Schlechte Evaluierung rentiert sich kaum: Lehren aus dem Bereich der finanziellen Bildung

Von Tim Kaiser und Lukas Menkhoff

Die Verbesserung finanzieller Bildung ist inzwischen weltweit ein etabliertes Ziel der Wirtschaftspolitik, das über vielfältige Initiativen erreicht werden soll. Es verfügen aber nur wenige Wirkungsevaluierungen über den wissenschaftlich wünschenswerten Stand, um die Leistungen dieser Initiativen sicher bewerten zu können. Dieser Bericht erläutert die verschiedenen praktizierten Formen der Evaluierung und zeigt, dass „schlechte" Evaluierung, zum Beispiel in Form von bloßen Vorher-Nachher-Vergleichen, zu verzerrten Bewertungen führt und typischerweise das Ergebnis beschönigt. Diese Schwächen bedeuten, dass die Träger finanzieller Bildungsmaßnahmen und die wirtschaftspolitisch Verantwortlichen die Wirkung ihrer Aktion eventuell überschätzen und die wahren Probleme nicht gut erkennen. Von methodisch schlechten Evaluierungen sollte abgesehen werden, um stattdessen fachgerechte Evaluierungsmethoden zu definieren und einzusetzen.

Die unzureichende finanzieller Bildung vieler Menschen ist ein identifiziertes und detailliert beschriebenes Problem.[1] Die OECD hat bereits im Jahr 2005 dazu aufgefordert, finanzielle Bildung in Schulen zu verankern[2] und die Erfassung finanzieller Kompetenz im Rahmen der international vergleichenden Schulleistungsstudie PISA wurde erstmals im Jahr 2012 durchgeführt. Zahlreiche Initiativen und Maßnahmen sollen zudem die Finanzbildung verbessern: So werden vielerorts am Arbeitsplatz Trainings zur Altersvorsorge angeboten, auch existieren individuelle Trainingsangebote zu Teilaspekte wie Sparen oder Schulden.[3] Ziel dieser Angebote ist es, die Kenntnis über Finanzprodukte und Zusammenhänge zu verbessern und auf das Finanzverhalten der einzelnen Akteure einzuwirken, um ihnen adäquate Konsum-, Kredit- oder Anlageentscheidungen zu ermöglichen.

Diesen verschiedenen Maßnahmen konnten bisher aber keine eindeutig erwiesenen Erfolge zugeschrieben werden.[4] So wurde ein „Mangel jeglicher zwingenden Evidenz einer positiven Wirkung [von finanzieller Bildung]" konstatiert.[5] Offensichtlich herrscht eine große Diskrepanz zwischen dem großen Aufwand für die Durchführung der Maßnahmen und dem gerin-

1 Annamaria Lusardi und Olivia S. Mitchell (2014): The economic importance of financial literacy: Theory and evidence. Journal of Economic Literature, 52(1), 5–44; Leora Klapper, Annamaria Lusardi und Peter van Oudheusden (2015): Financial literacy around the world: Insights from the Standard and Poor's Rating Services global financial literacy survey (online verfügbar, abgerufen am 19. Juni 2017. Dies gilt, insofern nicht anders vermerkt, für alle Online Quellen in diesem Bericht).

2 Vgl. OECD (2005): Recommendation on Principles and Good Practices for Financial Education, OECD Publishing.

3 Für eine detailliertere Auflistung möglicher Maßnahmen siehe Tim Kaiser und Lukas Menkhoff (2017): Does financial education impact financial literacy and financial behavior, and, if so, when? World Bank Economic Review (forthcoming) und DIW Discussion Paper 1562 (revised).

4 Justine S. Hastings, Brigitte C. Madrian und William L. Skimmyhorn (2013): Financial literacy, financial education, and economic outcomes. Annual Review of Economics, 5, 347–373.

5 Daniel Fernandes, John G. Lynch Jr. und Richard G. Netemeyer (2014): Financial literacy, financial education, and downstream financial behaviors. Management Science, 60(8), 1861–1883.

gen Wissen über ihre Wirksamkeit. Studien kommen sogar zu gegensätzlichen Urteilen über die Effektivität finanzieller Bildung, teilweise weil unterschiedliche Maßstäbe an den Erfolg angelegt werden. Diese Situation ist für Träger finanzieller Bildungsmaßnahmen und die wirtschaftspolitisch Verantwortlichen unbefriedigend.

Dieser Wochenbericht widmet sich der Beantwortung von drei Fragen: Worauf kommt es bei einer guten Evaluierung von Bildungsmaßnahmen an? Wie unterscheiden sich hier „schlechte" und „gute" Evaluierung? Kann man überhaupt eine Aussage darüber treffen, ob finanzielle Bildung eine erwünschte Wirkung hat? Anhand eines Samples von Studien zur Wirkung finanzieller Bildungsmaßnahmen wird gezeigt, dass falsche Evaluierung zu „geschönten" Ergebnissen führt. Wissenschaftlich überzeugende Evaluierung stellt weitaus höhere Anforderungen an Wirksamkeit.

Die Praxis der Wirkungsevaluierung ist sehr heterogen

Wirtschaftspolitische Maßnahmen werden nicht immer evaluiert: Wenn die Zusammenhänge gut bekannt sind, spart man sich lieber den Aufwand. Bei neuartigen Maßnahmen ist eine begleitende Wirkungsevaluierung dennoch sinnvoll, weil man zum einen wissen möchte, ob die Maßnahmen überhaupt zur Zielerreichung beitragen und zum anderen, wie man die Effektivität eventuell verbessern kann. Im Bereich der finanziellen Bildung, in dem seit mehreren Jahrzehnten verschiedene Träger aktiv sind, wird seit langem immer wieder evaluiert, häufig jedoch nicht nach den derzeit wissenschaftlich wünschenswerten Standards.

Einfache Ex post-Befragungen

Eine einfache und preiswerte Form der Evaluierung ist die Ex-post-Befragung der TeilnehmerInnen. Eine solche Methode schreibt sämtliche Änderungen den Ergebnissen der Intervention zu. Nach einem einwöchigen Training zur Finanzbildung werden beispielsweise die TeilnehmerInnen zum Abschluss nach ihren Erfahrungen gefragt. Die Fragen zielen darauf ab, zu erfahren, ob es Lernerfolge gab, worin diese bestehen, ob es zu geplanten Verhaltensänderungen kommt oder was man verbessern könnte. Das Hauptproblem solch einer Befragung ist offensichtlich, dass es sich um qualitative Selbsteinschätzungen handelt, die zudem noch unter dem unmittelbaren Eindruck des Trainings stehen, sodass die Wirkung tendenziell überschätzt wird. So gewonnene Erkenntnisse können hilfreich sein, sollten aber nicht mit einer Evidenz für eine kausale Wirkung der Maßnahme verwechselt werden.

Ex-ante- und Ex-post-Befragungen

Etwas aufwendiger, aber potentiell informativer sind zweimalige Befragungen, also einmal vor einem Training (ex ante) und ein zweites Mal danach (ex post). Aus der Befragung zu Beginn ergibt sich eine sogenannte Baseline-Information, die den Referenzpunkt für gemessene Veränderungen bildet. Allerdings bleibt unklar, ob während des Trainings – gerade wenn es sich über längere Perioden hinzieht – andere Dinge eintreten, die unabhängig von der Intervention einen Einfluss auf die gemessenen Werte haben können. Vielleicht haben die Medien Finanzthemen stark aufgegriffen, weil es zufällig – parallel zur Bildungsmaßnahme – starke Verwerfungen im Finanzbereich gab. Dann hätten sich die TeilnehmerInnen eventuell stark mit Finanzthemen beschäftigt und etwas gelernt, aber nur teilweise wegen der Intervention. Ein Kontrafaktum wird nicht beobachtet.

Evaluierung mit Vergleichsgruppe

Deshalb werden Evaluierungen bevorzugt mit einer Vergleichsgruppe durchgeführt (vgl. Kasten 1). Um im genannten Beispiel zu bleiben, würden neben den TeilnehmerInnen der Bildungsmaßnahme zur gleichen Zeit und in vergleichbarer Form auch andere Personen befragt, die kein Training erhalten. Man kann im Nachhinein die Veränderung bei der Zielgruppe mit Training mit einer möglichen Veränderung bei der Kontrollgruppe vergleichen. Die Selektion der TeilnehmerInnen ist hierbei die größte Herausforderung. Die Praxis zeigt, dass Menschen, die an Trainings teilnehmen, besondere Züge aufweisen: Sie haben beispielsweise ein besonders großes Interesse am Thema, sind überhaupt neugierig und aufgeschlossen für Neues. Es verwundert dann wenig, dass die Trainings mit solchen Personen gute Effekte hervorbringen, das ist aber nicht verallgemeinerungsfähig. Dem kann man partiell entgegen wirken, indem man für beobachtbare Charakteristika der Personen kontrolliert und gegebenenfalls korrigiert (sogenannte Matching-Verfahren).

Randomisierte Evaluierung

Ein probateres Mittel, um eine solche Kontrollgruppe zu identifizieren, und somit ein valides Kontrafaktum zu generieren, ist die Randomisierung: Hierbei entscheidet allein der Zufall, ob ein Individuum in der Treatment- oder der Kontrollgruppe landet und nicht etwa das Individuum selbst:[6] Beispielsweise wird eine Gruppe per Zufallsentscheidung, wie den Wurf eines Würfels, in zwei Untergruppen aufgeteilt. Dadurch kann – bei einer großen Zahl von Individuen – davon ausgegangen werden, dass beide

6 Vgl. Julian C. Jamison (2016): The entry of randomized assignment into the social sciences. Working Paper (online verfügbar).

Kasten 1

Anforderung der Messung kausaler Wirkung

Wirkungsevaluierung meint die Identifikation von kausalen Effekten einer Intervention auf eine oder mehrere Ergebnisvariable(n). Während die Identifikation von kausalen Wirkungszusammenhängen in den Naturwissenschaften in der Regel durch die Isolierung von Umwelteinflüssen im Labor ohne Gefahr von Fehlschlüssen (hohe interne Validität) zu realisieren ist, gestaltet sich Wirkungsevaluierung in den Sozialwissenschaften häufig schwieriger: Menschliches Handeln ist selten unter Laborbedingungen zu beobachten, und wenn dies der Fall ist (zum Beispiel bei computergestützten Labor-Experimenten zur Überprüfung von ökonomischen Theorien), dann stellt sich die Frage nach der Generalisierbarkeit der ermittelten Ergebnisse (externe Validität) auf Situationen außerhalb des Labors.

Wirkungsevaluierung von Politikmaßnahmen, wie zum Beispiel einer Bildungsmaßnahme im Bereich finanzieller Bildung, bedarf daher der Beobachtung von menschlichem Handeln im natürlichen Raum (Feld), ist dennoch an einer möglichst genauen Identifikation von kausalen Wirkungszusammenhängen (hohe interne Validität) interessiert. Um die Unterschiede in der Qualität und Glaubwürdigkeit der verschiedenen Forschungs-Designs zu verstehen, hilft es, die allgemeine Frage nach der kausalen Wirkung einer Maßnahme zu formalisieren und zwei Kernkonzepte der Wirkungsevaluierung (kausale Inferenz und Kontrafaktum) näher zu erläutern:[1]

Unter Wirkung (Δ) wird die Differenz zwischen einer Ergebnisvariable einer Person mit Intervention ($Y|\ I=1$) und derselben Ergebnisvariable derselben Person ohne Intervention ($Y|\ I=0$) verstanden: $\Delta = (Y|\ I=1) - (Y|\ I=0)$. $I=1$ könnte hierbei eine Maßnahme zur Stärkung finanzieller Bildung bedeuten; das interessierende Ergebnis Y wäre beispielsweise die Summe der Ersparnisse in Euro oder der Anteil von richtig beantworteten Fragen in einem Leistungstest zum finanziellen Wissen. Man würde also gerne die Ergebnisvariable (zum Beispiel die Ersparnisse) zum selben Zeitpunkt für dieselbe Person mit und ohne Intervention beobachten. Es ist offensichtlich, dass diese Art von Beobachtung nicht möglich ist, da zum Zeitpunkt der (Nicht-) Intervention die Entscheidung für einen Zustand getroffen wurde, der die gleichzeitige Beobachtung des alternativen Zustandes unmöglich macht.

Wissenschaftliche Wirkungsevaluierung versucht daher ein alternatives, valides *Kontrafaktum* zu identifizieren, um den kausalen Effekt der Intervention möglichst präzise zu schätzen. Hierzu wird die Wirkung auf Ebene von Gruppen gemessen: Existiert eine Gruppe mit Intervention (Treatment/Interventions-Gruppe) und eine Gruppe ohne Intervention (Kontroll/Vergleichs-Gruppe) so kann man die Mittelwerte für die Ergebnisvariable beider Gruppen miteinander vergleichen – wenn interne Validität gegeben ist, spiegelt die Mittelwertdifferenz den kausalen Effekt wieder. So betrachtet besteht ein Kernproblem der Wirkungsevakuierung darin, eine valide Kontrollgruppe zu finden.[2]

1 Vgl. Paul J. Gertler et al. (2016): Impact evaluation in practice. World Bank Publications.

2 Vgl. Paul J. Gertler et al. (2016), a.a.O., 52ff.

Gruppen im Mittel vollkommen identisch sind und zwar sowohl in Bezug auf beobachtbare Merkmale, aber auch – und dies ist wichtig – in Bezug auf nicht beobachtbare Merkmale. Der einzige systematische Unterschied zwischen den beiden Gruppen ist dann die experimentelle Intervention (zum Beispiel die Bildungsmaßnahme). Dies ist das ideale sozialwissenschaftliche Experiment und die sauberste Methode, kausale Effekte zu identifizieren. Ihre Bedeutung hat jüngst stark zugenommen, allerdings existieren auch Grenzen (vgl. Kasten 2).

Schlechte Evaluierung finanzieller Bildungsmaßnahmen überschätzt deren Effektivität

Von den vier beschriebenen Formen der Evaluierung wird im Folgenden nur noch auf die zwei oder drei besten Formen eingegangen – Ex-Post-Befragungen werden also außen vor gelassen. Konkret basiert dieser Bericht auf einer Analyse von 143 Arbeiten.[7] Diese Studien wurden in drei Gruppen klassifiziert, je nach wissenschaftlicher Qualität der Evaluierungsmethode. Gruppe 1 bilden hierbei Studien, die Vorher-Nachher-Vergleiche von nur einer Gruppe verwenden. Gruppe 2 sind Studien, die auf Quasi-Experimenten oder natürlichen Experimenten beruhen (beobachtetes Kontrafaktum). Gruppe 3 besteht aus Studien, die randomisierte Evaluierungen nutzen.

7 Vgl. Tim Kaiser und Lukas Menkhoff (2017), a. a. O. In dieser Meta-Analyse arbeiten wir mit einem Datensatz aus 126 Studien. Die 17 weiteren Studien wurden im Rahmen einer systematischen Literaturrecherche identifiziert und wegen Nicht-Erfüllung der Inklusionskriterien hinsichtlich der methodischen Qualität nicht in die Empirische Analyse eingeschlossen.

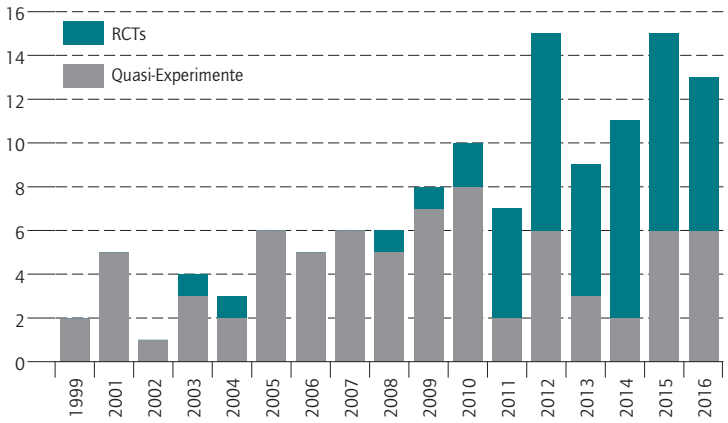
Kasten 2

Verbreitung und Grenzen randomisierter Evaluierung

Randomisierte Evaluierungen haben sich erst seit den neunziger Jahren in der Ökonomie durchgesetzt. Im Bereich der finanziellen Bildung gab es in den letzten zehn Jahren ein deutlicher Anstieg von experimentellen Evaluierungen (Abbildung). Pro Kalenderjahr sind alle aussagekräftigen Evaluierungsstudien ermittelt worden und davon der Anteil an randomisierten Studien, sogenannten RCTs (Randomized Controlled Trials).

Abbildung

Anzahl der Studien zur Wirkung finanzieller Bildung nach Forschungsdesign pro Jahr



Quelle: Eigene Berechnungen.

Randomized Controlled Trials machen seit fünf Jahren die Mehrheit der Studien aus.

RCTs gelten als besonders aussagekräftig im Hinblick auf kausale Aussagen. Allerdings sind die Anforderungen an RCTs hoch, was den Mitteleinsatz anbelangt, und sie sind spezifisch, was die Fragestellung anbelangt. So gibt es neben Interventionen, die lediglich Opportunitätskosten verursachen (wie Bildungsmaßnahmen) auch Interventionen (beispielsweise in der Medizin), die aus ethischen Gründen nicht am Menschen durchführbar sind. Daneben kann es Ereignisse geben, die nicht in Experimenten herbeigeführt werden können, wie Naturkatastrophen, Finanzkrisen oder große Innovationen. Manche Fragen erfordern für ihre Analyse auch lange Zeiträume, für die sich RCTs nicht eignen, unter anderem weil man die Umfeldbedingungen mit den Jahren immer schlechter kontrollieren kann.

Eine randomisierte Evaluierung ist vergleichsweise aufwendig und leider nicht immer durchführbar. Eine wichtige Bedingung ist dabei, dass die zu vergleichenden Gruppen (mit und ohne Intervention) ähnlich sind, was für eine räumliche Nähe spricht, aber gleichzeitig darf es keine Übertragung (Spillover-Effekte) von den Trainierten zu den Untrainierten geben. Weiterhin bestehen häufiger ethische oder praktische Vorbehalte, manche Personen in den Genuss einer Intervention kommen zu lassen und andere nicht. Manchmal wird die Intervention mit der anfangs unberücksichtigten (Kontroll-) Gruppe nachgeholt; dennoch erweist sich diese empfundene Ungleichbehandlung in der Praxis als ein Nachteil einer randomisierten Evaluierung.

Trotz all dieser Einschränkungen lässt sich aber feststellen, dass RCTs sehr geeignet sind, um kurz- und mittelfristige Effekte von Bildungsmaßnahmen zu evaluieren, was häufig der relevante Zeithorizont ist.

Um Vergleichbarkeit hinsichtlich der Wirksamkeit finanzieller Bildungsmaßnahmen herzustellen, werden für alle Studien sogenannte Effektstärken[8] berechnet, also ein standardisiertes Maß, das etwas über die Wirksamkeit der Bildungsmaßnahme im Hinblick auf den angestrebten Erfolg aussagt (Abbildung 1).

Die weniger rigorosen Gruppe-1-Studien weisen demnach zwei Eigenschaften auf: Erstens liegt ihre durchschnittliche Effektstärke deutlich über derjenigen der Gruppen 2 und 3. Breite Konfidenzintervalle bei den Ergebnissen weisen zweitens auf eine erhebliche Unsicherheit in Bezug auf die geschätzten Effekte hin. Offensichtlich liefern schwächere Evaluierungsmethoden tendenziell ein nach oben verzerrtes und zugleich unpräziseres Bild. Deshalb werden weitere Fragen nur noch anhand der 126 Studien der Gruppe 2 und 3 erläutert.

Selbst bei rigorosen Studien beeinflussen die Forschungsmethoden die Ergebnisse stark

Für die Studien der Gruppen 2 und 3 wird die Wirkung, die eine Variation der Messung auf die geschätzte Effek-

8 Mark W. Lipsey und David B. Wilson (2001). Practical meta-analysis. Sage, Thousand Oaks, CA.

tivität der Bildungsmaßnahme hat, untersucht. Es geht hier also ausschließlich um die Beeinflussung des Messergebnisses durch die Art der Messung. Hierzu werden vier mögliche Einflüsse betrachtet:

1. Studiendesign: Berücksichtigt wird, ob die Studie von ihrem Design her dem höchsten Anspruch gerecht wird und als Randomized Controlled Trial (RCT) angelegt ist. Da RCTs am stärksten positiven Effekten durch Selbstselektion entgegenwirken (siehe Kasten 2), wird hier ein negatives Vorzeichen des Koeffizienten erwartet.

2. Erfassung der Zielgruppe: Bei einer Bildungsintervention wird vorab definiert, wer zur Zielgruppe gehört. Bei der Messung der Effektivität hat man dann – sofern das Studiendesign diese Unterscheidung zulässt – zwei Möglichkeiten, die Zielgruppe zu definieren. Man misst entweder die Effektivität für diejenigen, die wirklich an der Maßnahme teilgenommen haben, das ist der sogenannte TOT-Effekt (*treatment on the treated*), oder man misst die Wirkung bei der ursprünglich angestrebten Zielgruppe, der sogenannte ITT-Effekt (*intended to treat*). Da typischerweise nicht alle Mitglieder einer Zielgruppe erreicht werden, weil manche nicht kooperieren, liegt der TOT-Effekt systematisch höher. Dieser höhere Effekt wäre weniger wichtig, wenn er zufällig auftreten würde. Tatsächlich jedoch zeigt die Praxis, dass gerade diejenigen durch Bildungsmaßnahmen schwer zu erreichen sind, die es am nötigsten hätten und deshalb ausdrücklich zur Zielgruppe gehören.

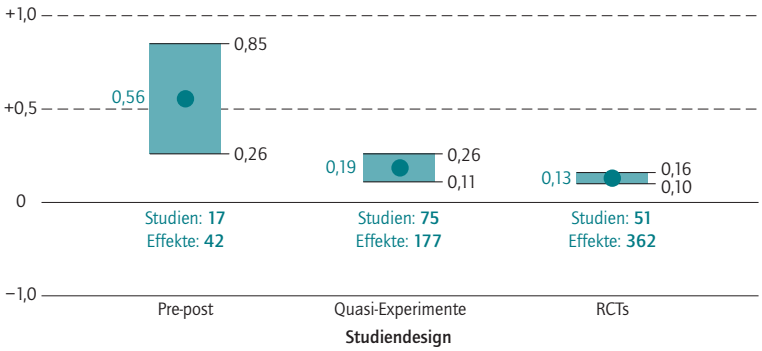
3. Messabstand: Typischerweise lässt die Erinnerung an Gelerntes mit der Zeit nach. Folglich wirkt sich der zeitliche Abstand zwischen einer Bildungsintervention und der Messung des Erfolgs nachteilig auf die gemessene Effektivität aus.

4. Studiengröße: Studien mit mehr TeilnehmerInnen weisen tendenziell kleinere Standardfehler für jede Schätzung aus. Insofern führen größere Studien zu mehr Verlässlichkeit in der Aussage.

In dem untersuchten Sample aus 126 Studien erklären die kodierten Variablen zum Studiendesign rund 16 Prozent der Varianz in den Effektstärken (Tabelle 1). Im Einzelnen wird ersichtlich, dass RCTs tendenziell kleinere Effektstärken berichten (negatives Vorzeichen), dass die Operationalisierung als TOT-Effekt mit höheren geschätzten Effektstärken einhergeht, dass ein größerer Abstand von Intervention und Messung höhere Effekte erwarten lässt (jedoch nicht mit einem besonders starken Einfluss), und dass Studien mit größeren Standardfehlern systematisch zu einer Verzerrung zugunsten von zu optimistisch eingeschätzter Wirksamkeit gelangen.

Abbildung 1

Empirischer Unterschied in den berichteten Effektstärken (Hedges' g) nach Studiendesign von Maßnahmen finanzieller Bildung



Quelle: Eigene Berechnungen.

Weniger rigorose Studiendesigns führen zu nach oben verzerrten und mit hoher Unsicherheit geschätzten Ergebnissen.

Tabelle 1

Einflüsse des Studiendesigns auf die geschätzten Effektstärken von Maßnahmen finanzieller Bildung

	(1)
	Abhängige Variable: Effektstärke
RCT	−0,073*
	(0,041)
TOT	0,092*
	(0,049)
Delay	−0,000***
	(0,000)
SE	1,132**
	(0,552)
Konstante	0,090**
	(0,035)
Adj. R2	0,160
n (Studien)	126
n (Effektstärken)	539

Anmerkung: Koeffizienten sind Ergebnisse einer OLS Regression. Abhängige Variable ist Effektstärke einer Intervention zur finanziellen Bildung auf finanzielles Wissen oder finanzielles Verhalten. Standardfehler (geclustered auf Studienebene) in Klammern unter den Koeffizienten.

** p<0.1, ** p<0.05, *** p<0.01*

Quelle: Eigene Berechnungen.

Tabelle 2

Zusammenfassung der durchschnittlichen Wirksamkeit finanzieller Bildung in der Literatur

Abhängige Variable:	Finanzielles Wissen		Finanzielles Verhalten	
Effektstärke	Alle Studien	Nur RCTs	Alle Studien	Nur RCTs
(1) OLS	0,263***	0,209***	0,086***	0,082***
	(0,041)	(0,033)	(0,012)	(0,014)
(2) WLS (1/SE)	0,191***	0,175***	0,026**	0,067***
	(0,025)	(0,024)	(0,011)	(0,013)
(3) WLS (1/SE2)	0,135***	0,155***	0,002*	0,051***
	(0.,019)	(0,015)	(0,001)	(0,010)
(4) Metareg	0,297***	0,234***	0,079***	0,075***
	(0,042)	(0,039)	(0,009)	(0,013)
(5) Robumeta	0,287***	0,237***	0,064***	0,078***
	(0,037)	(0,039)	(0,008)	(0,012)
n(Studien)	67	33	90	40
n(Effektstärken)	190	135	349	227

Anmerkung: Abhängige Variable ist die Effektstärke (Hedges' g).

Die Tabelle zeigt Ergebnisse verschiedener Meta-Analyse-Modelle. Ausgewiesen sind die (gewichteten) mittleren Effektstärken pro Outcome-Typ und Studiendesign. Standardfehler sind in Klammern.

** p<0.1, ** p<0.05, *** p<0.01*

Zeile (1) zeigt die Ergebnisse einer ungewichteten OLS-Regression. Zeile (2) zeigt Ergebnisse einer unrestricted weighted least squares regression mit dem inversen Standardfehler als Regressionsgewicht, vgl. Tom D. Stanley und Hristos Doucouliagos (2015). Neither fixed nor random: weighted least squares meta-analysis. Statistics in Medicine 34(13), 2115–2127.vgl. Zeile (3) zeigt eine identische Schätzung mit der inversen Varianz als Gewicht. Zeile (4) zeigt Ergebnisse einer Random-Effects-Metaregression mit korrigierten Standardfehlern nach Knapp und Hartung, vgl. Rebecca DerSimonian und Nan Laird (1986). Meta-analysis in clinical trials. Controlled clinical trials, 7(3), 177–188; Guido Knapp und Joachim Hartung (2003). Improved tests for a random effects meta- regression with a single covariate. Statistics in Medicine, 22(17), 2693–2710. Diese Schätzung arbeitet mit einer (synthetischen) Effektstärke pro Studie. Zeile (5) zeigt Ergebnisse einer Robust variance meta-regression with dependent effect size estimates, vgl. Larry V. Hedges, Elizabeth Tipton und Matthew C. Johnson. (2010). Research Synthesis Methods. 1(1), 39–65.

Der signifikante Koeffizient des Standardfehlers der Effektstärke (SE) gibt zudem einen Hinweis auf einen vorhandenen Publikationsbias: Wenn Effektstärken systematisch mit dem dazugehörigen Standardfehler im Zusammenhang stehen, bedeutet dies, dass Studien mit insignifikanten Ergebnissen seltener publiziert werden. Dies bezeichnet man als *file drawer problem*, weil ForscherInnen „schwache" Forschungsergebnisse „in eine Schublade verbannen".[9] Dieses Publikationsverhalten verzerrt eine Analyse der durchschnittlichen Wirksamkeit tendenziell nach oben, da die insignifikanten Ergebnisse nicht im Datensatz vorhanden sind. Eine andere Erklärung für diesen positiven Zusammenhang können allerdings auch ex-ante Power-Berechnungen sein, die die Samplegröße schon vor Beginn der Studie an die erwarteten Effektstärken anpassen.

Festhalten lässt sich, dass die gemessene Wirkung einer Maßnahme größer sein wird, wenn die Evaluierung nicht als RCT erfolgt, wenn der TOT-Effekt angegeben wird und/oder wenn der Messabstand möglichst geringgehalten wird. Schließlich sind Aussagen aus Studien mit geringerer Varianz (zugleich eher größere Studien) tendenziell verlässlicher.

Finanzbildung ist wirksam, finden auch rigorose Studien

Wie bereits erwähnt, sind die Urteile zur Wirksamkeit recht unterschiedlich. Neben den qualitativen (narrativen) Überblicksarbeiten, die auf subjektiver Auswahl und Interpretation der Literatur beruhen, gibt es quantitative Meta-Analysen. Diese haben den Anspruch, alle verfügbaren Forschungsarbeiten in einer einheitlichen und nachvollziehbaren Weise auszuwerten. Damit sind sie einerseits rigide, andererseits weniger subjektiv. Leider liegt für die Analyse der Wirksamkeit von Maßnahmen finanzieller Bildung bisher nur eine umfangreiche Meta-Analyse vor.[10]

Diese Arbeit wirkte wie ein Paukenschlag in der einschlägigen Literatur. Während bis dahin die Vorstellung vorherrschte, Bildungsinterventionen würden mehr oder weniger stark wirken, stellten die Autoren in den Raum, es gäbe keine nachweisbare Wirksamkeit. Im Durchschnitt aller von ihnen betrachteten Studien kommen sie auf eine Effektstärke von etwa 0,07. Effektstärken unterhalb von 0,2 werden als klein bezeichnet (und oberhalb von 0,8 als groß). Aber noch wesentlicher ist die Feststellung, dass selbst dieser geringe Wert noch zu hoch ist, wenn man methodisch sauber misst. Die Effektstärke von allen damals erfassten RCTs, als bester Messung, liegt nur noch bei 0,02 und ist damit klar insignifikant und ökonomisch bedeutungslos.

Eine Gruppe von AutorInnen der Weltbank hat daraufhin mit einer eigenen Meta-Analyse reagiert.[11] Auch sie weisen neben Erfolgen auf Probleme hin und darauf, dass die erwünschten Wirkungen, beispielsweise ein besse-

9 Vgl. Tom D. Stanley (2001): Wheat from chaff: Meta-Analysis as quantitative literature review. Journal of Economic Perspectives 15(3), 131–150.

10 Daniel Fernandes, John G. Lynch Jr. und Richard G. Netemeyer (2014), a. a. O.

11 Margret Miller et al. (2015): Can you help someone become financially capable? A meta-analysis of the literature. World Bank Research Observer, 30(2), 220–246.

res Verständnis finanzieller Produkte oder ein rationaleres Spar- oder Anlageverhalten, nicht immer eintreten. Gerade beim Kreditverhalten scheinen Bildungsmaßnahmen oft wirkungslos zu bleiben: Menschen nehmen immer noch zu viele Schulden auf und können Verbindlichkeiten nicht bedienen. Der Nachteil dieser Studie ist ihre geringe Abdeckung: Es wurden nur 19 Primärstudien berücksichtigt, so dass vertiefende Aussagen kaum generalisiert werden können.

Maßnahmen wirksam sowohl auf Finanzwissen als auch auf Finanzverhalten

Neue, eigene Untersuchungen kommen zu zwei wesentlichen Ergebnissen: Maßnahmen finanzieller Bildung sind meist wirksam, auch bei Verwendung rigoroser Evaluierungsmethoden, und die Streubreite der Wirkungen ist hoch.

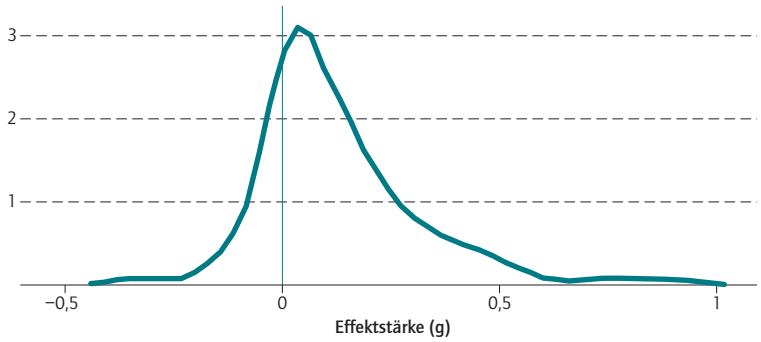
Interventionen der finanziellen Bildung beeinflussen sowohl das Wissen als auch das Verhalten positiv. Dieser Befund gilt auch dann, wenn höchste methodische Ansprüche an die Evaluierungen gestellt werden, das heißt, das Sample auf randomisierte Experimente beschränkt wird. Der mittlere Effekt von Interventionen auf das finanzielle Wissen im Sample aller 67 berücksichtigten Studien der Gruppe 2 und 3 liegt je nach Schätzmethode in der Meta-Analyse zwischen 0,135 und 0,297 (Tabelle 2). Betrachtet man nur die Ergebnisse von Studien der Gruppe 3 (RCTs) liegt der ermittelte Effekt zwischen 0,155 und 0,234. Damit wird deutlich, dass Maßnahmen zur Stärkung des finanziellen Wissens ähnlich wirksam sind wie Bildungsinterventionen im Bereich der Naturwissenschaften an Hochschulen.[12]

Die Effektstärke auf tatsächliches Finanzverhalten ist zwar deutlich geringer (zwischen 0,002 und 0,086 für alle Studien und zwischen 0,051 und 0,082 für RCTs), jedoch existiert eine signifikant positive, messbare Evidenz für einen kausalen Effekt der Bildungsmaßnahmen in sämtlichen durchgeführten Meta-Analysen. Zudem sind diese Effektstärken durchaus vergleichbar mit Interventionen aus anderen Domänen wie den Gesundheitswissenschaften, in denen es darum geht, Menschen zu einer Änderung ihres Verhaltens anzuregen.[13]

Abbildung 2

Verteilung (Dichteschätzung) der kodierten Effektstärken bei den untersuchten Studien zur Wirkung finanzieller Bildung

Für g<1



Quelle: Eigene Berechnungen.

Die empirischen Effektstärken in der Literatur zur finanziellen Bildung sind äußerst heterogen.

Ergebnisse sind sehr heterogen

Es muss allerdings konstatiert werden, dass eine recht breite Streuung in den geschätzten Effektstärken besteht. Dies wird an die Verteilung aller kodierten (ungewichteten) Effektstärken für das eingeführte Sample von 126 Studien deutlich (Abbildung 2). Neben den oben diskutierten Erklärungsfaktoren des Studiendesigns gibt es weitere Determinanten, auf die hier nicht eingegangen wird, und einen erheblichen unerklärten „Rest".[14] In jedem Fall verdeutlicht die große Spannbreite an ermittelten Wirkungen, dass man grundsätzlich von erfolgreichen Maßnahmen lernen kann; aber damit dies systematisch möglich ist, bedarf es einer hinreichend großen Zahl sorgfältig evaluierter Bildungsmaßnahmen.

Fazit: Rigorose Evaluierung notwendig

Finanzielle Bildung hat in den letzten Jahren erhöhte Aufmerksamkeit bekommen, sowohl von privaten Trägern als auch von Schulbehörden und der Wirtschaftspolitik. Es wird in diesem Bereich viel investiert, aber gründliche Evaluierungen der getroffenen Maßnahmen sind rar. Dabei kann man nur aus methodisch überzeugenden Studien verlässlich lernen. Einfache Methoden wie Vorher-Nachher-Befragungen haben zwar den Vorteil, schnell und preiswert zu sein, weisen aber zwei

12 Vgl. Scott Freeman et al. (2014): Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences, 111(23), 8410–8415.

13 David B. Portnoy et al. (2008): Computer-delivered interventions for health promotion and behavioral risk reduction: A meta-analysis of 75 randomized controlled trials, 1988–2007. Preventive Medicine, 47(1), 3–16.

14 Vgl. Tim Kaiser und Lukas Menkhoff (2017), a. a. O.

große Nachteile auf: Sie überschätzen die Wirksamkeit der Bildungsmaßnahmen systematisch und sind wenig präzise, so dass man kaum etwas über die Erfolgsbedingungen von Bildungsinitiativen lernen kann.

Will man wissen, welche Art finanzieller Bildung in welcher Form nachhaltig wirkt, sollte man Gelegenheiten zum Beispiel bei der Einführung neuer Curricula nutzen, um systematisch und verlässlich daraus für spätere Maßnahmen zu lernen.

Bildungsmaßnahmen wirken sich stärker auf das Finanzwissen als auf das Finanzverhalten aus. Fachgerechte Evaluierung würde es möglicherweise erlauben, diese Diskrepanz zu adressieren und gezielter auf die Verbesserung des Finanzverhaltens hinzuwirken.

Tim Kaiser ist wissenschaftlicher Mitarbeiter der Abteilung Weltwirtschaft am DIW Berlin | tkaiser@diw.de

Lukas Menkhoff ist Leiter der Abteilung Weltwirtschaft am DIW Berlin | lmenkhoff@diw.de

JEL: D 14, I 21

Keywords: Financial education, financial literacy, financial behavior, meta-analysis, meta-regression, impact evaluation

DIW Berlin — Deutsches Institut
für Wirtschaftsforschung e. V.
Mohrenstraße 58, 10117 Berlin
T +49 30 897 89 -0
F +49 30 897 89 -200
84. Jahrgang

Vertrieb
DIW Berlin Leserservice
Postfach 74
77649 Offenburg
leserservice@diw.de
Tel. (01806) 14 00 50 25
20 Cent pro Anruf
ISSN 0012-1304
ISSN 1860-8787 (Online)

Gestaltung
Edenspiekermann

Satz
eScriptum GmbH & Co KG, Berlin

Druck
USE gGmbH, Berlin