Appendix

to accompany

"Financial education affects financial knowledge and downstream behaviors"

Tim Kaiser, Annamaria Lusardi, Lukas Menkhoff and Carly Urban

Appendix A: Studies included in the meta-analysis

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

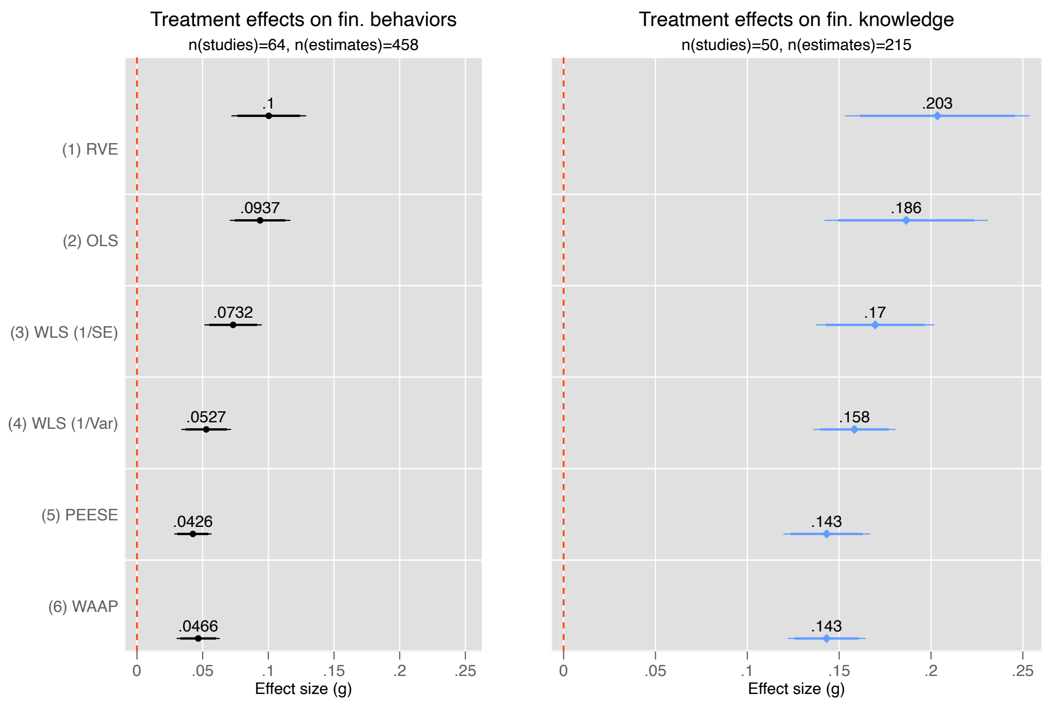
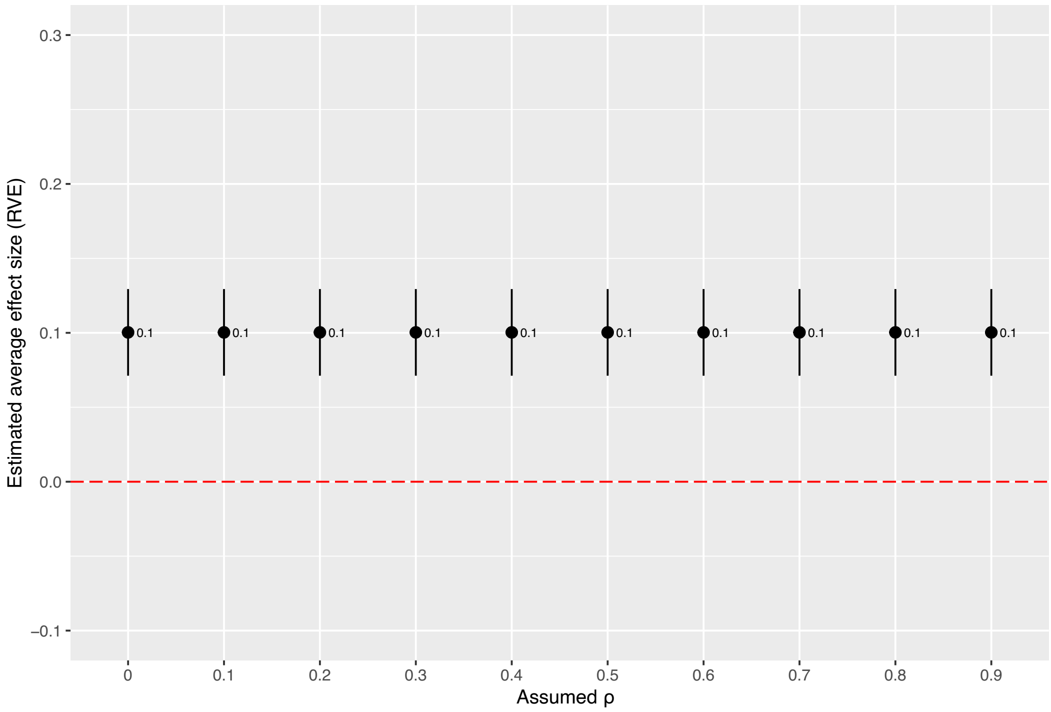
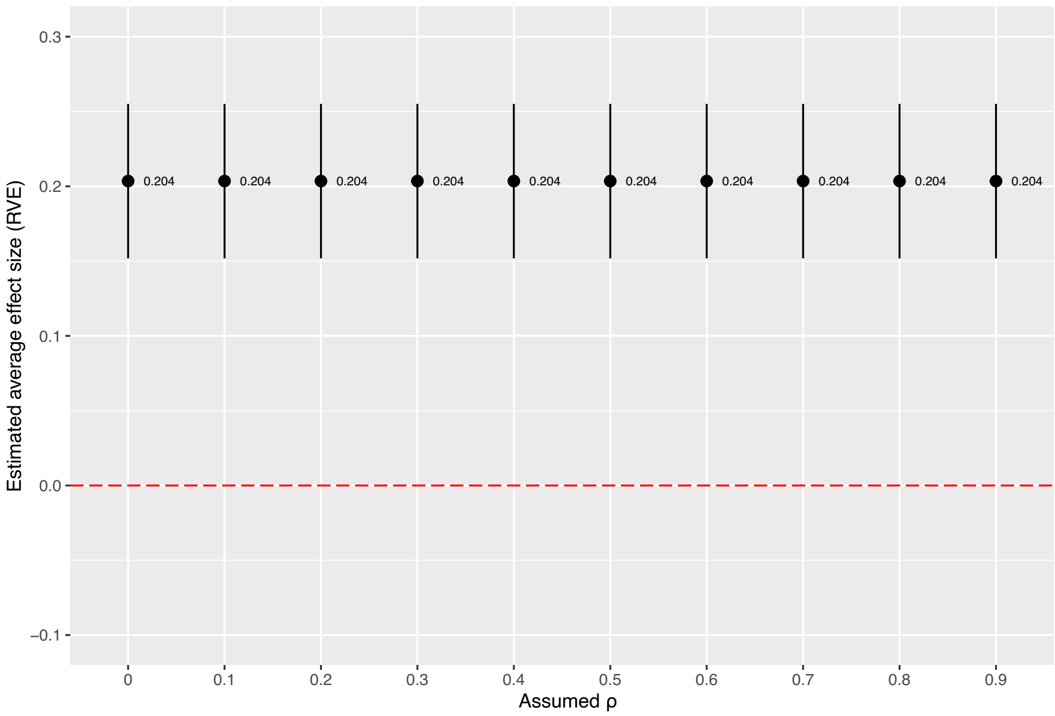


Figure B2: (Non-)Sensitivity of RVE estimate to the choice of ρ (treatment effects on financial behaviors)



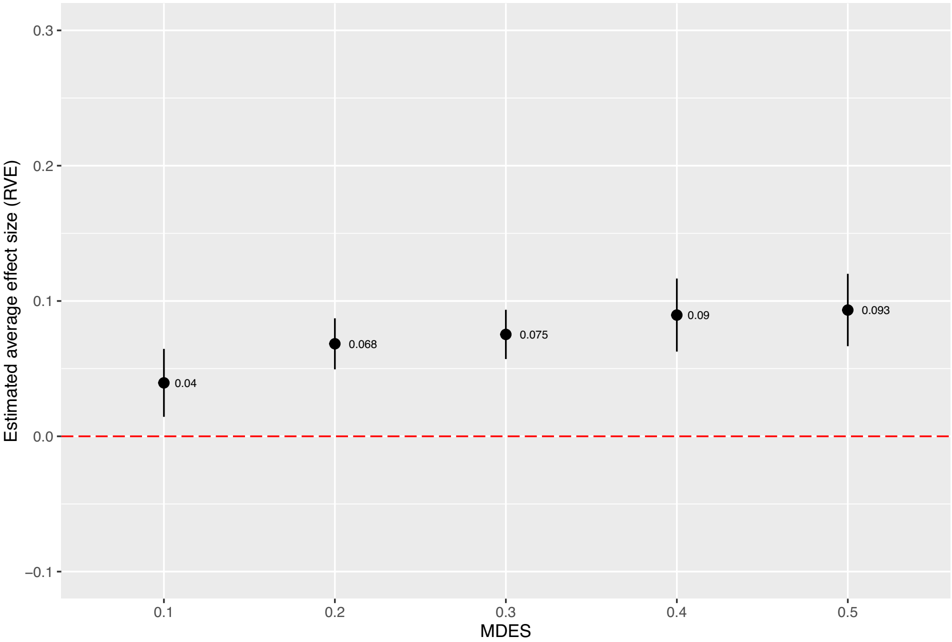
Notes: Figure shows results from (random effects) RVE for different choices of assumed ρ.

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



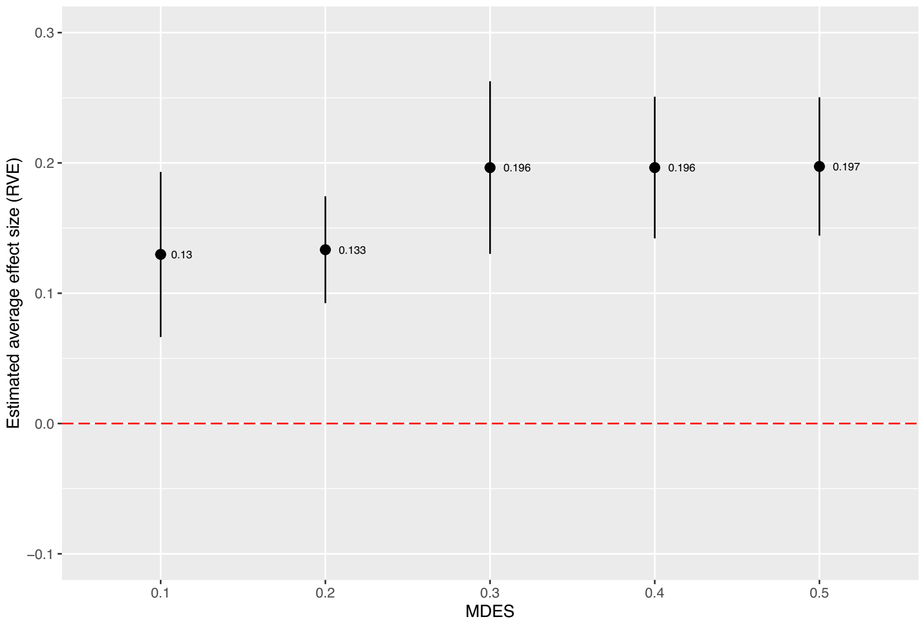
Notes: Figure shows results from (random effects) RVE for different choices of assumed ρ.

Figure B4: Power in the financial behavior sample



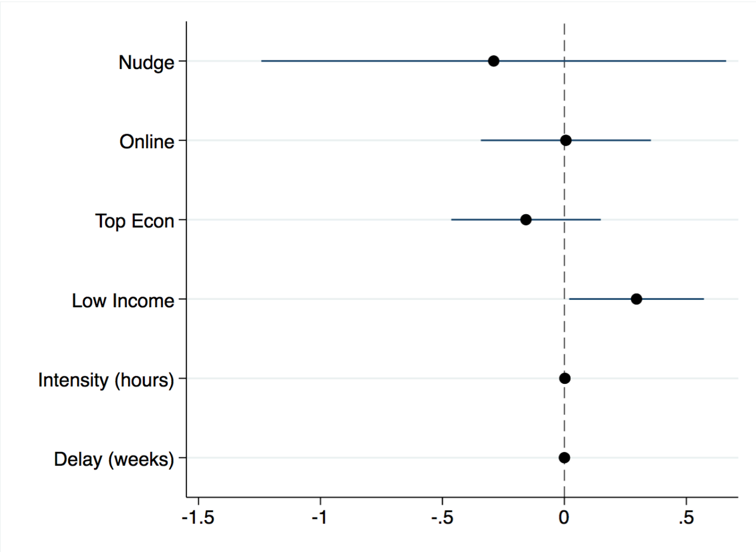
Notes: Average effect size of treatment effects on financial behaviors (from RVE) within the set of studies with the respective MDES. Minimum detectable effect size (MDES) at $\alpha = 0.05$ and $1 - \beta = 0.8$. The number of observations for the sample with a MDES of 0.1 is 60 effect sizes within 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



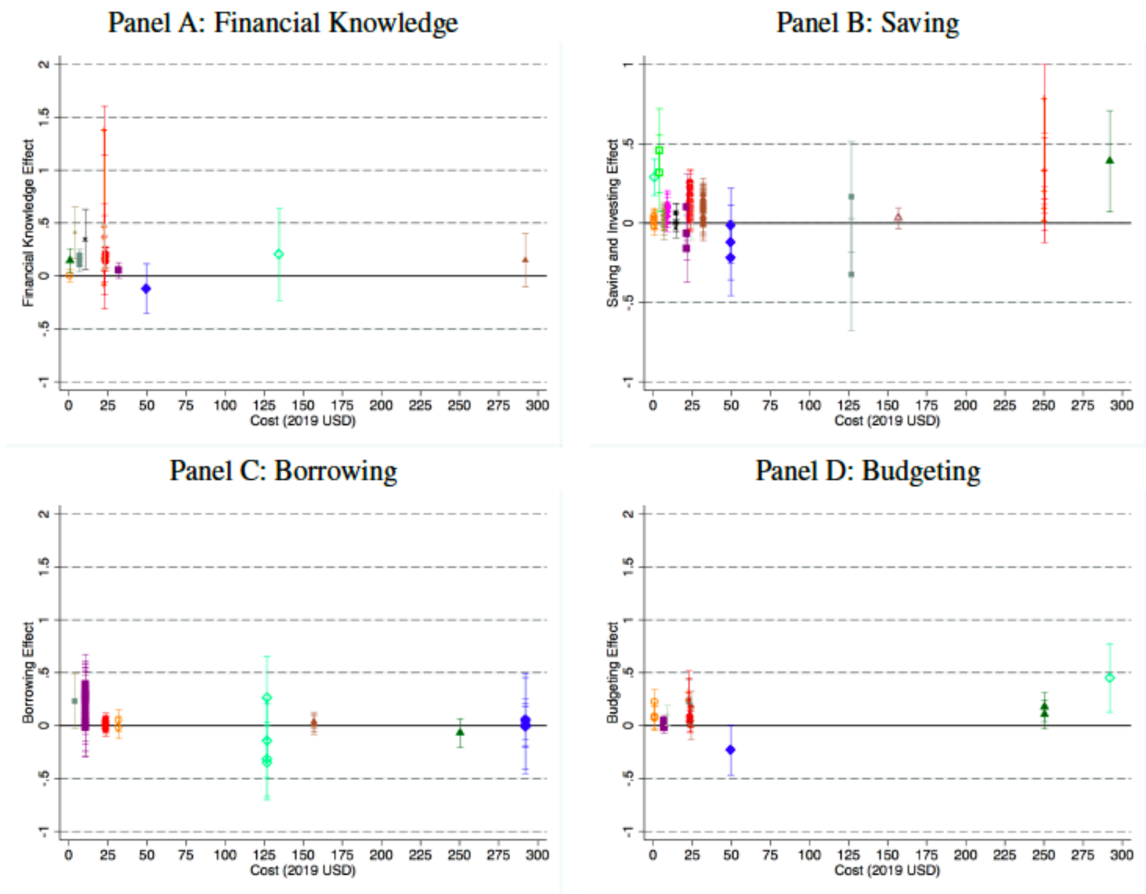
Notes: Average effect size of treatment effects on financial knowledge (from RVE) within the set of studies with the respective MDES. Minimum detectable effect size (MDES) at $\alpha = 0.05$ and $1 - \beta = 0.8$. The number of observations for the sample with an MDES of 0.1 is 12 effect sizes within 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) Unadjusted	(2) FAT-PET (1/SE)	(3) PEESE ($1/SE^2$)	(4) Unadjusted	(5) FAT-PET (1/SE)	(6) PEESE ($1/SE^2$)
SE		0.731*** (0.243)		0.187*** (0.022)	0.846 (0.524)	
SE^2			5.360*** (1.514)			4.538 (2.736)
Average effect	0.093*** (0.015)	0.032** (0.013)	0.0426*** (0.007)		0.113*** (0.030)	0.143*** (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)
[CI95]	[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)
[CI95]	[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)
[CI95]	[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)[1]	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)[2]	No	Yes
36)	Cole et al. (2011)	Yes (coding error)[3]	Yes	Yes
37)	Collins (2013)	Yes (2011 WP)	No	Yes
38)	Collins and Urban (2016)	No	No	No

[1] Wrongly classified as quasi-experiment and not included in the RCT sample (see Appendix D).
[2] Wrongly coded estimate (wrong sign and magnitude) and misclassified financial behavior as savings when it is in the insurance domain (see Appendix D).
[3] Wrongly coded multiple time-points within the same study as independent samples (see Appendix D).

39)	Custers (2011)	No	No	Yes
40)	Doi et al. (2014)	No	Yes (2012 WP)	Yes
41)	Drexler et al. (2014)	Yes (coding error)[4]	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)[5]	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)[6]	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)[7]	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

[4] Wrongly coded multiple treatments as independent samples even though they are compared to a common control group (see Appendix D).
[5] Wrongly classified as quasi-experiment and not included in the RCT sample (see Appendix D).
[6] Wrongly classified as quasi-experiment and not included in the RCT sample (see Appendix D).
[7] 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).[8]

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

[8] 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).[9]

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

[9] 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).[10]

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

[10] 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.[11] 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

[11] 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).[12] 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.

[12] 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.[13]

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.

[13] 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