial education on adolescents' intertemporal choices. *American Economic Journal: Economic Policy*, 10(3): 309–332.

Lusardi, A. and Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. *Journal of Economic Literature*, 52(1): 5–44.

Lusardi, A., Michaud, P.-C., and Mitchell, O. S. (2017). Optimal financial knowledge and wealth inequality. *Journal of Political Economy*, 125(2): 431–477.

Mangrum, D. (2019). Personal finance education mandates and student loan repayment. Working paper. https://www.danielmangrum.com/research.html

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1): 57–91.

Miller, M., Reichelstein, J., Salas, C., and Zia, B. (2015). Can you help someone become financially capable? A meta-analysis of the literature. *World Bank Research Observer*, 30(2): 220–246.

Noar, S. M., Benac, C. N., and Harris, M. S. (2007). Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. *Psychological Bulletin*, 133(4): 673–693.

OECD (2015). National strategies for financial education. OECD/INFE policy handbook, https://www.oecd.org/finance/National-Strategies-Financial-Education-Policy-Handbook.pdf.

Portnoy, D. B., Scott-Sheldon, L. A., Johnson, B. T., and Carey, M. P. (2008). Computer-delivered interventions for health promotion and behavioral risk reduction: A meta-analysis of 75 randomized controlled trials. *Preventive Medicine*, 47(1): 3–16.

Rooney, B. L. and Murray, D. M. (1996). A meta-analysis of smoking prevention programs after adjustment for errors in the unit of analysis. *Health Education Quarterly*, 23(1): 48–64.

Stanley, T. D. (2001). Wheat from chaff: Meta-analysis as quantitative literature review. *Journal of Economic Perspectives*, 15(3): 131–150.

Stoddard, C. and Urban, C. (2019) The effects of financial education graduation requirements on postsecondary financing decisions. *Journal of Money, Credit, and Banking,* forthcoming.

Tanner-Smith, E. E., and Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in STATA and SPSS. *Research Synthesis Methods*, 5(1): 13–30.

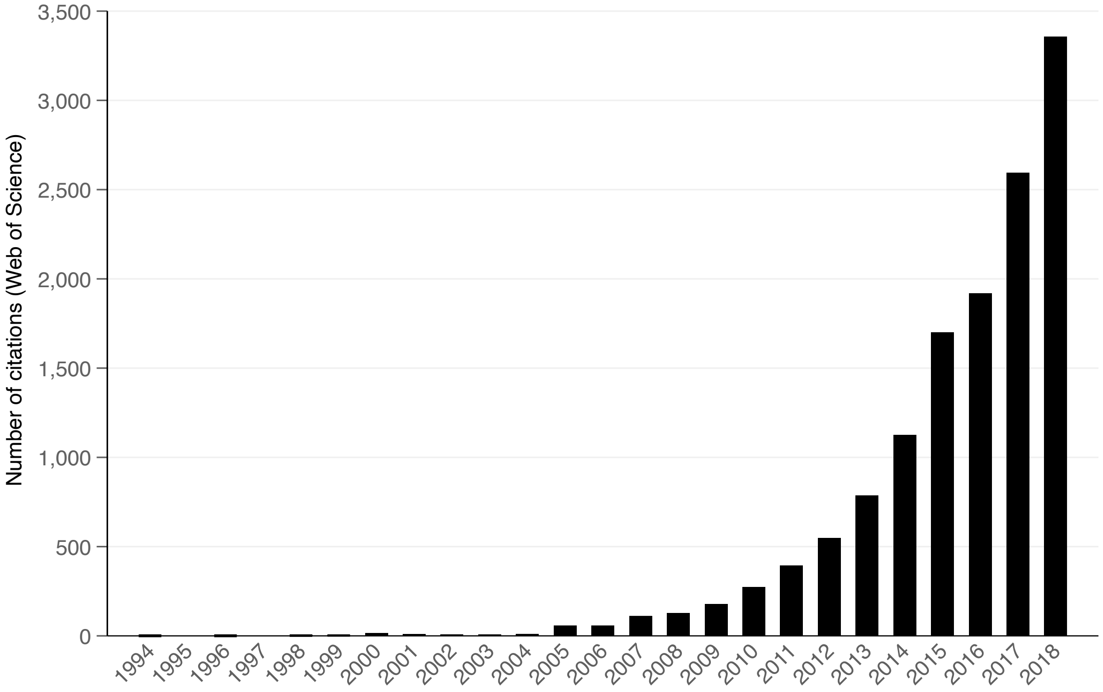
Tanner-Smith, E. E., Tipton, E., and Polanin, J. R. (2016). Handling complex meta- analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology*, 2(1): 85–112.

Urban, C., Schmeiser, M., Collins, J. M., and Brown, A. (2018). The effects of high school personal financial education policies on financial behavior. *Economics of Education Review*, forthcoming.

What Works Clearinghouse. (2014). WWC procedures and standards handbook (Version 3.0). *U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse.*

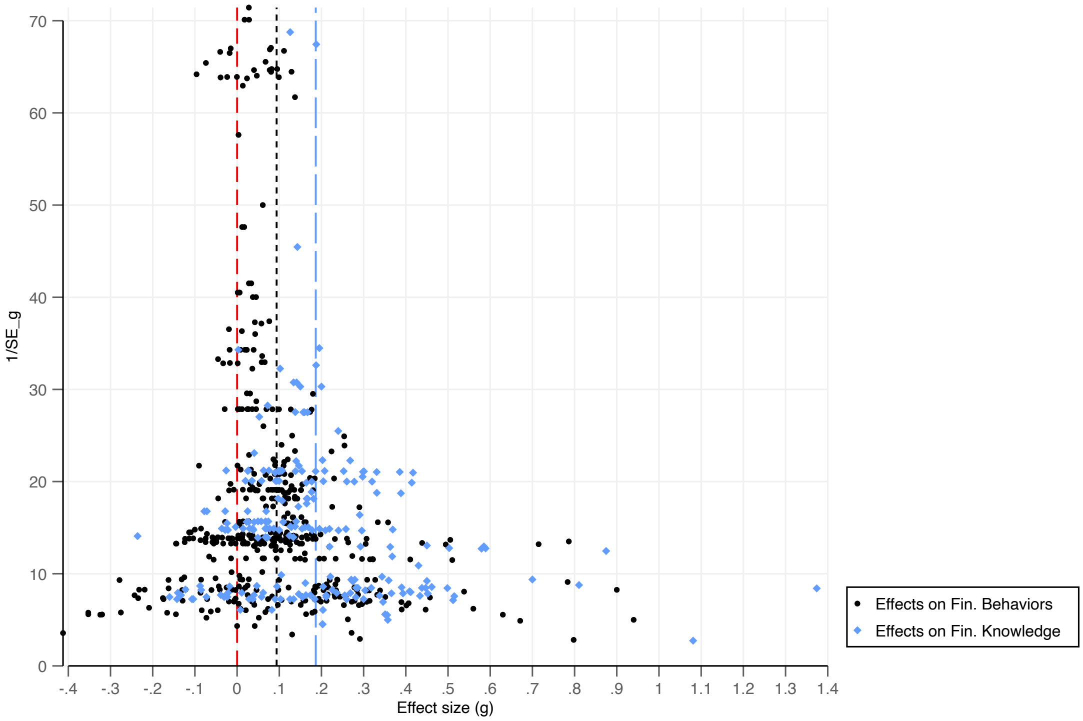
Xu, L., and Zia, B. (2012). Financial literacy around the world: An overview of the evidence with practical suggestions for the way forward. *World Bank Policy Research Working Paper No. 6107.*

Figure 1: Citations in the SSCI to the term "financial literacy" per year



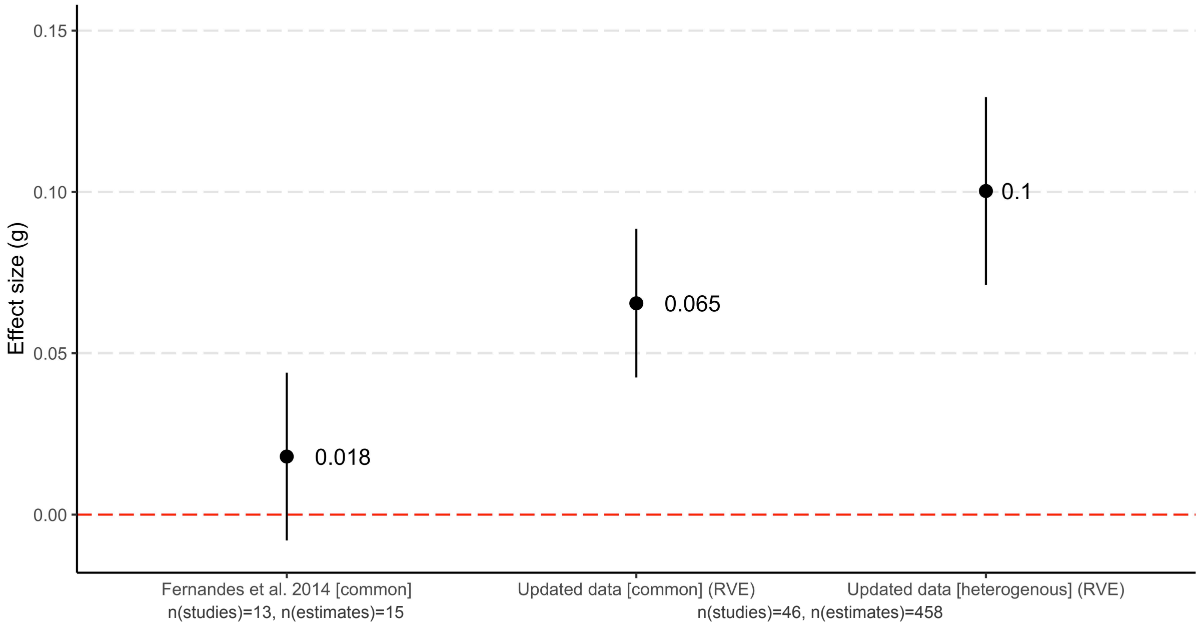
Notes: Number of citations within the social science citation index (Web of Science) to articles including the term "financial literacy" in the title or the abstract. Data from October 11, 2019.

Figure 2: Distribution of raw financial education treatment effects and their standard errors



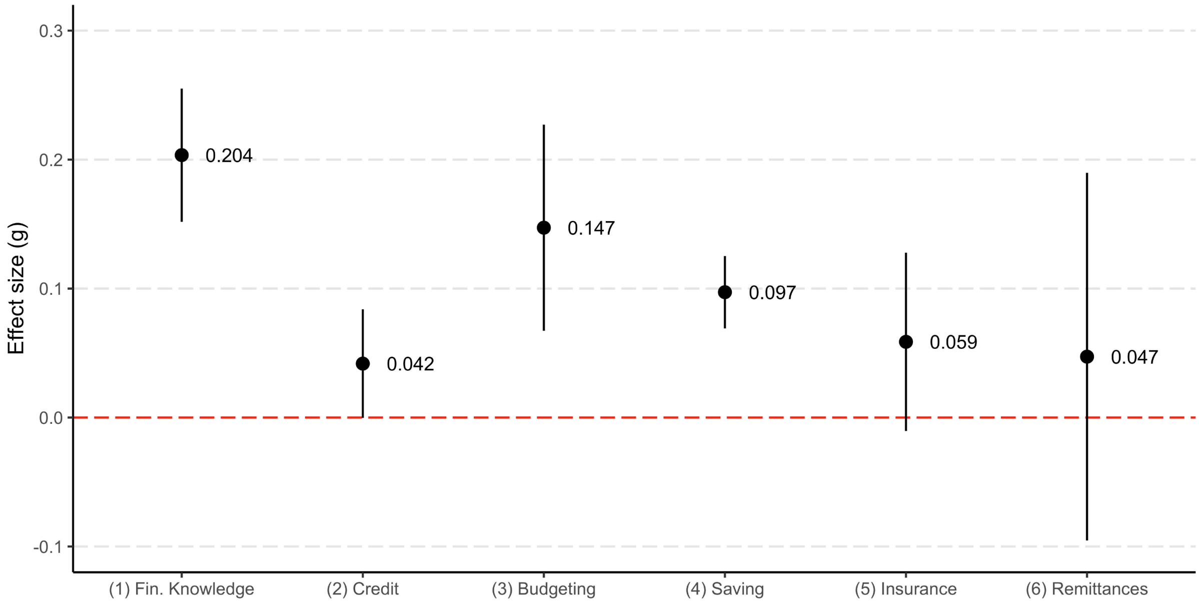
Notes: Effect size (g) is the bias corrected standardized mean difference (Hedges' g). 1/SE_g is its inverse standard error (precision). The number of observations in the treatment effects on financial behaviors sample is 458 effect size estimates from 64 studies. The number of observations in the treatment effects on financial knowledge sample is 215 effect size estimates from 50 studies. Thirty-eight studies report treatment effects on both types of outcomes. The mean effect size on financial behaviors is 0.0937 SD units, and the mean effect size on financial knowledge is 0.186 SD units.

Figure 3: Estimating the average effect of financial education treatment on financial behaviors in RCTs



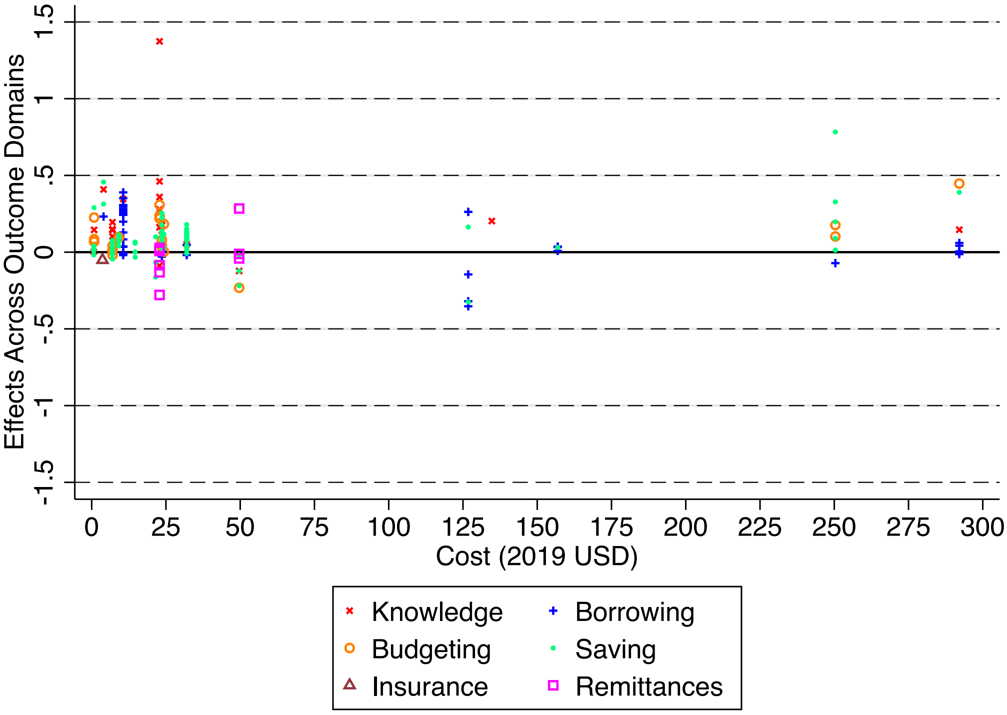
Notes: Fernandes et al. (2014) report weighted least squares estimates with inverse variance weights (common effect assumption). The results with updated data are from robust variance estimation in meta-regression with dependent effect size estimates (RVE) (Hedges et al. 2010) with $\tau^2 = 0$ in the common effect case, and τ^2estimated via methods of moments in the heterogeneous effects case. Fernandes et al. (2014) use within-study average effects and estimate the weighted average effect across 15 observations using inverse variance weights. Our estimates with updated data are based on multiple effect sizes per study and account for the statistical dependency (estimates within studies) by relying on robust variance estimation in meta-regression with dependent effect size estimates (Hedges et al. 2010). Dots show the point estimate, and the solid lines indicate the 95% confidence interval.

Figure 4: Financial education treatment effects by outcome domain



Notes: Results from robust variance estimation in meta-regression with dependent effect size estimates (RVE) (Hedges et al. 2010). The number of observations for the financial knowledge sample (1) is 215 effect size estimates within 50 studies. The number of observations for the credit behavior sample (2) is 115 within 22 studies. The number of effect size estimates for the budgeting behavior sample (3) is 55 within 23 studies. The number of observations in the saving behavior (4) sample is 253 effect size estimates within 54 studies. The number of observations in the insurance behavior sample (5) is 18 effect sizes within six studies. The number of observations on remittance behavior (6) is 17 effect size estimates reported within six studies. Dots show the point estimate, and the solid lines indicate the 95% confidence interval.

Figure 5: Cost of intervention and effect sizes



Notes: The graph depicts the cost and effect sizes for each outcome domain among the 20 experiments that report costs. Each data point is an effect size for an outcome studied. Figure B7 in Appendix B provides a graph for each outcome domain that contains standard errors of the estimates.

Table 1: Descriptive statistics

Variable	Obs.	Mean	Median	Std. Dev.	Min.	Max.
Hedges' *g*	677	0.123	0.098	0.183	-0.413	1.374
SE (*g*)	677	0.084	0.072	0.049	0.007	0.365
Time span (in weeks)	643	30.239	25.800	31.537	0.000	143.550
Intensity (in hours)	604	11.709	7.000	16.267	0.008	108.000
Mean age (in years)	650	33.480	38.300	12.480	8.500	55.000
Children (< age 14)	677	0.075	-	-	0.000	1.000
Youth (age 14 to 25)	677	0.201	-	-	0.000	1.000
Adults (> age 25)	677	0.724	-	-	0.000	1.000
Low income (yes=1)	677	0.725	-	-	0.000	1.000
Developing economy (yes=1)	677	0.604	-	-	0.000	1.000
Top econ journal (yes=1)	677	0.267	-	-	0.000	1.000

Note: Descriptive statistics at the estimate-level, i.e. we consider the total of 677 effects reported in 76 RCTs.

Table 2: Financial education treatment effects by subgroups of studies and populations

Subgroup	Effect size (g)	SE	95% CI Lower bound	95% CI Upper bound	n(Studies)	n(effects)
Panel A: Treatment effects on *financial behaviors*						
(a) By country income						
High income economies	0.1127	0.0316	0.0478	0.1777	32	129
Developing economies	0.0928	0.0130	0.0660	0.1195	32	329
(b) By respondent income						
Low income individuals	0.0993	0.0194	0.0600	0.1387	43	367
General population	0.1035	0.0219	0.0571	0.1500	21	91
(c) By age of participants						
Children (< age 14)	0.0640	0.0186	0.0188	0.1091	9	36
Youth (age 14 to 25)	0.1203	0.0415	0.0250	0.2155	11	92
Adults (> age 25)	0.1068	0.0205	0.0653	0.1483	44	330
(d) By type of publication						
Top econ. journals	0.0833	0.0235	0.0325	0.1342	15	161
Other publications	0.1075	0.0183	0.0704	0.1445	49	297
(e) By delay between treatment and measurement of outcomes						
Delay of < 6 months	0.0991	0.0169	0.0645	0.1337	34	180
Delay of ≥ 6 months	0.0710	0.0137	0.0425	0.0995	28	260
Delay of ≥ 12 months	0.0878	0.0200	0.0450	0.1308	18	134
Delay of ≥ 18 months	0.0653	0.0192	0.0209	0.1098	10	49
Delay of ≥ 24 months	0.0574	0.0225	0.0013	0.1136	7	32
Panel B: Treatment effects on *financial knowledge*						
(a) By country income						
High income economies	0.2591	0.0415	0.1738	0.3443	29	135
Developing economies	0.1392	0.0218	0.0934	0.1851	21	80
(b) By respondent income						
Low income individuals	0.2238	0.0395	0.1428	0.3049	30	120
General population	0.1835	0.0310	0.1183	0.2486	20	95
(c) By age of participants						
Children (< age 14)	0.2763	0.1098	0.0076	0.5450	7	15
Youth (age 14 to 25)	0.1859	0.0390	0.1015	0.2703	16	40
Adults (> age 25)	0.2001	0.0282	0.1418	0.2583	28	160
(d) By type of publication						
Top econ. journals	0.1572	0.0379	0.0648	0.2497	8	46
Other publications	0.2142	0.0299	0.1537	0.2746	42	169
(e) By delay between treatment and measurement of outcomes						
Delay of < 6 months	0.2305	0.0319	0.1654	0.2956	36	142
Delay of ≥ 6 months	0.1408	0.0289	0.0775	0.2041	15	57
Delay of ≥ 12 months	0.1406	0.0367	0.0166	0.2646	5	5
Delay of ≥ 18 months	-	-	-	-	0	0
Delay of ≥ 24 months	-	-	-	-	0	0

Notes: This table reports average effects of financial education treatment on financial behaviors (Panel A) and financial knowledge (Panel B) estimated via RVE. Ten studies with 34 effect size estimates are missing information about the delay between treatment and measurement of outcomes.

Appendix

(online appendix not intended for print publication)

to accompany

“Financial education affects financial knowledge and downstream behaviors”

Appendix A: Included studies

Table A1: Overview of included experiments

	Experiment	Country	Sample mean age	Sample Size	Outcomes	Cost
1	Abarcar et al. (2018)	Philippines	42	1,808	A, B, D	NR
2	Abebe et al. (2018)	Ethiopia	37	508	A, B, D	NR
3	Alan and Ertac (2018)	Turkey	9	1,970	D	NR
4	Ambuehl et al. (2014)	USA	29	504	A	NR
5	Angel (2018)	Austria	18	296	A, D	NR
6	Attanasio et al. (2019)	Colombia	39	3,136	A, B, C, D	23.6
7	Barcellos et al. (2016)	USA	51	370	A, D	NR
8	Barua et al. (2012)	Singapore	37	408	A, C, D, F	43.5
9	Batty et al. (2015) [independent sample 1]	USA	9	703	A, C, D	NR
10	Batty et al. (2015) [independent sample 2]	USA	9	277	A, C, D	NR
11	Batty et al. (2017)	USA	9	1,972	A, C, D	NR
12	Becchetti and Pisani (2012)	Italy	18	3,820	A	NR
13	Becchetti et al. (2013)	Italy	18	1,063	A, D	NR
14	Berg and Zia (2017)	South Africa	32	1,031	A, B, D	NR
15	Berry et al. (2018)	Ghana	11	5,400	A, B, D	0.62
16	Bhattacharya et al. (2016)	USA	15	84	A	121.5
17	Bhutoria and Vignoles (2018)	India	32	1,281	A, C, D	0.76
18	Billari et al. (2017)	Italy	44	1,436	A	NR
19	Bjorvatn and Tungodden (2010)	Tanzania	39	211	A	NR
20	Bonan et al. (2016)	Senegal	52	360	E	3.15
21	Bover et al. (2018)	Spain	15	3,070	A, D	NR
22	Boyer et al. (2019)	Canada	44	3,005	A, D	NR
23	Brugiavini et al. (2015) [independent sample 1]	Italy	23	104	A, D	NR
24	Brugiavini et al. (2015) [independent sample 2]	Italy	23	642	A, D	NR
25	Bruhn and Zia (2013)	Bosnia and Herzegovina	28	445	A, B, C, D	245
26	Bruhn et al. (2016)	Brazil	16	25,000	A, B, C, D	NR
27	Bruhn et al. (2014)	Mexico	33	2,178	A, B, D	NR
28	Calderone et al. (2018)	India	45	3,000	A, B, D	28
29	Carpena et al. (2017)	India	39	1,328	A, B, C, D, E	NR
30	Carter et al. (2016)	Mozambique	46	1,534	B, D	NR
31	Choi et al. (2010) [independent sample 1]	USA		391	D	NR
32	Choi et al. (2010) [independent sample 2]	USA		252	D	NR
33	Choi et al. (2010) [independent sample 3]	USA		87	D	NR
34	Clark et al. (2014)	USA	35	4,111	D	NR
35	Cole et al. (2013)	India	48	1,047	E	NR
36	Cole et al. (2011)	Indonesia	41	564	D	17
37	Collins (2013)	USA	39	144	B, D	100
38	Collins and Urban (2016)			1,001	B, C, D	210
39	Custers (2011)	India	34	667	A	NR
40	Doi et al. (2014)	Indonesia	44	400	A, D, F	NR
41	Drexler et al. (2014)	Dominican Republic	41	1,193	C, D	19.6
42	Duflo and Saez (2003)	USA	38	4,879	D	9.8
43	Elbogen et al. (2016)	USA	NA (adults)	184	A, D	NR
44	Field et al. (2010)	India	32	597	B, D	NR
45	Flory (2018)	Malawi	41	2,011	D	NR
46	Frisancho (2018)	Peru	15	25,980	A, C, D	6.6
47	Furtado (2017)	Brazil	12	14,655	A, D	NR
48	Gaurav et al. (2011)	India	50	597	E	NR
49	Gibson et al. (2014)	New Zealand	NA (adults)	344	A, C, F	22.9

	[independent sample 1]					
50	Gibson et al. (2014) [independent sample 2]	New Zealand	NA (adults)	352	A, C, F	22.9
51	Gibson et al. (2014) [independent sample 3]	Australia	NA (adults)	209	A, C, F	NR
52	Gine and Mansuri (2013)	Pakistan	38	3,494	B, D	126
53	Gine et al. (2013)	Kenya	49	904	E	NR
54	Han et al. (2009)	USA	41	840	D	NR
55	Haynes et al. (2011)	USA	55	228	A	NR
56	Heinberg et al. (2014)	USA	35	2,920	A	NR
57	Hetling et al. (2016)	USA	36	300	B	NR
58	Hinojosa et al. (2010)	USA	9 / 15	8,594	A	NR
59	Jamison et al. (2014)	Uganda	25	2,810	A, B, C, D	NR
60	Kaiser and Menkhoff (2018)	Uganda	36	1,291	A, B, C, D, E	NR
61	Kajwij et al. (2017)	Netherlands	10	2,321	A, D	NR
62	Lusardi et al. (2017)	USA	50	892	A	NR
63	Lührmann et al. (2018)	Germany	14	914	A, D	NR
64	Migheli and Moscarola (2017)	Italy	9	213	D	NR
65	Mills et al. (2004)	USA	36	840	B, D	NR
66	Modestino et al. (2019)	USA	24	300	A, B	10
67	Postmus et al. (2015)	USA	38	195	B	NR
68	Reich and Berman (2015)	USA	30	33	A, B, D	NR
69	Sayinzoga et al. (2016)	Rwanda	40	341	A, B, D	3.5
70	Seshan and Yang (2014)	Qatar	40	232	D, F	NR
71	Shephard et al. (2017)	Rwanda	15	1,750	A, C, D	NR
72	Skimmyhorn et al. (2016)	USA	19	991	A	NR
73	Song (2012)	China	45	1,104	A, D	NR
74	Seinert et al. (2018)	South Africa	49	552	B, D	NR
75	Supanataroek et al. (2016)	Uganda	13	1,746	C, D	8
76	Yetter and Suiter (2015)	USA	24	1,982	A	NR

Notes: Costs are converted to 2019 USD. NR denotes that the costs are not reported in the paper.

Table A2: Extracted estimates by country of financial education intervention

Country	Number of estimates	Percent
Australia	7	1.03
Austria	6	0.89
Bosnia and Herzegovina	8	1.18
Brazil	29	4.28
Canada	4	0.59
China	16	2.36
Colombia	28	4.14
Dominican Republic	4	0.59
Ethiopia	16	2.36
Germany	10	1.48
Ghana	7	1.03
India	123	18.17
Indonesia	30	4.43
Italy	14	2.07
Kenya	1	0.15
Malawi	3	0.44
Mexico	7	1.03
Mozambique	13	1.92
Netherlands	2	0.3
New Zealand	18	2.66
Pakistan	4	0.59
Peru	28	4.14
Philippines	22	3.25
Qatar	6	0.89
Rwanda	8	1.18
Senegal	1	0.15
Singapore	8	1.18
South Africa	14	2.07
Spain	8	1.18
Tanzania	1	0.15
Turkey	13	1.92
USA	185	27.33
Uganda	33	4.87
Total	677	100

Table A3: Types of outcomes coded

	Outcome category	Definition	Freq.
A	*Financial knowledge (+)*	Raw score on financial knowledge test Indicator of scoring above a defined threshold Indicator of solving a test item correctly	215 (31.76%)
B	*Credit behavior*		119 (17.58%)
	1) Reduction of loan default within a certain time-frame (+)	Binary indicator	
	2) Reduction of delinquencies within certain time frame (+)	Binary indicator	
	3) Better credit score (+)	Continuous measure of credit score	
	4) Reduction in informal borrowings (+)	Binary indicator of informal loan or reduction in number of informal loans	
	5) Lower cost of credit / interest rate (+)	Sum of real interest amount or interest rate and (if applicable) cost of fees	
	6) Any debt (-) / (+) (depending on intervention goal)	Binary indicator	
	7) Any formal loan (+)	Binary indicator	
	8) Total amount borrowed (-) / (+) (depending on intervention goal)	Continuous measure (or log) of borrowed amount	
	9) Outstanding debt (-) / (+) (depending on intervention goal, e.g. loan repayment)	Continuous measure of total debt or percentage repaid over time period	
	10) Borrowing index (+)	Study-specific index of survey items to measure borrowing amount, frequency, and repayment	
	11) Uses credit card up to limit (-)	Binary indicator	
	12) Take-up of formal loan (as opposed to informal loan)	Binary indicator	
	13) Reduction in borrowing for consumption (+)	Binary indicator or loan amount	
	14) Increase in borrowing for productive purposes (+)	Binary indicator or loan amount	
C	*Budgeting behavior*		55 (8.12 %)
	1) Having a written budget (+)	Binary indicator	
	2) Positive sentiment toward budgeting (+)	Binary indicator	
	3) Having a financial plan or long-term aspirations (+)	Binary indicator	
	4) Keeping separate records for business and household (+)	Binary indicator	
	5) Seeking information before making financial decisions (+)	Binary indicator	
	6) Self-rating of adherence to budget (+)	Study-specific scale	
D	*Saving & retirement saving behavior*		253 (57.46 %)
	1) Amount of savings (+)	Continuous measure (or log) of savings amount (in currency or number of valuable assets) or categorical variable indicating amount within range	
	2) Savings rate or savings within timeframe (+)	Savings relative to income Amount over defined time-frame	
	3) Savings index (+)	Study-specific index of survey items designed to measure savings amount and frequency	
	4) Any savings (+)	Binary indicator	
	5) Has formal bank (savings) account (+)	Binary indicator	
	6) Investments into own or other business (stocks) (+)	Continuous measure of amount invested	
	7) Holds any stocks or bonds (+)	Binary indicator	
	8) Has any retirement savings (+)	Binary indicator	
	9) Participating in retirement savings plan (e.g. 401k) (+)	Binary indicator	
	10) Amount of retirement savings (+)	Continuous measure of retirement savings amount	
	11) Retirement savings rate (+)	Retirement savings relative to income	
	12) Positive sentiment towards investing in (retirement-) funds (+)	Binary indicator or rating-scale	
	13) Reduction of excess risk in retirement fund (+)	Continuous measure of retirement savings amount allocated to risky assets	
	14) Reduction of cost of savings product (fees / taxes paid) (+)	Continuous measure of fee amount paid / estimate of welfare loss	
	15) Contribution rate to retirement savings plan (+)	Indicator of increase or continuous measure of amount increase	
	16) Net wealth (+)	Continuous measure of net wealth	

	17)	Amount saved in allocation task (+)	Continuous measure of amount saved in allocation task	
	18)	Amount allocated to delayed payment date in experimental elicitation task (+)	Continuous measure of amount delayed to be paid out at a later date within an experimental elicitation task	
	19)	Meeting savings goals (+)	Meeting a pre-defined savings goal (survey response)	
	20)	Reduction in spending on temptation goods (+)	Continuous measure or relative measure (to income) of amount spent on temptation goods (e.g. alcohol, tobacco)	
E	*Insurance behavior*			18 (2.51 %)
	1)	Any formal insurance (+)	Binary indicator	
F	*Remittance behavior*			17 (2.56 %)
	1)	Lower cost of remittance product (+)	Continuous measure of cost or binary choice of lower cost product	
	2)	Lower remittance frequency and higher amount (lower cost) (+)	Measure of remittance frequency within timeframe and continuous amount remitted	
	3)	More control over remitted funds (+)	Study-specific scale to measure control over remitted amount	

Notes: When necessary, outcomes are reverse-coded so that positive signs reflect positive financial education treatment effects (e.g., when the dependent variable is coded as the probability of default, we transform this to the reduction in probability of default in order to be able to assign a positive sign reflecting desirable treatment effects).

References of included experiments

1) Abarcar, P., Barua, R., and Yang, D. (2018). Financial education and financial access in transnational households: Field experimental evidence from the Philippines, *Economic Development and Cultural Change*, forthcoming.
2) Abebe, G., Tekle, B., and Mano, Y. (2018). Changing saving and investment behaviour: The impact of financial literacy training and reminders on micro-businesses. *Journal of African Economies*, 27(5), 587–611.
3) Alan, S. and Ertac, S. (2018). Fostering patience in the classroom: Results from randomized educational intervention. *Journal of Political Economy*, 126(5), 1865–1911.
4) Ambuehl, S., Bernheim, B. D., and Lusardi, L. (2014). The effect of financial education on the quality of decision making. *NBER Working Paper 20618.*
5) Angel, S. (2018). Smart tools? A randomized controlled trial on the impact of three different media tools on personal finance. *Journal of Behavioral and Experimental Economics*, 74, 104–111.
6) Attanasio, O., Bird, M., Cardona-Sosa, L., and Lavado, P. (2019). Freeing financial education via tablets: Experimental evidence from Colombia. *NBER Working Paper No. w25929.*
7) Barcellos, S. H., Carvalho, L. S., Smith, J. P., and Yoong, J. (2016). Financial education interventions targeting immigrants and children of immigrants: Results from a rRandomized control trial. *Journal of Consumer Affairs*, 50(2), 263–285.
8) Barua, R., Shastry, G.K., and Yang, D. (2012). Evaluating the effect of peer-based financial education on savings and remittances for foreign domestic workers in Singapore. Working Paper. Singapore Management University, Wellesley College, and University of Michigan.
9) Batty, M., Collins, J.M., and Odders-White, E. (2015). Experimental evidence on the effects of financial education on elementary school students' knowledge, behavior, and attitudes. *Journal of Consumer Affairs*, 49(1): 69–96. [Independent Sample 1]
10) Batty, M., Collins, J.M., and Odders-White, E. (2015). Experimental evidence on the effects of financial education on elementary school students' knowledge, behavior, and attitudes. *Journal of Consumer Affairs*, 49(1): 69–96. [Independent Sample 2]
11) Batty, M., Collins, M., O'Rourke, C. and Elizabeth, O. (2017). Evaluating Experiential Financial Capability Education. A Field Study of My Classroom Economy. *Working Paper.*
12) Becchetti, L. and Pisani, F. (2012). Financial education on secondary school students: The randomized experiment revisited. Facolta di Economia di Forli, Working Paper No. 98.
13) Becchetti, L., Caiazza, S., and Coviello, D. (2013). Financial education and investment attitudes in high schools: Evidence from a randomized experiment. *Applied Financial Economics*, 23(10): 817–836.
14) Berg, G. and Zia, B. (2017). Harnessing emotional connections to improve financial decisions. Evaluating the impact of financial education in mainstream media. *Journal of the European Economic Association*, 15(5): 1025–1055.
15) Berry, J., Karlan, D., and Pradhan, M. (2018). The impact of financial education for youth in Ghana. *World Development*, 102: 71–89.
16) Bhattacharya, R., Gill, A., and Stanley, D. (2016). The effectiveness of financial literacy instruction: The role of individual development accounts participation and the intensity of instruction. *Journal of Financial Counseling and Planning*, 27(1): 20–35.

17) Bhutoria, A. and Vignoles, A. (2018). Do financial education interventions for women from poor Hhuseholds impact their financial behaviors? Experimental evidence from India. *Journal of Research on Educational Effectiveness*, 11(3): 409–432.

18) Billari, F. C., Favero, C. A. and Saita, F. (2017). Nudging financial and demographic literacy: Experimental evidence from an Italian pension fund (November 1, 2017). BAFFI CAREFIN Centre Research Paper No. 67. Available at SSRN: https://ssrn.com/abstract=3095919 or http://dx.doi.org/10.2139/ssrn.3095919

19) Bjorvatn, K, and Tungodden, B. (2010). Teaching business in Tanzania: Evaluating participation and performance. *Journal of the European Economic Association*, 8 (2-3): 561–570.

20) Bonan, J., Dagnelie, O., LeMay-Boucher, P., and Tenikue, M. (2016). The impact of insurance literacy and marketing treatments on the demand for health microinsurance in Senegal: A randomised evaluation. *Journal of African Economies*, 26(2): 169-191.

21) Bover, O., Hospido, L., and Villanueva, E. (2018). The impact of high school financial education on financial knowledge and choices: Evidence from a randomized trial in Spain. *IZA Discussion Papers 11265.*

22) Boyer, M. Martin, d'Astous, P. and Michaud, P.C. (2019). Tax-sheltered retirement accounts: Can financial education improve decisions? *NBER Working Paper No. 26128.*

23) Brugiavini, A., Cavapozzi, D., Padula, M., and Pettinicchi, Y. (2015). Financial education, literacy and investment attitudes. SAFE Working Paper No. 86. University of Venice and SAFECenter, University of Frankfurt. [independent Sample 1]

24) Brugiavini, A., Cavapozzi, D., Padula, M., and Pettinicchi, Y. (2015). Financial education, literacy and investment attitudes. SAFE Working Paper No. 86. University of Venice and SAFECenter, University of Frankfurt. [independent Sample 2]

25) Bruhn, M. and Zia, B. (2013). Stimulating managerial capital in emerging markets: The impact of business training for young entrepreneurs. *Journal of Development Effectiveness*, 5(2): 232–266.

26) Bruhn, M., de Souza Leao, L., Legovini, A., Marchetti, R., and Zia, B. (2016). The impact of high school financial education: Evidence from a large-scale evaluation in Brazil. *American Economic Journal: Applied Economics*, 8(4): 256–295.

27) Bruhn, M., Ibarra, G.L. and McKenzie, D. (2014). The minimal impact of a large-scale financial education program in Mexico city. *Journal of Development Economics*, 108: 184–189.

28) Calderone, M., Fiala, N., Mulaj, F. Sadhu, S., and Sarr, L. (2018). Financial education and savings behavior: Evidence from a randomized experiment among low income clients of branchless banking in India. *Economic Development and Cultural Change*, forthcoming.

29) Carpena, F., Cole, S., Shapiro, J., and Zia, B. (2017). The ABCs of financial education. Experimental evidence on attitudes, behavior, and cognitive biases. *Management Science*, https://doi.org/10.1287/mnsc.2017.2819.

30) Carter, M.R., Laajaj, R., and Yang, D. (2016). Savings, subsidies, and technology adoption: Field experimental evidence from Mozambique. Unpublished working paper.

31) Choi, J.J., Laibson, D., and Madrian, B.C. (2010). Why does the law of one price fail? An experiment on index mutual funds. *Review of Financial Studies*, 23(4): 1405–1432 [independent sample 1].

32) Choi, J.J., Laibson, D., and Madrian, B.C. (2010). Why does the law of one price fail? An experiment on index mutual funds. *Review of Financial Studies*, 23(4): 1405–1432 [independent sample 2].

33) Choi, J.J., Laibson, D., and Madrian, B.C. (2010). Why does the law of one price fail? An experiment on index mutual funds. *Review of Financial Studies*, 23(4): 1405–1432 [independent sample 3].

34) Clark, R.L., Maki, J.A., and Morrill, M.S. (2014). Can simple informational nudges increase employee participation in a 401(k) plan? *Southern Economic Journal*, 80(3): 677–701.

35) Cole, S., Gine, X., Tobacman, J., Topalova, P., Townsend, R., and Vickery, J. (2013). Barriers to household risk management: Evidence from India. *American Economic Journal: Applied Economics*, 5(1): 104–135.

36) Cole, S., Sampson, T., and Zia, B. (2011). Prices or knowledge? What drives demand for financial services in emerging markets? *Journal of Finance*, 66(6): 1933–1967.

37) Collins, J.M. (2013). The impacts of mandatory financial education: Evidence from a randomized field study. *Journal of Economic Behavior and Organization*, 95: 146–158.

38) Collins, J. M., and Urban, C. (2016). The role of information on retirement planning: Evidence from a field study. *Economic Inquiry,* 54(4): 1860–1872.

39) Custers, A. (2011). Furthering financial literacy: Experimental evidence from a financial literacy program for microfinance clients in Bhopal, India. LSE International Development Working Paper 11-113, London.

40) Doi, Y., McKenzie, D., and Zia, B. (2014). Who you train matters: Identifying combined effects of financial education on migrant households. *Journal of Development Economics*, 109: 39–55.

41) Drexler, A., Fischer, G., and Schoar, A. (2014). Keeping it simple: Financial literacy and rules of thumb. *American Economic Journal: Applied Economics*, 6(2): 1–31.

42) Duflo, E. and Saez, E. (2003). The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *Quarterly Journal of Economics*, 118(3): 815–842.

43) Elbogen, E. B., Hamer, R. M., Swanson, J. W., and Swartz, M. S. (2016). A randomized clinical trial of a money management intervention for veterans with psychiatric disabilities. *Psychiatric Services*, 0(0):appi.ps.201500203. PMID: 27181733.

44) Field, E., Jayachandran, S., and Pande, R. (2010). Do traditional institutions constrain female entrepreneurship? A field experiment on business training in India. *American Economic Review: Papers and Proceedings*, 100(2): 125–29.

45) Flory, J. A. (2018). Formal finance and informal safety nets of the poor: Evidence from a savings field experiment. *Journal of Development Economics*, 135: 517–533.

46) Frisancho, V. (2018). The Impact of School-Based Financial Education on High School Students and their Teachers: Experimental Evidence from Peru. *Inter-American Development Bank Working Paper No. 871.*

47) Furtado, I., Legovini, A., and Piza, C. (2017). How early one should start financial education: Evidence from a large-scale experiment. World Bank, DIME Financial and PSD Program in Brief.

48) Gaurav, S., Cole, S., and Tobacman, J. (2011). Marketing complex financial products in emerging markets: Evidence from rainfall insurance in India. *Journal of Marketing Research*, 48(SPL): S150–S162.

49) Gibson, J., McKenzie, D., and Zia, B. (2014). The impact of financial literacy training for migrants. *World Bank Economic Review*, 28(1): 130–161. [Independent Sample 1]

50) Gibson, J., McKenzie, D., and Zia, B. (2014). The impact of financial literacy training for migrants. *World Bank Economic Review*, 28(1): 130–161. [Independent Sample 2]

51) Gibson, J., McKenzie, D., and Zia, B. (2014). The impact of financial literacy training for migrants. *World Bank Economic Review*, 28(1): 130–161. [Independent Sample 3]

52) Gine, X. and Mansuri, G. (2014). Money or ideas? A field experiment on constraints to entrepreneurship in rural Pakistan. World Bank Policy Research Working Paper 6959.

53) Gine, X., Karlan, D., and Ngatia, M. (2013). Social networks, financial literacy and index insurance. World Bank, Washington, DC.

54) Han, C.-K., Grinstein-Weiss, M., and Sherraden, M. (2009). Assets beyond savings in individual development accounts. *Social Service Review*, 83(2): 221–244.

55) Haynes, D.C., Haynes, G., and Weinert, C. (2011). Outcomes of on-line financial education for chronically ill rural women. *Journal of Financial Counseling and Planning*, 22(1): 3–17.

56) Heinberg, A., Hung, A.A., Kapteyn, A., Lusardi, A., Samek, A.S., and Yoong, J. (2014). Five steps to planning success. Experimental evidence from U.S. households. *Oxford Review of Economic Policy*, 30(4): 697-724.

57) Hetling, A., Postmus, J. L., and Kaltz, C. (2016). A randomized controlled trial of a financial literacy curriculum for survivors of intimate partner violence. *Journal of Family and Economic Issues*, 37(4): 672-685.

58) Hinojosa, T., Miller, S., Swanlund A, Hallberg K, Brown, M., and O'Brien, B. (2010). The impact of the stock market game on financial literacy and mathematics achievement. Results from a national randomized controlled trial. Working Paper. Evanston, IL: Society for Research on Educational Effectiveness.

59) Jamison, J.C., Karlan, D, and Zinman, J. (2014). Financial education and access to savings accounts: Complements or substitutes? Evidence from Ugandan youth clubs. *NBER Working Paper 20135.*

60) Kaiser, T. and Menkhoff, L. (2018). Active learning fosters financial behavior: Experimental evidence. *DIW Discussion Paper No. 1743.*

61) Kalwij, A.S., Alessie, R., Dinkova, M., Schonewille, G., van der Schors, A., and van der Werf, M. (2017). The effects of financial education on financial literacy and savings behavior: Evidence from a controlled field experiment in Dutch primary schools. *Working Papers 17-05, Utrecht School of Economics.*

62) Lührmann, M., Serra-Garcia, M., and Winter, J. (2018). The impact of financial education on adolescents' intertemporal choices. *American Economic Journal: Economic Policy*, 10(3), 309–332.

63) Lusardi, A., Samek, A.S., Kapteyn, A., Glinert, L., Hung, A., and Heinberg, A. (2017a). Visual tools and narratives: New ways to improve financial literacy. *Journal of Pension Economics and Finance*, 16(3): 297–323.

64) Migheli, M. and Coda Moscarola, F. (2017). Gender differences in financial education: Evidence from primary school. *De Economist*, 165(3), 321–347.

65) Mills, G., Patterson, R. Orr, L., and DeMarco, D. (2004). Evaluation of the American dream demonstration: Final evaluation report, Cambridge, MA.

66) Modestino, A.S., Sederberg, R., and Tuller, L. (2019). Assessing the effectiveness of financial coaching: Evidence from the Boston Youth Credit Building Initiative. *Unpublished Working Paper*.

67) Postmus, J. L, Hetling,A., and Hoge, G. L. (2015). Evaluating a financial education curriculum as an intervention to improve financial behaviors and financial well-being of survivors of domestic violence: Results from a longitudinal randomized controlled study. *Journal of Consumer Affairs,* 49 (1): 250–66.

68) Reich, C. M. and Berman, J.S. (2015). Do financial literacy classes help? An experimental assessment in a low-income population. *Journal of Social Service Research*, 41(2): 193–203.

69) Sayinzoga, A., Bulte, E. H., and Lensink, R. (2016). Financial literacy and financial behaviour: Experimental evidence from rural Rwanda. *Economic Journal*, 126(594): 1571–1599.

70) Seshan, G. and Yang, D. (2014). Motivating migrants: A field experiment on financial decision-making in transnational households. *Journal of Development Economics,* 108: 119–127.

71) Shephard, D. D., Kaneza, Y. V., and Moclair, P. (2017). What curriculum? Which methods? A cluster randomized controlled trial of social and financial education in Rwanda. *Children and Youth Services Review*, 82: 310–320.

72) Skimmyhorn, W. L., Davies, E. R., Mun, D., and Mitchell, B. (2016). Assessing financial education methods: Principles vs. rules-of-thumb approaches. *Journal of Economic Education*, 47(3): 193–210.

73) Song, C. (2012). Financial illiteracy and pension contributions: A field experiment on compound interest in China. *Unpublished Manuscript.*

74) Steinert, J. I., Cluver, L. D., Meinck, F., Doubt, J., and Vollmer, S. (2018). Household economic strengthening through financial and psychosocial programming: Evidence from a field experiment in South Africa. *Journal of Development Economics*, 134: 443–466.

75) Supanantaroek, S., Lensink, R., and Hansen, N. (2016). The impact of social and financial education on savings attitudes and behavior among primary school children in Uganda. *Evaluation Review*, forthcoming.

76) Yetter, E.A. and Suiter, M. (2015). Financial literacy in the community college classroom: A curriculum intervention study. *Federal Reserve Bank of St. Louis Working Paper 2015-001.*

Appendix B: Considering alternative models, publication bias, power, and additional cost results

We complement our analysis presented in the main text by comparing the estimation results from the random-effects model to alternative approaches to meta-analysis (see Figure B1).

The first row of figure B1 repeats the results from the random-effects model (RVE) discussed in the main text of the manuscript. Panel A shows the effect on financial behaviors (0.1SD units) and Panel B shows the results for treatment effects on financial knowledge (0.2 SD units). We probe the robustness of this result by changing the assumed within-study correlation of estimates (see Figures B2 and B3). The results are identical irrespective of the assumed correlation.

Row 2 of figure B1 reports an unweighted average effect of financial education by estimating an ordinary least squares (OLS) model where each study contributes multiple effect sizes (see Kaiser and Menkhoff 2017; Card et al., 2017 for such an approach). We cluster the standard errors at the study level. This approach represents a description about the literature to date, without inferring an estimate of a possible true effect of financial education in the broader set of possible studies. The results are similar to the random-effects model reported in row 1. Rows 3 and 4 show results from a fixed effects approach to meta-analysis. This corresponds to the same model as in row 2 but weights each effect size estimate by its inverse standard error or the inverse variance, respectively. This unrestricted weighted least squares (WLS) estimation is advocated by Stanley and Doucouliagos (2012, 2015). Effect sizes are deflated in these estimations, since these models place extreme weight on larger studies reporting small effect size estimates with small standard errors while assuming that each estimate relates to a single true effect. Thus, evidence from comparatively smaller studies is strongly discounted since any variation in the observed effect size estimates is considered to be due to measurement error and not possible heterogeneity in true effects. We have argued in section 2 of the main text that this assumption is highly unreasonable in the context of the literature on financial education impact evaluations, since the underlying programs are very heterogeneous in multiple dimensions. Yet, estimates from these models may serve as a lower-bound estimate of the average effect of financial education: The weighted average effect on financial behaviors is estimated to be 0.073 and 0.053 SD units, respectively. The average effect on financial knowledge is estimated to be 0.17 and 0.158 SD units. The 95% confidence intervals clearly rule out zero effects. Note that the estimates in rows 1 to 3 are not statistically different from each other and that the estimate reported in row 4 is not statistically different from the estimate reported in row 3.

Next, we probe the robustness of the estimated financial education treatment effects to the possibility of publication selection bias being present in this empirical literature. Specifically, we investigate whether there is a mechanism that results in the selection of estimates by their statistical significance at conventional levels. If researchers and journal editors tend to favor reporting and publishing statistically significant results over estimates which do not pass tests for significance (i.e., the file drawer problem), the weighted average of this body of evidence is biased. Given the assumption of a single true empirical effect, the standard error of its estimate should be orthogonal to the reported effect sizes in a given literature. If this is not the case, we observe so-called funnel asymmetry. A graphical investigation of the funnel plot in Figure 2 in the main text shows that the distribution of effect sizes is near symmetrical around the estimated true effects for both types of outcomes up until effect sizes of about 0.4 to 0.5 SD units. Effect sizes larger than 0.5 SD units appear to be selected for statistical significance. In row 5, we report results from "precision-effect estimate with standard error" (PEESE) models as suggested by Stanley and Doucouliagos (2012) (see also Table B1 for an implementation of the full FAT-PET-PEESE procedure). The estimate on financial behaviors (0.0426) is statistically not different from the estimate from the unrestricted weighted least squares model with inverse variance weights (row 4), and thus, indicates that the possibility of publication bias does not affect the conclusions drawn from this literature. The estimate on financial knowledge is not significantly different from the estimate relying on unrestricted weighted least squares model with inverse variance weights, as well.

Next, we study the power of studies in the financial education literature. We follow the approach by Ioannidis et al. (2017) and restrict the sample to those estimates that are adequately powered to detect small effects. Assuming conventional levels of statistical significance (α = 0.05) and 80% power ($1-\beta$ = 0.8), the "true effect" will need to be 2.8 standard errors away from zero to reject the zero. The value of 2.8 is the sum of the conventional threshold of 1.96 (at α = 0.05) and 0.84, which is the standard normal value needed to reach the 80th percentile in its cumulative distribution (cf. Gelman and Hill 2006, p. 441). Thus, the standard error of an estimate needs to be smaller than the absolute value of the underlying true effect divided by 2.8 (at $1-\beta$ = 0.8 and α = 0.05). Since the true effect (or the mean of a distribution of true effects) is unknown, we started with the default rule of thumb value for small statistical effect sizes proposed by Cohen (1977) and chose 0.2 SD units as a possible true effect. Note that the median study in this literature (Carpena et al. 2017) has eighty percent power to detect effect sizes of 0.2 SD units, and the average study is powered to have an MDES of 0.23 SD units. Only two studies are able to detect effects as small as 0.05 SD units (Bruhn et al. 2016; Frisancho 2018). The least powered study has 80 percent power to detect effect sizes of approximately one standard deviation (Reich and Berman 2015).

Estimating the unrestricted weighted least squares model with inverse variance weights (i.e., a common true effect assumption) on those studies adequately powered to detect an effect of 0.2 results in the *weighted average of the adequately powered (WAAP)* (Ioannidis et al. 2017) of 0.0466 SD units on financial behaviors in a sample of 198 effect size estimates within 31 studies (see row 6 in Figure B1). Thus, this estimate is still more than twice as large as the estimate reported in Fernandes et al. (2014), clearly different from zero, and near identical to the PEESE or the unrestricted WLS estimate. Similarly, the weighted average effect on financial knowledge in a sample of 115 estimates within 25 studies adequately powered to detect an effect of 0.2 is estimated to be 0.143 SD units.

Next, we use the more appropriate random-effects assumption accounting for the possibility of heterogeneity in true effects between studies and start with the same assumed effect of 0.2 SD units as the mean of the distribution of true effects. We find that the estimate on financial behaviors is now 0.068 SD units (see Figure B4), i.e., 46 percent larger than the estimate with a common effect assumption, and 3.8 times larger than the estimate reported in Fernandes et al. (2014).

Since an estimate of 0.2 SD units appears to be an adequate lower bound of effects on knowledge (see Kaiser and Menkhoff 2017, 2018) an assumed effect of 0.2 may be considered too large regarding the effect on financial behaviors. Thus, we decrease the assumed true effect and rely only on those studies with adequate power to identify an assumed true effect of 0.1 SD units (close to the simple average estimate in a previous meta-analysis by Kaiser and Menkhoff 2017). We estimate the RVE model discussed in the main text. The number of observations for the sample with an MDES of 0.1 is 60 effect sizes within 7 studies. Using only the information from these studies results in an estimated mean of distribution of true effects of 0.0395 SD units. Increasing the assumed mean of the distribution of true effects to above 0.2, on the other hand, leads to larger estimates in this larger sample of studies with adequate power to detect effects of 0.3, 0.4, and 0.5 SD units, respectively (see Figure B4). The same is true for effect sizes on financial knowledge (see Figure B5). We draw two general lessons: First, the effect(s) of financial education appear to be robust and clearly different from zero, even when restricting the sample to only studies with adequate power, and, second, given an estimated mean of the distribution of true effects of 0.1 or smaller, future studies need to have substantial sample sizes to be able to identify these effects if they are present. Assuming individual-level randomization and equal sample sizes in treatment and control groups, studies need to have at least 3,142 observations to identify an effect with 80 percent power. Assuming an effect of 0.05 (and individual-level randomization and a T/C ratio of 1:1) requires a sample size of 12,562. Thus, studies with smaller sample size (such as the earlier literature) do not have adequate power to detect typical effects of financial education, even if they are present.

Next, we probe the sensitivity of results to the decision to include multiple estimates per study in the analyses. Thus, we create one synthetic effect size per study by taking the inverse variance weighted average. Table B3 shows the result for the sample of treatment effects on financial behaviors. The results are similar to the more sophisticated analyses allowing for multiple effect sizes per study.

Finally, we complement these analyses with additional robustness checks. Table B4 shows treatment effects on financial behaviors without the set of papers that do not report intention-to-treat effects (Column 1), without studies by any of the authors of the paper (Column 2), and for those studies that do or do not include a measure of program cost (Columns 3 and 4). Neither of these are statistically different from each other. Table B5 repeats these exercises for the sample of studies that focus on financial knowledge as the outcome. The conclusions are identical.

Appendix B References

Gelman, A., and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models (Analytical Methods for Social Research)*. Cambridge: Cambridge University Press.

Ioannidis, J. Stanley, T.D., and Doucouliagos, H. (2017). The Power of Bias in Economics Research. *Economic Journal* 127(605), F236–65.

Stanley, T. D. and Doucouliagos, H. (2012). *Meta-regression analysis in economics and business*, Routledge, New York, NY.

Stanley, T. D. and Doucouliagos, H. (2015). Neither fixed nor random: weighted least squares meta-analysis. *Statistics in Medicine* 34(13): 2115–2127.

Figure B1: Robustness of financial education treatment effects to different meta-analysis models

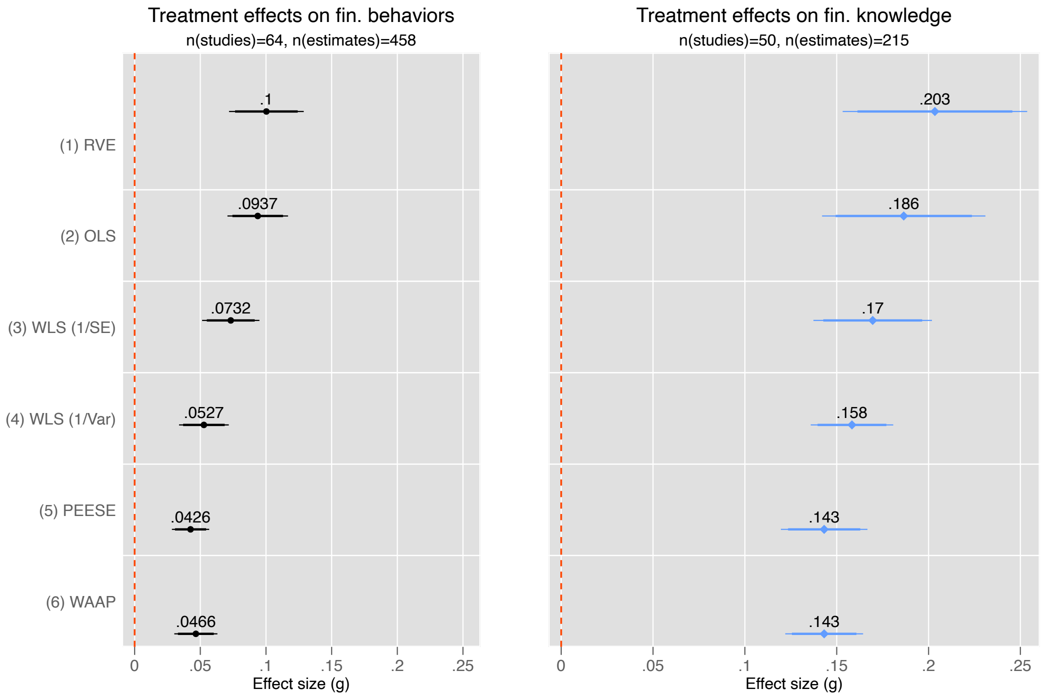
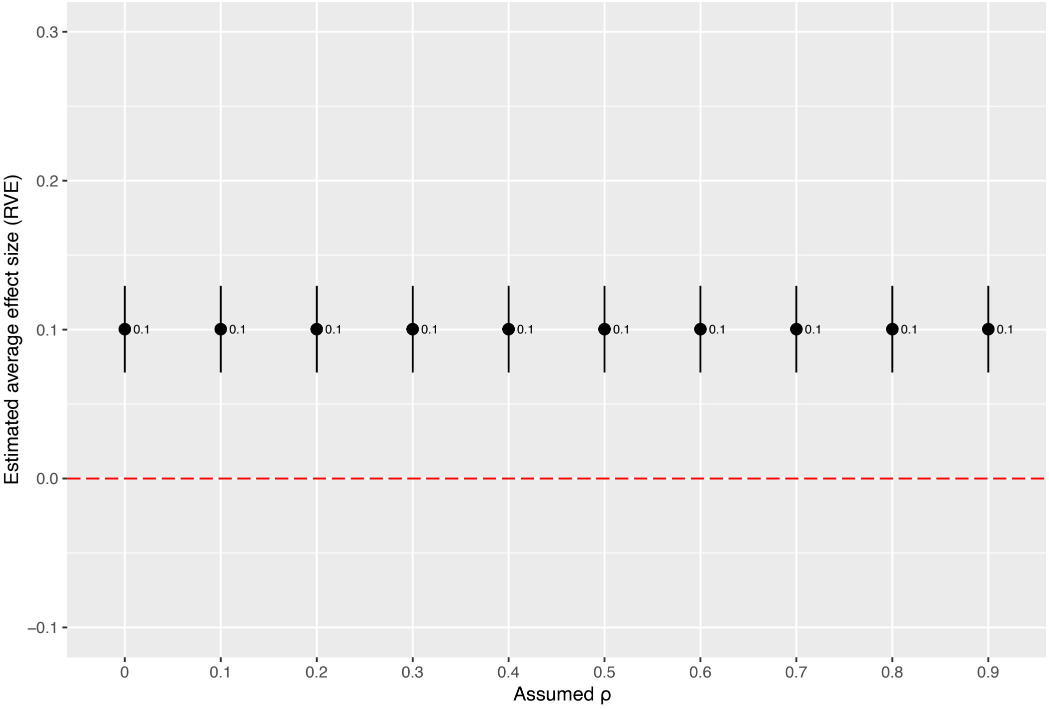
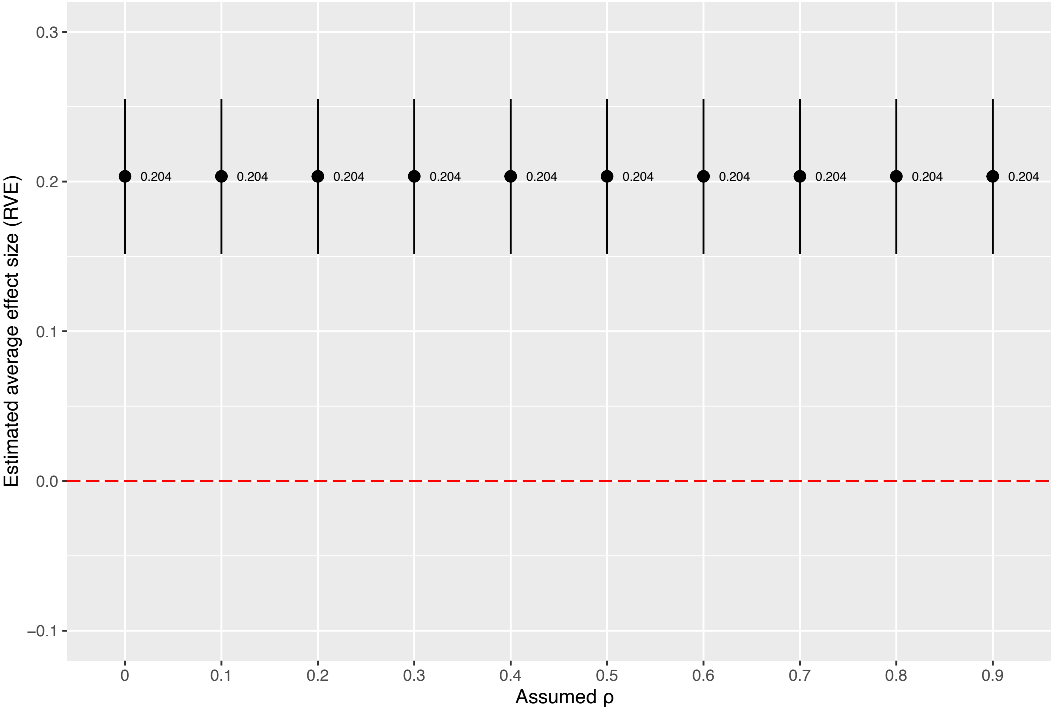


Figure B2: (Non-)Sensitivity of RVE estimate to the choice of ρ (treatment effects on financial behaviors)



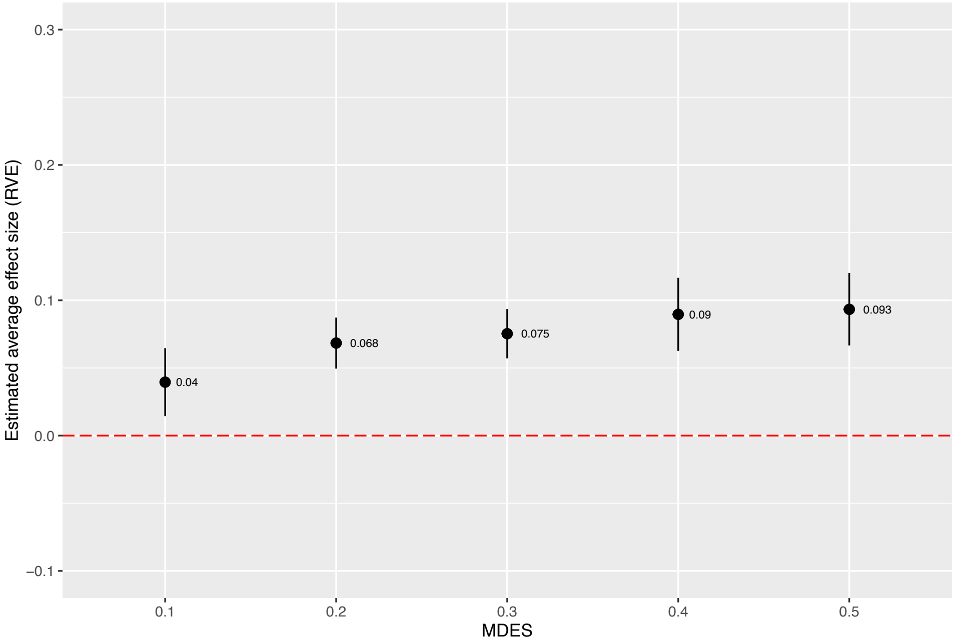
Notes: Figure shows results from (random effects) RVE for different choices of assumed ρ.

Figure B3: (Non-)Sensitivity of RVE estimate to the choice of ρ (treatment effects on financial knowledge)



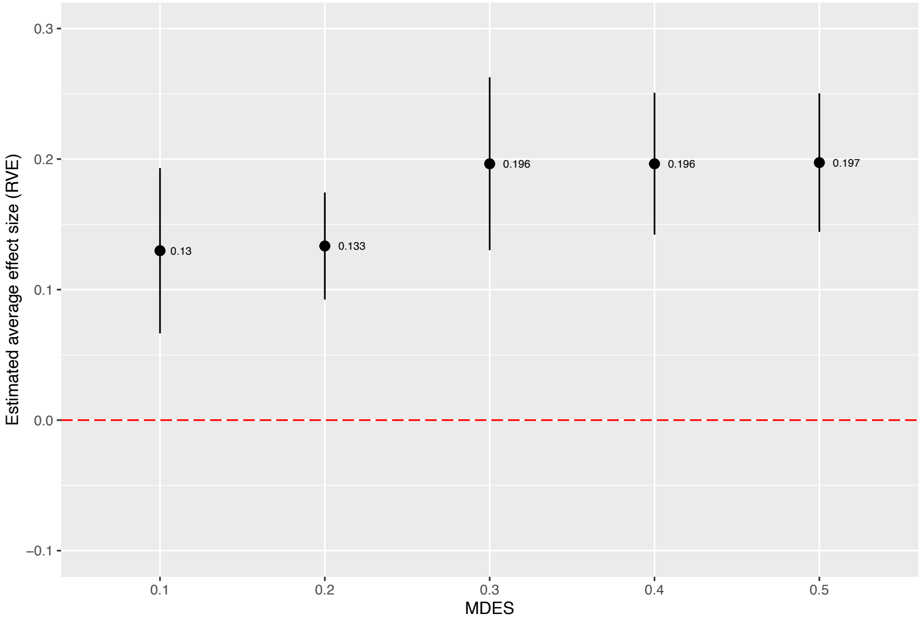
Notes: Figure shows results from (random effects) RVE for different choices of assumed ρ.

Figure B4: Power in the financial behavior sample



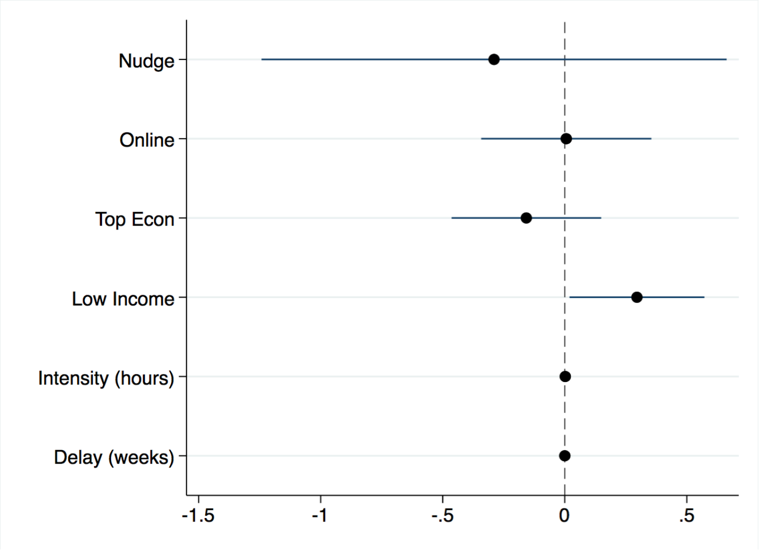
Notes: Average effect size of treatment effects on financial behaviors (from RVE) within the set of studies with the respective MDES. Minimum detectable effect size (MDES) at $\alpha = 0.05$ and $1 - \beta = 0.8$. The number of observations for the sample with a MDES of 0.1 is 60 effect sizes within 7 studies. For MDES=0.2, the sample size is 198 effect size estimates within 31 studies. For MDES=0.3, the sample size is 326 effect sizes in 45 studies. For MDES=0.4, it is 402 effect sizes within 53 studies. For MDES=0.5, it is 443 effect size estimates within 60 studies. The mean MDES in the entire sample is 0.23 SD units. The median MDES in the entire sample of effect sizes is 0.2 SD units (Carpena et al. 2017). The smallest MDES is 0.04 SD units (Frisancho 2018). The largest MDES is 1 SD unit (Reich and Berman 2015). Dots show the point estimate, and the solid lines indicate the 95% confidence interval.

Figure B5: Power in the financial knowledge sample



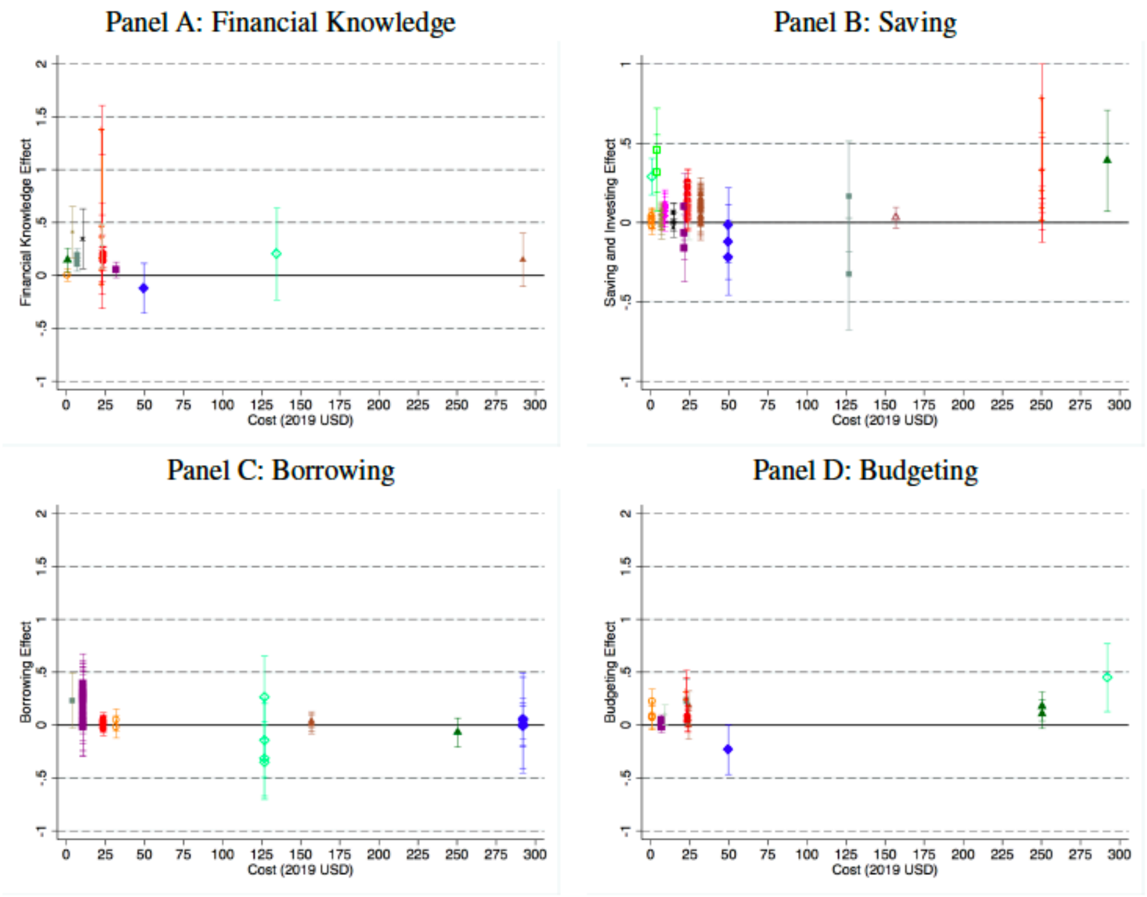
Notes: Average effect size of treatment effects on financial knowledge (from RVE) within the set of studies with the respective MDES. Minimum detectable effect size (MDES) at $\alpha = 0.05$ and $1 - \beta = 0.8$. The number of observations for the sample with an MDES of 0.1 is 12 effect sizes within 7 studies. For MDES=0.2, the sample size is 115 effect size estimates within 25 studies. For MDES=0.3, the sample size is 136 effect sizes in 33 studies. For MDES=0.4, it is 205 effect sizes within 43 studies. For MDES=0.5, it is 209 effect sizes estimates within 45 studies. Dots show the point estimate, and the solid lines indicate the 95% confidence interval.

Figure B6: Which experiments report costs?



Notes: Each point depicts regression coefficients and 95% confidence intervals from linear probability models, where the dependent variable is whether or not the experiment reported the per-participant cost of the intervention. The model includes all covariates depicted at once. The reference groups are Classroom or Counseling intervention, not published in a top Economics Journal, and non-low-income sample. Both the intensity and delay coefficients are precisely estimated zeros. Each data point in the regression is an experiment sample.

Figure B7: Effect sizes by cost for each outcome domain



Notes: Each panel depicts all effect sizes and 95% confidence intervals, as well as the cost per participant in 2019 USD for each of the 20 studies reporting costs in the financial knowledge, saving, borrowing, and budgeting domains. Each color represents effects from a different experiment within that domain. We omit remittances and insurance since there are so few studies in each of those categories.

Table B1: Testing for publication selection bias (FAT-PET-PEESE)

	Financial behaviors			*Financial knowledge*		
	(1)	(2)	(3)	(4)	(5)	(6)
	Unadjusted	FAT-PET	PEESE	Unadjusted	FAT-PET	PEESE
		(1/SE)	(1/SE^2)		(1/SE)	(1/SE^2)
SE		0.731***		0.187***	0.846	
		(0.243)		(0.022)	(0.524)	
SE^2			5.360***			4.538
			(1.514)			(2.736)
Average effect	0.093***	0.032**	0.0426***		0.113***	0.143***
	(0.015)	(0.013)	(0.007)		(0.030)	(0.012)
R^2		0.055	0.054		0.035	0.025
n (Studies)	64	64	64	50	50	50
n (Effect sizes)	458	458	458	215	215	215

Notes: Standard errors (clustered at the study-level) in parentheses. ***, ** and * denote significance at the 1%, 5%, and 10% level.

Table B2: Financial education treatment effects by outcome domain and model

Outcome domain	Treatment effect (g)	Standard Error	95% CI Lower bound	95% CI Upper bound	n(Studies)	n(Effect sizes)
Panel A: RVE						
(1) Fin. Knowledge	0.2035	0.0256	0.1518	0.2551	50	215
(2) Credit	0.0418	0.0199	-0.0003	0.0839	22	115
(3) Budgeting	0.1472	0.0383	0.0673	0.2271	23	55
(4) Saving	0.0972	0.0139	0.0691	0.1252	54	253
(5) Insurance	0.0587	0.0263	-0.0105	0.1278	6	18
(6) Remittances	0.0472	0.0551	-0.0953	0.1897	6	17
Panel B: OLS						
(1) Fin. Knowledge	.1864942	.0221258	.1420307	.2309578	50	215
(2) Credit	.0658676	.0300674	.0033391	.1283961	22	115
(3) Budgeting	.1851885	.0467036	.0883311	.2820459	23	55
(4) Saving	.0934569	.0153911	.0625863	.1243275	54	253
(5) Insurance	.0374928	.0174763	-.0074315	.0824172	6	18
(6) Remittances	.0497656	.0513203	-.0821574	.1816887	6	17
Pancel C: WLS (1/SE_g)						
(1) Fin. Knowledge	.1696031	.0160508	.1373478	.2018585	50	215
(2) Credit	.028473	.0253898	-.024328	.0812741	22	115
(3) Budgeting	.1340191	.0465008	.0375823	.2304558	23	55
(4) Saving	.0809610	.013229	.0544271	.1074950	54	253
(5) Insurance	.0383468	.017276	-.0060625	.0827562	6	18
(6) Remittances	.0364145	.0504204	-.0931953	.1660243	6	17
Panel D: WLS (1/Var_g)						
(1) Fin. Knowledge	.1583121	.0111803	.1358445	.1807797	50	215
(2) Credit	-.0089557	.019686	-.0498949	.0319835	22	115
(3) Budgeting	.0863279	.0324534	.0190236	.1536322	23	55
(4) Saving	.0692363	.0148824	.039386	.0990865	54	253
(5) Insurance	.0388070	.0170809	-.005101	.0827149	6	18
(6) Remittances	.0235143	.0492582	-.103108	.1501366	6	17
Panel E: PEESE						
(1) Fin. Knowledge	.1431734	.0117256	.1196100	.1667368	50	215
(2) Credit	-.0254595	.0119766	-.0503661	-.0005529	22	115
(3) Budgeting	.0516157	.0234908	.0028988	.1003327	23	55
(4) Saving	.0637163	.0165028	.0306159	.0968167	54	253
(5) Insurance	.0752942	.0649840	-.0917526	.2423409	6	18
(6) Remittances	-.3395076	.0923237	-.5768332	-.1021820	6	17
Panel F: WAAP (MDES of 0.2)						
(1) Fin. Knowledge	.1431727	.0102640	.1219889	.1643565	25	115
(2) Credit	-.0221165	.0138866	-.0535302	.0092972	10	31
(3) Budgeting	.0548868	.0190372	.0118217	.0979519	10	24
(4) Saving	.0640661	.0156289	.0347707	.0987992	29	141
(5) Insurance	.0408968	.0369759	-.4289263	.5107200	2	2
(6) Remittances	-	-	-	-	0	0

Table B3: Using only one synthetic effect size per study (treatment effects on financial behaviors)

	(1) OLS	(2) Unrestricted WLS	(3) Fixed-effect Meta-Analysis	(4) Random-effects (REML)
β_0	0.116	0.055	0.055	0.090
(SE)	(0.021)	(0.006)	(0.002)	(0.012)
[CI$_{95}$]	[0.074, 0.157]	[0.043, 0.066]	[0.050, 0.059]	[0.066, 0.113]
Q-statistic	-	-	464.71	464.71
I^2	-	-	86.44%	94.91%
n (Studies)	64	64	64	64
n (Effect sizes)	64	64	64	64

Notes: Column (1) presents results from a simple OLS regression. Column (2) presents results from unrestricted weighted least squares with inverse variance weights (see Stanley and Doucouliagos, 2015). Column (3) presents results from (restricted) fixed-effect meta-analysis with inverse variance weights. Column (4) presents results from random-effects meta-analysis (using restricted maximum likelihood).

Table B4: Additional robustness checks (treatment effects on financial behaviors)

	(1) ITT estimates only	(2) Excluding authors' experiments	(3) Experiments reporting costs	(4) Experiments not reporting costs
β_0	0.0792	0.0988	0.0629	0.1203
(SE)	(0.0101)	(0.0148)	(0.0159)	(0.0205)
[CI$_{95}$]	[0.1391, 0.2442]	[0.0690, 0.1286]	[0.0288, 0.0969]	[0.0788, 0.1618]
n (Studies)	57	62	19	45
n (Effect sizes)	448	439	167	291

Table B5: Additional robustness checks (treatment effects on financial knowledge)

	(1) ITT estimates only	(2) Excluding authors' experiments	(3) Experiments reporting costs	(4) Experiments not reporting costs
β_0	0.1916	0.1979	0.1573	0.2174
(SE)	(0.0261)	(0.0267)	(0.0408)	(0.0309)
[CI$_{95}$]	[0.1391, 0.2442]	[0.1440, 0.2518]	[0.0659, 0.2487]	[0.1546, 0.2803]
n (Studies)	46	46	12	38
n (Effect sizes)	211	176	23	192

Table B6: Analysis of intensity and delay in measurement (treatment effects on financial behaviors)

	(1) Effect size (g)
Intensity	0.0043
	(0.0024)
Intensity× Intensity	-0.0000
	(0.0000)
Delay	-0.0018
	(0.0052)
Delay × Delay	-0.0000
	(0.0002)
Intensity × Delay	-0.0001
	(0.0003)
n (Studies)	52
n (Effect sizes)	419

Note: This table reruns the main analysis of the result presented in Figure 4 in Fernandes et al. (2014) with updated data. Intensity is (mean-centered) number of hours of instruction, Delay is delay between treatment and measurement of outcomes in months. Results from RVE (random-effects assumption). Robust standard errors in parentheses. Assumed $\rho = 0.8$. Estimated τ^2=0.0111.

Appendix C: Comparing our data to previous quantitative meta-analyses

Table C1: Comparison of datasets

	RCT	Fernandes et al. (2014)	Miller et al. (2015)	Kaiser and Menkhoff (2017)
1)	Abarcar et al. (2018)	No	No	No
2)	Abebe et al. (2018)	No	No	No
3)	Alan and Ertac (2018)	No	No	No
4)	Ambuehl et al. (2014)	No	No	Yes
5)	Angel (2018)	No	No	No
6)	Attanasio et al. (2019)	No	No	No
7)	Barcellos et al. (2016)	No	No	Yes (2012 WP)
8)	Barua et al. (2012)	No	No	Yes
9)	Batty et al. (2015) [independent sample 1]	No	No	Yes
10)	Batty et al. (2015) [independent sample 2]	No	No	Yes
11)	Batty et al. (2017)	No	No	No
12)	Becchetti and Pisani (2012)	No	No	No
13)	Becchetti et al. (2013)	Yes	No	Yes
14)	Berg and Zia (2017)	No	Yes	Yes
15)	Berry et al. (2018)	Yes (2013 WP)	No	Yes
16)	Bhattacharya et al. (2016)	No	No	No
17)	Bhutoria and Vignoles (2018)	No	No	No
18)	Billari et al. (2017)	No	No	No
19)	Bjorvatn and Tungodden (2010)	No	No	Yes
20)	Bonan et al. (2016)	No	No	No
21)	Bover et al. (2018)	No	No	No
22)	Boyer et al. (2019)	No	No	No
23)	Brugiavini et al. (2015) [independent sample 1]	No	No	Yes
24)	Brugiavini et al. (2015) [independent sample 2]	No	No	Yes
25)	Bruhn and Zia (2013)	No	No	Yes
26)	Bruhn et al. (2016)	No	Yes (2013 WP)	Yes
27)	Bruhn et al. (2014)	Yes (2013 WP)	Yes (2012 WP)	Yes
28)	Calderone et al. (2018)	No	No	No
29)	Carpena et al. (2017)	No	No	Yes (2015 WP)
30)	Carter et al. (2016)	No	No	No
31)	Choi et al. (2010) [indendent sample 1]	Yes (coding error)[19]	No	Yes
32)	Choi et al. (2010) [indendent sample 2]	No	No	No
33)	Choi et al. (2010) [indendent sample 2]	No	No	No
34)	Clark et al. (2014)	Yes (2012 WP)	No	Yes
35)	Cole et al. (2013)	Yes (coding error)[20]	No	Yes
36)	Cole et al. (2011)	Yes (coding error)[21]	Yes	Yes

[19] Wrongly classified as quasi-experiment and not included in the RCT sample (see Appendix D).

[20] Wrongly coded estimate (wrong sign and magnitude) and misclassified financial behavior as savings when it is in the insurance domain (see Appendix D).

[21] Wrongly coded multiple time-points within the same study as independent samples (see Appendix D).

37)	Collins (2013)	Yes (2011 WP)	No	Yes
38)	Collins and Urban (2016)	No	No	No
39)	Custers (2011)	No	No	Yes
40)	Doi et al. (2014)	No	Yes (2012 WP)	Yes
41)	Drexler et al. (2014)	Yes (coding error)[22]	Yes	Yes
42)	Duflo and Saez (2003)	Yes	No	Yes
43)	Elbogen et al. (2016)	No	No	Yes
44)	Field et al. (2010)	No	No	Yes
45)	Flory (2018)	No	No	Yes (2016 WP)
46)	Frisancho (2018)	No	No	No
47)	Furtado (2017)	No	No	No
48)	Gaurav et al. (2011)	Yes	No	Yes
49)	Gibson et al. (2014) [independent sample 1]	No	Yes (2012 WP)	Yes
50)	Gibson et al. (2014) [independent sample 2]	No	Yes (2012 WP)	Yes
51)	Gibson et al. (2014) [independent sample 3]	No	Yes (2012 WP)	Yes
52)	Gine and Mansuri (2013)	No	Yes (2011 WP)	Yes
53)	Gine et al. (2013)	No	No	Yes
54)	Han et al. (2009)	Yes (coding error)[23]	No	Yes
55)	Haynes et al. (2011)	No	No	Yes
56)	Heinberg et al. (2014)	No	No	Yes
57)	Hetling et al. (2016)	No	No	No
58)	Hinojosa et al. (2010)	No	No	No
59)	Jamison et al. (2014)	No	No	Yes
60)	Kaiser and Menkhoff (2018)	No	No	No
61)	Kajwij et al. (2017)	No	No	No
62)	Lührmann et al. (2018)	No	No	No
63)	Lusardi et al. (2017)	No	No	Yes (2015 WP)
64)	Migheli and Moscarola (2017)	No	No	No
65)	Mills et al. (2004)	Yes (coding error)[24]	No	Yes
66)	Modestino et al. (2019)	No	No	No
67)	Postmus et al. (2015)	No	No	No
68)	Reich and Berman (2015)	No	No	Yes
69)	Sayinzoga et al. (2016)	No	No	Yes
70)	Seshan and Yang (2014)	Yes (2012 WP / coding error)[25]	No	Yes
71)	Shephard et al. (2017)	No	No	No
72)	Skimmyhorn et al. (2016)	No	No	Yes
73)	Song (2012)	No	No	Yes
74)	Seinert et al. (2018)	No	No	No
75)	Supanataroek et al. (2016)	No	No	Yes
76)	Yetter and Suiter (2015)	No	No	Yes

[22] Wrongly coded multiple treatments as independent samples even though they are compared to a common control group (see Appendix D).

[23] Wrongly classified as quasi-experiment and not included in the RCT sample (see Appendix D).

[24] Wrongly classified as quasi-experiment and not included in the RCT sample (see Appendix D).

[25] Wrongly coded estimate on savings (wrong sign) (see Appendix D).

Appendix D:
Replicating Fernandes et al. (2014)

While the analysis by Fernandes et al. (2014) includes evidence from randomized trials, quasi-experiments, and observational studies, it is most often cited for the lack of impact of financial education interventions (i.e., what Fernandes et al. (2014) term "manipulated financial literacy"). Our paper does not take a stance on the internal validity of observational studies in the present literature. Also, we do not disagree that quasi-experiments in this literature (which also are highly heterogenous with regard to their internal validity) may report inflated effect sizes relative to RCTs, which have higher internal validity, on average. We disagree, however, that there are no effects of financial education treatments on financial behaviors, as evidenced by the large number of randomized experiments.

Despite newer data presented in the main paper, we would like to understand the result by Fernandes et al. (2014) on the early set of RCTs. Thus, we attempt to replicate their original result regarding RCTs and document the differences between our analysis and theirs.

Our analysis includes twenty of the reported effect size estimates in Fernandes et al. (2014). Specifically, we compare our extracted estimates to the reported "effect size(s) (partial r)" in Table WA1 ("Studies of Manipulated Financial Literacy with Randomized Experiments") and, in five wrongly classified cases, to estimates reported in Table WA2 ("Studies of Manipulated Financial Literacy with Pre-Post or Quasi-Experiments").

Our attempt to replicate the result by Fernandes et al. (2014) is not entirely successful. We begin by clarifying that Fernandes et al. (2014) choose to include 15 observations from 13 papers in their analysis of RCTs. In doing so, they average across multiple reported treatment effects within studies and create one effect (one observation) per study to be used in the analysis. While we disagree with the approach to average effect sizes across outcome domains into one effect-size per study, we follow this approach here to be able to compare the results.

Unfortunately, the manuscript by Fernandes et al. (2014) lacks details about their exact method. What we can infer from their text is the following:

(i) Fernandes et al. (2014) create one effect size (r) per study:

> "*Most studies reported multiple effect sizes across dependent variables. We averaged the effect sizes for each study that manipulated financial literacy and for each study that measured financial literacy*" (Fernandes et al. 2014, p.1863).

What remains unclear, however, is whether this is a simple average (i.e., the arithmetic mean of the effect sizes and their standard errors) or a weighted average. The textbook meta-analysis literature clearly cautions against the use of simple averages (cf. Borenstein et al. 2009).

(ii) Fernandes et al. (2014) conduct a meta-analysis using the inverse variance of the extracted estimates as weights:

> *"Because sample size affects the correspondence between the estimated relationship between variables and true relationship [sic!], we first weighted effects by the inverse variance. Empirically in our sample, smaller studies reported larger effect sizes. Given that it requires a larger effect size to reach statistical significance with a smaller N, this might suggest a publication bias favoring significant results. We examined significance for the mean effect size by calculating the confidence intervals of the effect sizes to determine whether the confidence interval includes 0."* (Fernandes et al. 2014, p.1864).[26]

While this paragraph implies Fernandes et al. (2014) use a common-effect assumption in their approach to meta-analysis (the weights are solely defined by the within-study sampling variances), the calculation of the standard error for the "mean effect size" is not disclosed. Note that unrestricted weighted least squares (Stanley and Doucouliagos 2015) and the more common and canonical "common-effect" (sometimes also called "fixed-effect") meta-analysis which restricts the multiplicative constant to be one (cf. Stanley and Doucouliagos 2015, p. 20) and is implemented in most meta-analysis packages, may lead to very different estimates of the

[26] Conflicting with this description of the method in the main text, the Appendix to Fernandes et al. (2014) state that the estimated mean effect sizes are "sample weighted" (See Table WA1). While the within-study variances are obviously inversely related to sample size, we note that they are not a direct function of total N. Instead the estimated within-study standard errors will also depend on the choice of econometric model (i.e., clustering of standard errors, regression-adjustment by including pre-treatment covariates such as the lacked outcome). Thus, these alternative approaches (weights based on sample-size and inverse-variance weights) will produce different estimates of both the (weighted) average effect size and its confidence interval.

standard error of the (weighted) average effect size. Thus, we estimate both approaches in the later comparison of results.

Agreement in coding of studies and effect sizes.

We start by noting that our dataset agrees with four out of fifteen extracted estimates where we get identical signs and magnitudes. These experiments are Berry et al. (2013 [2018]), Clark et al. (2012 [2014]), Gine et al. (2013), and Gaurav et al. (2011).[27]

Another two estimates have identical signs and similar magnitudes. These papers are Becchetti et al. (2013), in which both the dataset by Fernandes et al. (2014) and our dataset include an estimate on "savings" but different magnitudes (*r* of 0.04 vs 0.06), and Bruhn et al. (2013 [2014]), in which both their and our dataset include effects on "savings" and "debt" (*r* of 0.01 vs. 0.02). We are unable to tell exactly why these differences in magnitude arise. In the case of Becchetti et al. (2013) we code the estimate from Table 9 (see Becchetti et al. 2013, p. 826) but there are also alternative specifications regarding the same effect reported in Tables 15 to 17, which arrive at different magnitudes. This is a likely source of the difference in results. In the case of Bruhn et al. (2013 [2014]), we note that we use the 2014 version of the paper published in the *Journal of Development Economics* 108 (pp. 184-189) whereas Fernandes et al. (2014) rely on an earlier working paper from 2013. However, we find that the reported estimates do not differ (see Bruhn et al. 2013, Tables 5 and 7; Bruhn et al. 2014, Table 2). A likely source of difference may lie in the fact that we only code the reported ITT estimates from table two, whereas Fernandes et al. (2014) state that they code the TOT for 46 percent of the experiments (Fernandes et al. 2014, p. 1865). It is possible that they chose to code the LATE estimate reported in tables 5 to 7 in Bruhn et al. (2013) that are generally larger in magnitude (and also the negative effects related to credit outcomes). Another possibility relates to the

[27] Note that the outcome domain "insurance" appears to be termed "plan" in Fernandes et al. (2014), since both Gine et al. (2013) and Gaurav et al. (2011) include estimates only on insurance purchase decisions.

decision of which variables to code. We rely on the results of aggregated indices reported in Table 2 and do not code redundant effects of the single components present in the appendix. In total, we think that it is fair to say that we generally agree with six out of fifteen extracted estimates.

Disagreement in coding of studies and effect sizes.

Next, we document six cases where we disagree with how studies have been coded. First, we note that we generally disagree with the approach by Fernandes et al. (2014) to count multiple observations from the same experiment (i.e., when multiple treatments are compared to a common control group, as in Drexler et al. (2014), or when there is a longer term follow-up on the original experimental sample, as in Cole et al. (2011) as two separate studies. This is deeply problematic, as it clearly violates the assumption of independent estimates required for the model chosen by Fernandes et al. (2014).[28]

Specifically, we disagree with counting the estimates in Cole et al. (2011) as two separate studies. One set of estimates is concerned with the short-term treatment effects (see Table 5, C1 and C2) and another set of estimates reports on long-term results (see Table 8, C1 and C2; Table 10, C1 and C2) after two years *on the same experimental sample* (albeit with substantial attrition). These estimates can never be included as independent in any meta-analysis. In addition to this difference, we note, again, that we chose to code the reduced form estimates in Tables 5 and 8 whereas it is likely that Fernandes et al. (2014) rely on the LATE estimates for the short-term result in Tables 7.

Additionally, we disagree with including Drexler et al. (2014) twice in this meta-analysis. The paper by Drexler et al. (2014) compares two different financial education

[28] Note that the correct inclusion of these estimates is easily implemented in an analysis relying on RVE, or alternatively (if one insists on explicitly not modeling between-study heterogeneity in true effects) on an unrestricted WLS regression with multiple effect sizes and cluster-robust standard errors at the study level.

treatments (differing in their content) to *a common control group*. Thus, again, these are not independent experiments and can never be counted twice in any meta-analysis that uses only one observation per study. Note that we agree with the sign and magnitude when averaging these two experimental treatments into one synthetic estimate.

Regarding the paper by Duflo and Saez (2003), we arrive at an estimate of similar magnitude but with an opposite sign. Digging deeper into this paper, we note that this is likely the result of different coding decisions that have to be debated. Duflo and Saez (2003) estimate the effect of informational events on the enrollment decisions of employees in a retirement plan. They specifically set up the experiment to study social interactions (i.e., identifying spill-over effects). They randomize invitation to the informational event both at the department and the individual level. Their results clearly suggest that untreated individuals in treated departments (i.e., employees working in a department where a random subset of employees have received an invitation to the fair) are as likely to respond to the treatment as treated individuals in treated departments (i.e., employees receiving an invitation themselves). Thus, comparing only those employees who received the invitation themselves to the pure control group (i.e., employees working in a department where no one received an invitation) leads to a biased estimate of the treatment effect, since the positive externality of interacting with a treated peer in a treated department is masked in such an analysis.[29] This appears to be exactly the source of the different sign in our data and the data presented in Fernandes et al. (2014). Only when an analyst exclusively codes the effect of the "letter-dummy," either in the reduced form analysis in Table 2 (Columns 2 and 3) or only the results from the IV-regression (i.e, the effect of fair attendance) in Table 3, does one gets an overall negative sign. Coding both the "department treatment" (Table 2) and the "letter and department treatment" results in an overall positive sign. Given

[29] See Duflo and Saez (2003, p.835): "The naive estimate would underestimate the overall effect of the fair (since part of the "control" group is actually treated) and overestimate the direct effect on those who received the letter. This shows the potential bias in randomized trials that ignores externalities."

that the experiment is specifically set up to identify treatment externalities and that the biases arising from ignoring them are discussed at length in the paper, it appears controversial to not consider the effects of being in a treated department. We reached out to Fernandes et al. and they confirmed they chose to only code the effect of fair attendance.

Additionally, we are puzzled by the fact that the two (short and longer term) estimates from Duflo and Saez (2003) are now (correctly) aggregated only as one observation whereas in the logic of the coding applied to the study by Cole et al. (2011), Duflo and Saez (2003) had to appear twice, as well. Thus, the coding appears to be inconsistent across studies.

Next, we extracted different estimates from Collins (2013) than Fernandes et al. (2014) did from an earlier version of the paper (Collins 2011). While we are unable to tell the exact source of difference in the synthetic effect size, we note that Collins (2013) includes a multitude of reported treatment effects, including reduced form results, the treatment effect on the treated, results from propensity score matching, and results from a Heckman 2-stage specification. The paper reports a total of 66 treatment effect estimates, including both self-reported behaviors and results from administrative data. Our estimates rely only on the reduced form (intention to treat) estimates presented in Table 4. The effects are clearly negative when aggregated (*r* of -0.065 in our data vs. +0.02 in Fernandes et al. 2014). This overall effect appears to be consistent with what is being advertised in Collins' abstract.

Next, we document a coding discrepancy regarding Seshan and Yang (2012) (subsequently published as Seshan and Yang 2014, *Journal of Development Economics*). Fernandes et al. (2014) report in Table WA1 the average effect on "savings" to be negative; however, Seshan and Yang (2012) report positive (insignificant) estimates on total household savings both in the earlier working paper version coded by Fernandes et al. (2014) (see Table 7, Columns 4 and 8) and in the updated and published version (see Table 3, Columns 1 and

2).[30] We reached out to Fernandes et al. and they stated that they did not code the estimate on total household savings (Table 7, Columns 4 and 8 in Seshan and Yang (2012)) but "[the] estimates on the savings of the person and not the savings with a spouse". While there is indeed an early version of the paper that shows a negative sign on this singular savings estimate (Column 1) the table clearly indicates that this is not the total estimate of the savings-effect but that Column 4 represents the aggregate impact on total household savings (sum of Columns 1 to 3). Consistent with this interpreatation, later versions of the paper only report aggregated (positive) impacts on household savings.

Finally, we disagree with including the study by Carpena et al. (2013) in this meta-analysis, as no financial behaviors are considered in the study. The paper reports treatment effects on financial knowledge and attitudes, but not on actual behaviors. In a later paper on the same experiment, Carpena et al. (2017) collect data on actual financial behaviors. Thus, we included this paper in our analysis of the updated data. We contacted one of the authors, and he confirmed that Carpena et al. (2017) was the appropriate experiment to include and that the earlier paper did not include any estimates of treatment effects on financial behaviors.

As a general remark, we note that we find it worrysome that Fernandes et al. (2014) state that they chose to focus on the *treatment effect on the treatment* for eight out of fifteen experiments (see Fernandes et al. 2014, p. 1865) and code the *intention to treat effects* for seven experiments. There is not a single experiment in this set that reports the TOT and does not at the same time report reduced form results (ITT). When both are available, we suggest that comparing the ITT across studies is the more appropriate comparison, or alternatively use variation within studies to code both types of effects and include an indicator in a meta-regression model.

[30] Note, that the paper also includes treatment effect estimates on budgeting behavior (financial practices) and remittances, which we code for our analysis with updated data but not for the purpose of this replication.

Coding errors in Fernandes et al. (2014)

While we have thus far documented agreement in coding and cases where we disagree, the disagreements do not necessarily constitute errors in coding, but they reflect decisions that are subject to researcher degrees of freedom present in any meta-analysis. In contrast, we now document four cases that constitute factual errors. We distinguish between two types of coding errors: (i) errors in the coding of effect sizes, and (ii) errors in the classification of studies and effect sizes.

First, we document coding errors for Cole et al. (2012) (subsequently published as Cole et al. 2013, *AEJ: Applied*). Fernandes et al. (2014) state in Table WA1 that Cole et al. (2012) report negative treatment effects on "savings." However, this experiment exclusively reports effects on insurance take-up in response to financial education. We contacted two of the authors of this paper, and they confirmed that there was never a version of this paper reporting treatment effects on savings. Additionally, and more importantly, the effect size has been wrongly coded. The baseline effects (Columns 1-3 of Table 5 in Cole et al. 2013) of the education treatment on take-up of the rainfall insurance product in Andhra Pradesh are clearly positive (albeit noisy). One may speculate whether an analyst coding the paper included estimates in the presence of the interaction terms reported in columns 4 to 6 of Table 5 without considering the net effect, or whether an analyst simply averaged across all columns of Table 5 without considering the net effect with interactions, which could falsely "result" in a negative overall effect of "Education Module" on the outcome (which is then classified as "savings" when it is actually "insurance take-up"). We asked two of the authors of the paper about their opinion on the coding and they agreed their paper was miscoded in Fernandes et al. (2014). We subsequently reached out to Fernandes et al. and they confirmed that our estimate was the appropriate one to include.[31]

Next, we note that three papers seem to have been misclassified to be quasi-experimental studies when they are actually randomized experiments. Fernandes et al. (2014) coded the paper Choi et al. (2008) (subsequently published as Choi et al. 2010, *Review of Financial Studies*) as a "Quasi-Experiment" (see Fernandes et al. 2014, Table WA2). However, this paper clearly presents evidence from randomized experiments: *"We randomly divided our participants into four information conditions"* (Choi et al. 2010, p. 1409). Additionally, we are puzzled by the decision to aggregate the evidence from the three experiments that are presented in the paper into one synthetic effect size. In contrast to the cases where papers have been included twice in the analysis before, this paper clearly presents evidence from three separate small-scale experiments *with an independent control group each*; some of them are even conducted in different years (one experiment on MBA students at Wharton, one experiment on college students at Harvard, and one experiment on Harvard staff (see Choi et al. 2010, p.1416)). Thus, we include the three experiments in our analysis.

Next, Fernandes et al. (2014) code Han et al. (2007) as a "Quasi-Experiment" (see Table WA2) even though the paper clearly leverages a *"[...]randomized longitudinal experimental design [...]"* (Han et al. 2007, p.16). However, one may argue that this paper should not be included in the meta-analysis at all, since financial education is confounded with IDA participation: *"[...] only the treatment group participated in the IDA program and received the required financial education classes"* (Han et al. 2007, p. 16). Since Fernandes et al. (2014) chose to include the paper in their analysis, however, we include it for the sake of comparability.

[31] They also clarified that the estimate on "insurance take-up" was classified as "savings" in this case due to a lack of a category for estimates in the "insurance domain". Note, however, that the outcome domain "insurance" appears to be coded as the outcome-category "plan" in the case of Gine et al. (2013) and Gaurav et al. (2011). Both of these studies exlusively include estimates on the take-up of index based insurance products. Thus, the classification of these estimates does not appear to be entirely consistent across studies.

Note that Fernandes et al. (2014) chose to include two non-independent estimates as two separate "studies" ("Study 1 and Study 2", Table WA2). However, the paper reports only results from one experiment and presents both ITT results (Table 5) and "efficacy subset" results (Table 6), which are essentially TOT results. We only code the reduced form estimates from Table 5 on p.13, and strongly disagree with including these non-independent estimates as two separate studies. Note that we agree on the direction and exact magnitude of effect size when the two estimates in Fernandes et al. (2014) are combined.

Finally, Fernandes et al. (2014) include Mills et al. (2004) in their quasi-experimental sample. This paper is also situated in context of IDA participation, and, again, financial education treatment is confounded with the other features of the IDA program: *"Prior to a matched withdrawal, participants were required to take 12 hours of general financial education and (in most instances) additional training specific to the type of intended asset purchase."* (Han et al. 2007, p. iii). Despite this fact, the paper uses a randomized experiment to estimate the treatment effects: *"To allow unbiased estimation of program effects, program applicants were randomly assigned to a treatment group, which was allowed to enter the program, or to a control group, which was not"* (Han et al. 2007, p.1). Thus, this paper should either not be included at all or be included as an RCT. It is definitely not a quasi-experiment (even though there appears to be differential attrition). Finally, we disagree with including estimates from two time points as two separate studies. The paper includes data from one experiment but at multiple follow-ups.

Do these differences matter for the estimated average effect?

We now compare the difference in results with our data as discussed above to the analysis presented in Fernandes et al. (2014), Table WA1. We first use one (synthetic) observation per study and estimate both (1) unrestricted weighted least squares and (2) a fixed effect meta-analysis, since these models are comparable with the original strategy outlined in Fernandes et al. (2014). Additionally, we estimate (3) a random-effects model with one synthetic effect size per study. To probe the sensitivity of results to the decision to create within-study average effect sizes, we estimate (4) unrestricted weighted least squares with multiple effect sizes per study and cluster-robust standard errors at the study level, (5) robust variance estimation with dependent effect size estimates (RVE) using "fixed-effect" weights, and (6) RVE with weights that account for the heterogeneity in true effects (see Section 4).

Table D2 shows results for the different models. We start with noting that the original result by Fernandes et al. (2014) results in an overall effect of r=0.009 (g=0.018) with the 95 percent confidence interval including zero. In our replication, the smallest effect size (see column 1, Panel A) is about 30 percent larger and clearly rules out zero effects in its 95 percent CI. Adding the falsely classified estimates from Table WA2 to the sample increases the average effect by a factor of 3 (relative to the original result presented in Fernandes et al. (2014)). This result is similar, irrespective of the model used. We next compare the results to the more sophisticated RVE model, which also serves as a sensitivity check to the practice of creating within-study averages. We find that the overall effect with a fixed-effect assumption (column 5 of Panel B) is r=0.018 (g=0.036), i.e., precisely double the effect reported in Fernandes et al. (2014). Relaxing the assumption to allow for heterogeneity in true effects results in an effect of r=0.023 (g=0.046). Thus, while it is true that the estimated treatment effects from the earlier literature are smaller than the recent studies, the effect size reported in RCTs was at least 30 to 50 percent larger than stated in Fernandes et al. (2014) and also significantly different from zero.

Appendix D References

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley. http://dx.doi.org/ 10.1002/9780470743386

Fernandes, D., Lynch Jr., J.G., and Netemeyer, R.G. (2014). Financial literacy, financial education, and downstream financial behaviors. *Management Science*, 60(8): 1861–1883.

Table D1: Replication attempt of the Fernandes et al. (2014) result on RCTs

Fernandes et al. (2014) (Table WA1)				Our data			
	Study	Effect size (r)	Outcomes coded	Year	Effect size (r) (SE)	Outcomes Coded	Notes
1	Becchetti et al. (2013)	0.04	Save	2013	0.063 (0.035)	D (save)	Agreement in sign
2	Berry et al. (2013)	0.01	Save, Plan	2018	0.008 (0.004)	B (credit), D (save)	Agreement in sign and magnitude
3	Bruhn et al. (2013)	0.01	Save, Debt	2014	0.020 (0.013)	B (credit), D (save)	Agreement in sign
4	Carpena et al. (2013)	0.02	Cash flow	-	-	-	Not included
5	Clark et al. (2012)	0.02	Invest	2014	0.023 (0.017)	D (save/invest)	Agreement in sign and magnitude
6	Cole et al. (2012)	-0.03	Save	2013	0.003 (0.033)	E (insurance)	Coding error in sign magnitude, and classification
7	Cole et al. (2011) ["sample 1"]	-0.03	Cash flow	2012	-0.023 (0.035)	D (savings)	Agreement in sign
8	Cole et al. (2011) ["sample 2"]	-0.07	Cash flow	-	-	-	Counted as two RCTs
9	Collins (2011)	0.02	Save, debt, invest	2013	-0.065 (0.054)	B (credit), D (save/invest)	Disagreement
10	Drexler et al. (2011) ["sample 1"]	0.02	Save, Cash flow, Invest	2014	0.041 (0.021)	C (Budgeting), D (save/invest)	Agreement in sign (and magnitude if averaged)
11	Drexler et al. (2011) ["sample 2"]	0.06	Save, Cash flow, Invest	-	-	-	Counted as two RCTs
12	Duflo and Saez (2003)	-0.01	Plan active	2003	0.012 (0.012)	D (save/retirement)	Disagreement
13	Gaurav et al. (2011)	0.08	Plan	2011	0.080 (0.041)	E (insurance)	Agreement in sign and magnitude
14	Gine et al. (2013)	0.04	Plan	2013	0.0399 (0.0345)	E (insurance)	Agreement in sign and magnitude
15	Seshan and Yang (2012)	-0.04	Save	2014	0.0344 (0.0139)	D (save)	Coding error in sign and magnitude
RCTs wrongly coded as quasi-experiments in Fernandes et al. (2014) (Table WA2)							
[25]	Choi et al. (2008)	0.02	Invest	2010	-	D (save/invest)	Coding error (three independent experiments)
	Choi et al. (2008) [study 1]	-	-		0.050 (0.049)		
	Choi et al. (2008) [study 2]	-	-		0.084 (0.190)		
	Choi et al. (2008) [study 3]	-	-		-0.034 (0.171)		
[40]	Han et al. 2007 (study 1)	0.06	Save	2009	0.064 (0.005)	D (Save)	Agreement in sign and magnitude
[41]	Han et al. 2007 (study 2)	0.06	Save		-		Counted as two studies
[75]	Mills et al. (2004) (sample 1)	-0.02	Save, Plan	2004	-0.033 (0.019)	B (Credit), D (Save)	Agreement in sign
[76]	Mills et al. (2004) (sample 2)	0.03	Save, Plan		-	B (Credit), D (Save)	Counted as two independent samples

Notes: This table compares our data to the extracted estimates reported in Fernandes et al. (2014) (Tables WA1 and WA 2). The measure of effect size is (partial) *r* as in Fernandes et al. (2014).

Table D2: Replication result

		Panel A: Replication of Table WA1					
	Fernandes et al. (2014, p.1864)	(1) Unrestricted WLS	(2) Fixed-effect Meta-Analysis	(3) Random-effects (REML)	(4) WLS (Cluster-robust SE)	(5) RVE (Fixed-Effect)	(6) RVE (Random-Effects)
r	0.009	0.012	0.012	0.018	0.013	0.017	0.021
(Std. Err.)	(0.0066)	(0.004)	(0.003)	(0.005)	(0.003)	(0.005)	(0.006)
[CI$_{95}$]	[-0.004, 0.022]	[0.004, 0.021]	[0.006, 0.019]	[0.007, 0.028]	[0.006, 0.021]	[0.004, 0.031]	[0.006, 0.035]
g	0.018	0.025	0.025	0.035	0.026	0.035	0.041
(Std. Err.)	(0.013)	(0.008)	(0.007)	(0.011)	(0.007)	(0.010)	(0.012)
[CI$_{95}$]	[-0.008, 0.044]	[0.008, 0.042]	[0.012, 0.037]	[0.014, 0.056]	[0.011, 0.041]	[0.008, 0.061]	[0.012, 0.071]
n (RCTs)	15	12	12	12	12	12	12
n (ES)	15	12	12	12	36	36	36
		Panel B: Adding falsely classified studies from Table WA2					
r	-	0.028	0.028	0.023	0.012	0.018	0.023
(Std. Err.)	-	(0.007)	(0.003)	(0.008)	(0.004)	(0.006)	(0.008)
[CI$_{95}$]	-	[0.013, 0.042]	[0.023, 0.033]	[0.007, 0.039]	[0.003, 0.021]	[0.004, 0.032]	[0.006, 0.040]
g	-	0.055	0.055	0.046	0.024	0.036	0.046
(Std. Err.)	-	(0.013)	(0.005)	(0.017)	(0.009)	(0.011)	(0.015)
[CI$_{95}$]	-	[0.027, 0.085]	[0.045, 0.066]	[0.013, 0.079]	[0.006, 0.042]	[0.008, 0.064]	[0.012, 0.079]
n (RCTs)	-	17	17	17	17	17	17
n (ES)	-	17	17	17	51	51	51