PNAS

RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

OPEN ACCESS

A national megastudy shows that email nudges to elementary school teachers boost student math achievement, particularly when personalized

Angela L. Duckworth[a,b,1], Ahra Ko[c], Katherine L. Milkman[b], Joseph S. Kay[c], Eugen Dimant[d], Dena M. Gromet[c], Aden Halpern[c,e], Youngwoo Jung[c], Madeline K. Paxson[c], Ramon A. Silvera Zumaran[c], Ron Berman[f], Ilana Brody[g], Colin F. Camerer[h], Elizabeth A. Canning[i], Hengchen Dai[g], Marcos Gallo[h], Hal E. Hershfield[g], Matthew D. Hilchey[j], Ariel Kalil[k], Kathryn M. Kroeper[l], Amy Lyon[m], Benjamin S. Manning[n], Nina Mazar[o], Michelle Michelini[k], Susan E. Mayer[k], Mary C. Murphy[p], Philip Oreopoulos[q], Sharon E. Parker[r], Renante Rondina[j], Dilip Soman[j], and Christophe Van den Bulte[f]

Affiliations are included on p. 8.

Edited by Timothy Wilson, University of Virginia, Charlottesville, VA; received September 11, 2024; accepted February 12, 2025

In response to the alarming recent decline in US math achievement, we conducted a national megastudy in which 140,461 elementary school teachers who collectively taught 2,992,027 students were randomly assigned to receive a variety of behaviorally informed email nudges aimed at improving students' progress in math. Specifically, we partnered with the nonprofit educational platform Zearn Math to compare the impact of 15 different interventions with a reminder-only megastudy control condition. All 16 conditions entailed weekly emails delivered to teachers over 4-wk in the fall of 2021. The best-performing intervention, which encouraged teachers to log into Zearn Math for an updated report on how their students were doing that week, produced a 5.06% increase in students' math progress (3.30% after accounting for the winner's curse). In exploratory analyses, teachers who received any behaviorally informed email nudge (vs. a reminder-only megastudy control) saw their students' math progress boosted by an average of 1.89% during the 4-wk intervention period; emails referencing personalized data (i.e., classroom-specific statistics) outperformed emails that did not by 2.26%. While small in size, these intervention effects were consistent across school socioeconomic status and school type (public, private, etc.) and, further, persisted in the 8-wk post-intervention period. Collectively, these findings underscore both how difficult it is to change behavior and the need for large-scale, rigorous, empirical research of the sort undertaken in this megastudy.

megastudy | math achievement | nudging | field experiment

American students continue to fall behind in math. Compared to students in other developed countries, Americans have ranked in the bottom 25% of students globally on standardized tests of mathematics for decades (1). During the COVID-19 pandemic, math scores for students in elementary and middle school experienced their most dramatic drop since the National Assessment of Educational Progress began annually evaluating student performance in representative national samples in 1990 (2). By the spring of 2022, the average US public school student in grades 3 to 8 had lost the equivalent of a half-year of learning in math (3).

Learning math depends in part on student motivation, and behaviorally informed interventions targeting students' motivation have previously proven effective at boosting math performance. For instance, asking students to give motivational advice improves their math grades the following marking period (4). Likewise, learning that intelligence can grow improves math grades, especially for underperforming students (5). However, the pace at which students' progress in math also depends on contextual factors not directly under their control—such as how their classroom teachers allocate instructional time and what teachers say and do in the classroom (6). Thus, teachers are a logical target for behaviorally informed interventions (7). Unfortunately, randomly assigning teachers to interventions, as opposed to randomly assigning students, massively increases the scale required for a fully powered experiment. The cost of mounting fully powered, large-scale intervention studies targeting teacher behavior is substantial and has limited past research of this type.

In this article, we present the results of a megastudy aimed at improving math achievement. A megastudy is a large field experiment that simultaneously tests a variety of interventions developed by different scientific teams targeting theoretically distinct mechanisms, providing an apples-to-apples comparison of efficacy by doing so in the same sample and with the same preregistered outcome (8, 9). The megastudy paradigm has

Significance

American students continue to fall behind in math. Aiming to improve math achievement among nearly 3 million elementary students, we conducted a megastudy with more than 140,000 teachers in partnership with Zearn Math, a nonprofit educational platform. The most effective of 15 email interventions prompted teachers to log into the platform's dashboard weekly to check their students' progress. This intervention increased students' math progress by 5.06% during the 4-wk intervention (3.30% after adjusting for the winner's curse). Emails that referenced data specific to a teacher's students (vs. those without) boosted progress by 2.26%, with effects lasting 8-wk post-intervention. These findings underscore both the difficulty of changing behavior and the need for large-scale, rigorous, empirical research.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: duckwort@wharton.upenn.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2418616122/-/DCSupplemental.

Published March 24, 2025.

Table 1. Description of megastudy conditions

Condition	Description
1. Nudging Weekly Logins	Teachers are prompted to log in to check their students' progress weekly (sent on Wednesdays)
2. Comparing This Week to Last Week	Teachers receive a comparison of their students' progress this week vs. last week
3. Nudging Friday Logins	Teachers are prompted to log in on Fridays to check their students' progress weekly (sent on Fridays)
4. Empathy	Teachers are prompted to view the platform from their students' perspective and to empathize with students
5. Comparing Own Students to Other Students	Teachers receive a comparison of their students' progress this week vs. progress of other students in the same grade
6. Weekly Classroom Dashboard	Teachers receive a weekly report on their students' progress that week (without Zearn Math Giveaway mentioned)
7. Performance Goals	Teachers are given tips for encouraging students to reach performance goals
8. Digital Swag w/ Celebrity Endorsement	Teachers are offered classroom posters and other materials with selfies of high-profile celebrities
9. Math Teaching Tips	Teachers are given tips on how to most effectively teach math
10. Planning Prompt w/ Printout	Teachers are prompted to plan their Zearn Math lessons for the week ahead with a printable daily lesson scheduler
11. Weekly Classroom Dashboard w/ Giveaway Mentioned	Teachers receive a weekly report on their students' progress that week (with Zearn Math Giveaway mentioned)
12. Learning Goals	Teachers are given tips for encouraging students to reach learning goals
13. Alerts about Students Who Are Struggling	Teachers are prompted to log in and check which students are struggling with a lesson
14. Digital Swag w/o Celebrity Endorsement	Teachers are offered classroom posters and other materials (without selfies of high-profile celebrities)
15. Planning Prompt w/o Printout	Teachers are prompted to plan their Zearn Math lessons for the week ahead (without a printable daily lesson scheduler)
Reminder-Only Megastudy Control	Teachers are reminded that reading Zearn Math emails earned them raffle tickets

Note. See *SI Appendix*, Table S1 for the theoretical rationale for each condition.

previously been applied to noneducational outcomes, including vaccination (10–12), physical exercise (13), and democratic attitudes (14).

In this megastudy, 156 behavioral scientists were invited to attend an informational webinar and then to submit ideas for nudges (i.e., interventions aimed at changing behavior without mandates, bans, or substantial financial incentives) to be delivered via four weekly emails to elementary school teachers. Invited scientists were members of Behavior Change For Good (BCFG), an interdisciplinary network of scholars who specialize in positive behavior change. Proposals were screened for feasibility and, subsequently, fully developed in collaboration with BCFG and Zearn Math staff. As shown in Table 1 and elaborated in *SI Appendix*, Table S1, this process led to a variety of theoretically distinct, behaviorally informed interventions. One intervention, for example, employed planning prompts, encouraging teachers to plan their Zearn Math lessons for the week ahead using a printable calendar; another appealed to teachers' identities as empathic, encouraging them to take their students' perspectives when accessing the Zearn Math portal. See *SI Appendix*, Table S1 for the theoretical rationale of each treatment.

In fall 2021, more than 140,000 elementary school teachers instructing nearly 3 million students were randomly assigned to receive one of 15 different sets of behaviorally informed emails (or a reminder-only megastudy control email) over 4-wk. The primary outcome was the average number of lessons these teachers' students completed on the Zearn Math platform during this intervention period. Zearn Math is a nonprofit, online math instruction platform used by roughly 25% of US elementary school students (15). Teachers using Zearn Math assign students to complete lessons on the platform during regular school hours. Upon logging in, students engage in guided lesson activities designed to enhance their understanding of mathematical concepts and procedures. All Zearn Math lessons include fluency games, videos led by on-screen teachers with pause points to solve math problems, and a closing mastery-based quiz on which students must score 100% in order to advance to the next lesson. Quasi-experimental evidence shows that programmatic use of Zearn Math predicts increases in state-wide standardized math achievement test scores (16).

Methods

This research study was granted exempt status by The University of Pennsylvania's Institutional Review Board (IRB). No data Zearn Math provided for this research included personally identifying information about teachers or students.*

A total of 19 behavioral scientists and 2 educators designed 15 different sets of weekly intervention emails to be sent to Zearn Math teachers once per week during our megastudy's 4-wk intervention period (shown in Table 1). Each week, email content was repeated in the "What's New" section of the Zearn Math teacher dashboard, a feature of the Zearn Math platform that teachers can log into for announcements and to view personalized data specific to their students. Note that by design, students can access Zearn Math without their teachers logging into this dashboard. See *SI Appendix* for details on the content and theoretical rationale of the 15 intervention conditions and the reminder-only megastudy control condition.

To incentivize teachers to read email messages, all teachers on the Zearn Math platform received two study-wide messages on September 1 and 8, 2021,

*Note that in accordance with Zearn Math's Privacy Policy and partner contracts, we did not receive any students' personal information (such as individual students' Zearn Math progress), nor did we receive identifiable teacher information. To ensure the confidentiality of students' personal information, we only received aggregate, classroom-level information.

informing them that they had been enrolled in the "Zearn Math Giveaway" and that for every email from Zearn Math they opened until October 12, 2021, they would earn raffle tickets. The more raffle tickets teachers earned, the higher their chances of winning prizes, including autographed children's books, stickers, and gift cards (see *SI Appendix* for details).

Teachers were randomized into one of the 15 intervention conditions or the megastudy control condition, which simply reminded teachers to "keep Zearning" and that reading Zearn Math emails earned them raffle tickets. As shown in Fig. 1, all email messages were sent weekly, starting Wednesday, September 15, 2021, and concluding on Tuesday, October 12, 2021. Zearn Math provided daily classroom-level data on students' average Zearn Math lessons completed during the 4-wk intervention (i.e., weeks 1 to 4) and 8-wk post-intervention periods (i.e., weeks 5 to 12).

Megastudy Analyses. Following our megastudy's preregistered analysis plan (https://osf.io/dgpkn), we restricted analyses to Zearn Math teachers who were assigned to one of our megastudy's 16 conditions and who taught in at least one classroom with at least one student during the 4-wk intervention period. As preregistered, we excluded teachers who: a) did not receive any emails because they had inactive accounts, invalid email addresses, or had opted out of receiving messages, b) neither logged in to the Zearn Math platform nor had an associated student who logged in to the platform between March 1, 2021, and September 14, 2021, c) had more than 150 students associated with their Zearn Math account as of October 18, 2021,[†] or d) had more than 6 classes associated with their Zearn Math account as of October 18, 2021. We further excluded Zearn Math classrooms that e) were associated with more than one Zearn Math teacher and f) had grade levels corresponding to high school or post-high school. The number of teachers excluded following these preregistered criteria did not differ by condition; see *SI Appendix* for details.

A total of N = 140,461 teachers instructing 2,992,027 students in 161,719 classrooms across over 22,000 schools were assigned to one of 15 intervention conditions (n_{min} = 7,238, n_{mean} = 7,397, n_{max} = 7,578) or the reminder-only megastudy control condition (n = 29,513), which included more teachers than the intervention conditions in order to increase statistical power. Unexpectedly, for n = 16,372, or 11.66% of teachers, at least one of two errors arose in the emails sent by Zearn Math during the intervention period: a) none of the intervention emails we intended to send them were sent (n = 13,568, or 9.66% of teachers), or b) an intervention email we did not intend to send was sent (i.e., its content was from an intervention condition other than the one to which the teacher was assigned; n = 2,804, or 2.00% of teachers). Because whether a teacher experienced an email error was systematically related to condition assignment [$\chi^2(15) = 33.01$, $P = 0.005$], we followed an intent-to-treat approach in our main analyses to avoid endogeneity. See *SI Appendix* for details on the prevalence of email problems by condition, as well as robustness analyses excluding and adjusting for teachers who were subject to these email errors (*SI Appendix*, Tables S4 and S5); in general, supplementary analyses yielded similar point estimates of intervention effects to those in our main analyses, but standard errors were larger.

Following our preregistered analysis plan, we fit a weighted ordinary least squares (OLS) regression model to assess the effect of condition assignment on the primary outcome of interest: average Zearn Math lessons completed by students in a given teacher's classroom(s) during the 4-wk intervention period. In this regression model, observations were weighted proportionally to the total number of students in teachers' classroom(s). The primary predictors were indicators for each of our megastudy's 15 intervention conditions (with the reminder-only megastudy control condition omitted as the comparison group). The regression model also included the following preregistered control variables, which did not differ by condition: a) school fixed effects,[‡] b) an indicator for whether the teacher's account was free (vs. paid) as of October 18, 2021,[§] c) the number of times the teacher had logged in to Zearn Math in the month-and-a-half prior to the megastudy (from August 1 to September 14, 2021, inclusive), d) the total number of students in the teacher's classroom(s) as of October 18, 2021, e) the number of classrooms associated with the teacher as of October 18, 2021, f) the number of days since the teacher obtained a Zearn Math account prior to the megastudy's launch, g) the number of days separating the megastudy's launch and the start of the teacher's school year, h) the total average lessons completed by a teacher's students preintervention,[¶] i) an indicator for whether the teacher opened our September 1, 2021, email announcing the upcoming Zearn Math Giveaway, j) an indicator for whether the teacher opened our September 8, 2021, email reminding them of the upcoming Zearn Math Giveaway, and k) separate controls for the percentage of a teacher's students in each grade, respectively (omitting the largest category, third grade, to avoid multicollinearity).[#] We calculated P values with and without adjustment for multiple comparisons using the Benjamini–Hochberg (BH) procedure (17).

As preregistered, we also examined the durability of intervention effects after the conclusion of the 4-wk intervention period (i.e., weeks 1 to 4). To do this, we reran our primary weighted OLS regression models to predict the average number of math lessons students completed in a given teacher's classrooms during weeks 5 to 8 and during weeks 9 to 12 (see Fig. 1 for an intervention timeline).

As preregistered, to examine heterogeneity in intervention effects across subgroups, we also reran our primary weighted OLS regression models by subgroup (e.g., fourth grade vs. fifth grade) and, separately, reran our primary models with added interaction terms between intervention condition indicators and indicators for subgroups.

Finally, we examined two preregistered secondary dependent variables using the same primary OLS regressions described above. These secondary dependent variables were a) student engagement, measured as the average minutes teachers' students spent on the Zearn Math platform during the 4-wk intervention and b) teacher engagement, measured as the number of times teachers logged in to their Zearn Math dashboards during the 4-wk intervention period.

Attribute and Forecasting Analyses. Email content from each condition varied on several underlying attributes. To characterize the subjective attributes of these conditions, we recruited 1,752 raters from Prolific. Each rater was randomly assigned to evaluate one of the 15 intervention conditions or the reminder-only megastudy control condition. Based on four weekly emails, raters evaluated their assigned condition on 8 different dimensions (using a rating scale ranging from 1 = strongly disagree to 5 = strongly agree): usefulness, expectedness, positive mood, emotional connection, surprisingness, casual tone, forwardability, and specificity (i.e., the degree to which "these emails are specific to a teacher and their classroom"; see *SI Appendix*, Table S8 for details). One-way random-effect intraclass correlation coefficients for all dimensions were higher than 0.80 [ICC(1, k) = 0.81]—indicative of good reliability (see *SI Appendix*, Table S9 for details).

To identify objective attributes, two trained raters coded conditions on 13 different features: a) word count, b) the number of exclamation marks, c) the number of questions, d) Flesch-Kincaid reading level, e) whether emails were sent on Wednesdays, f) whether emails were sent on Sundays, g) whether emails mentioned the Zearn Math Giveaway, h) whether any emails included links to printable material, i) whether any emails included graphics/icons, j) the number of weeks that emails included teacher testimonials, k) the number of weeks that emails mentioned science or research, l) the number of weeks that emails mentioned a "tip," and m) whether any emails referenced personalized data that were specific to a teacher's students either directly in the email body or indirectly by linking to personalized resources. Our trained raters achieved 100% agreement on these objective attributes for each condition (see *SI Appendix*, Table S7 for details).

Finally, to assess the ex-ante predictability of this megastudy's results, we also collected forecasts of each intervention's impact on the primary outcome (the

[†] Note that in some schools, class rosters and teacher assignments are automatically generated through a database integration with Zearn Math. As a result, Zearn Math's "teacher" database may include some individuals who are actually administrators. To ensure the accuracy of our analyses, we exclude any accounts associated with more than 150 students or more than 6 classes, as these are likely to belong to administrators rather than classroom teachers.

[‡] Each school with more than one teacher and a Zearn Math school ID has a unique fixed effect. Schools with only one teacher were assigned a common fixed effect, homeschools were assigned a common fixed effect, international schools without a Zearn Math school ID were assigned a common fixed effect, and unknown/other schools without a Zearn Math school ID were assigned a common fixed effect.

[§] Note that teachers with a "free account" were teachers who independently signed up for the Zearn Math platform, whereas teachers with a "paid account" were teachers whose schools or districts had purchased a license for the platform.

[¶] The preintervention period began on the first day of the teacher's school year and ended on September 14, 2021. If the teacher's school year's start date was unknown, we a) substituted the earliest start date among all teachers in the study, which was July 14, 2021, and b) created a separate indicator variable noting that their start date was missing.

[#] If a single classroom was associated with multiple grades between 1 and 8, we a) set the classroom grade to 3—the modal grade—and b) created a separate indicator variable noting that at least one of the teacher's classrooms was associated with multiple grades.

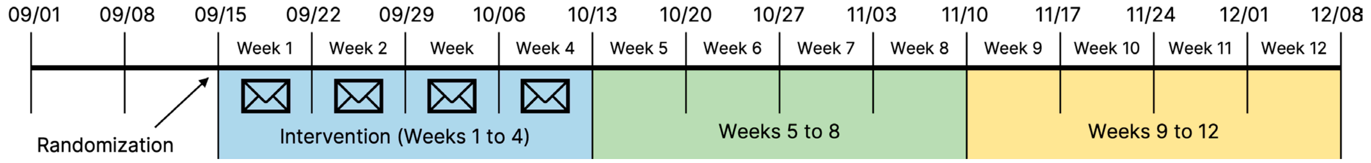


Fig. 1. Timeline of the megastudy.

average number of Zearn Math lessons all of a teacher's students completed per week during the 4-wk intervention period). We collected these forecasts from a) 15 behavioral scientists involved in designing the interventions (who had not yet been made aware of the megastudy's results in Table 2), b) 29 Zearn Math staff members (preregistration: https://osf.io/2efnw), and c) 100 US elementary school teachers recruited through Qualtrics (preregistration: https://osf.io/3mqfh). See *SI Appendix* for details.

In one megastudy intervention condition (Alerts about Students Who Are Struggling), teachers received a different email message depending on whether any of their students were stalled in their progress on the Zearn Math platform in the week prior. Because stalled progress is potentially endogenous to condition assignment (i.e., the type of message received depends on whether stalled progress existed in any particular classroom), we excluded this intervention arm from attribute and forecasting analyses presented in the main text. As a robustness check, however, we ran the same analyses including these observations (weighting attributes and forecasts by the number of messages sent out for a given version of the email intervention in the Alerts about Students Who Are Struggling condition; see *SI Appendix* for details), and we confirmed that our results were not meaningfully changed.

Results

Teachers in our megastudy taught an average of 21.30 students (SD = 15.31) across an average of 1.15 classrooms (SD = 0.59). During the six weeks prior to our intervention (from August 1 to September 14, 2021), teachers logged in to the Zearn Math dashboard a total of 3.62 times, on average (SD = 8.49). On average, teachers opened 25.08% of emails (SD = 33.87%), and open rates did not differ significantly by condition ($P = 0.16$).[‖] For complete summary statistics describing our study sample, see *SI Appendix*. As shown in *SI Appendix*, Table S2, megastudy control variables did not differ by condition (P-values from all F tests > 0.05).

Comparison of Interventions. In a preliminary, non-preregistered analysis reported in Model 1 in Table 2, megastudy intervention messages led students to complete an average of 0.03 more lessons than students in the megastudy control condition, a 1.89% increase (Cohen's $d = 0.01$, $P = 0.02$). Specifically, students in the reminder-only megastudy control condition completed a regression-estimated average of 1.78 lessons during the 4-wk intervention period,[**] whereas students in the 15 intervention conditions completed a regression-estimated average of 1.81 lessons during this period. Preregistered F-tests allow us to reject the null hypothesis that the effects of all 15 interventions are equal to zero [$\chi^2(15) = 27.09$, $P = 0.03$], and to reject the null hypothesis that 15 effects have the same true value at marginal significance [$\chi^2(14) = 21.89$, $P = 0.08$].

As shown in Fig. 2 and Model 2 in Table 2, four interventions produced significant benefits before adjusting for multiple hypothesis testing. However, only one intervention remained significant after applying a BH-correction for testing multiple hypotheses: Nudging Weekly Logins. This intervention encouraged teachers to log into their Zearn Math dashboard weekly to get an updated report on how their students were performing (see Fig. 3). While reliable, the effect of this intervention was nevertheless small in size, generating only an estimated 0.09 extra completed lessons in four weeks, or a 5.06% increase in lessons completed over the reminder-only megastudy control ($d = 0.02$, unadjusted $P = 0.002$, BH-adjusted $P = 0.03$). When we apply the James–Stein shrinkage procedure to adjust for the winner's curse (i.e., the maximum of the 15 intervention effects estimated is upwardly biased), we still estimate that this intervention produced 0.06 extra lessons completed, or a 3.30% increase in lessons completed over the reminder-only megastudy control condition (18). However, we cannot rule out with 95% confidence that a different intervention with a positive regression-estimated effect (from Interventions 2 to 12) was the true top performer (see *SI Appendix*, Table S25 for details).

Attribute Analyses. Because our megastudy conditions differed on multiple attributes, in preregistered exploratory analyses, we examined sets of correlations between a) 15 of our 16 regression-estimated megastudy condition effects (see Model 2 in Table 1, and note that the omitted control condition's estimated effect is 0.00)[††] and b) each of the 21 message attributes coded by at least two independent raters.

The only objectively coded attribute that showed a statistically significant association with a condition's regression-estimated effect was whether emails referenced personalized data specific to a teacher's students, either directly in the email body (e.g., "Last week, your students completed 2.1 lessons and spent 15.3 min on Zearn Math. Let's raise the bar this week!") or indirectly by hyperlinking to the Zearn Math teacher dashboard (e.g., "Check the Pace Report for your students today!") [$r(13) = 0.70$, $P = 0.004$]. As might be expected, Prolific raters evaluated interventions that referenced personalized data as more "specific to a teacher and their classroom" ($M = 3.81$, SD = 0.56) than those that did not [$M = 2.62$, SD = 0.22, $t(13) = 4.92$, $d = 2.59$, $P < 0.001$]. Notably, this specificity rating was the only subjectively coded attribute that showed a significant association with regression-estimated intervention effects [$r(13) = 0.55$, $P = 0.03$; Fig. 4].[‡‡]

In an exploratory analysis that was not preregistered, interventions that referenced personalized data led students to complete 0.04 more lessons (2.26% increase; $d = 0.02$, $P = 0.001$) during the 4-wk intervention period than interventions that did not. See Model 3 in Table 2.

[‖] Note that email open rates may not be accurate because some email platforms do not provide reliable data on whether the recipient actually opened the email. Additionally, some platforms have a preview function that allows recipients to read the email without opening it.

[**] All analyses reflected student-weighted estimates (i.e., weighted by the number of students per teacher), incorporated fixed effects for schools, and controlled for mean-centered covariates using the mean of the megastudy control.

[††] As noted in Methods, we excluded the intervention Alerts about Students Who Are Struggling in order to avoid endogeneity problems because the receipt of one version of this message would signal to a rater that a student was performing poorly.

[‡‡] In robustness checks, we confirmed that these analyses yielded similar results whether emails included personalized data directly or, instead, included a hyperlink to personalized data on the Zearn Math dashboard.

Table 2. Regression-estimated impact of 15 interventions on the average number of Zearn Math lessons students completed in a given Zearn Math teacher's classroom(s) during the 4-wk intervention period, either pooling all interventions (Model 1), breaking out all interventions individually (Model 2), or pooling the interventions that referenced personalized data specific to a teacher's students (Model 3)

	Model 1		Model 2			Model 3	
	B	*P-value*	*B*	*P-value*	*Adjusted P-value*	*B*	*P-value*
Assigned Any Intervention	0.034* (0.015)	0.023					
1. Nudging Weekly Logins[1]			0.090** (0.029)	0.002	0.029		
2. Comparing This Week to Last Week[1]			0.069* (0.030)	0.021	0.107		
3. Nudging Friday Logins[1]			0.067* (0.029)	0.021	0.107		
4. Empathy[1]			0.063* (0.029)	0.033	0.125		
5. Comparing Own Students to Other Students[1]			0.055 (0.030)	0.064	0.177		
6. Weekly Classroom Dashboard[1]			0.053 (0.029)	0.071	0.177		
7. Performance Goals[1]			0.047 (0.030)	0.111	0.239		
8. Digital Swag w/ Celebrity Endorsement[2]			0.026 (0.029)	0.379	0.583		
9. Math Teaching Tips[2]			0.026 (0.029)	0.379	0.583		
10. Planning Prompt w/ Printout[2]			0.025 (0.029)	0.389	0.583		
11. Weekly Classroom Dashboard w/ Giveaway Mentioned[1]			0.021 (0.029)	0.477	0.650		
12. Learning Goals[1]			0.004 (0.029)	0.897	0.897		
13. Alerts about Students Who Are Struggling[1]			−0.011 (0.029)	0.709	0.760		
14. Digital Swag w/o Celebrity Endorsement[2]			−0.012 (0.029)	0.679	0.760		
15. Planning Prompt w/o Printout[2]			−0.015 (0.029)	0.620	0.760		
Assigned an Intervention Referencing Personalized Data[1]						0.040*** (0.012)	0.001
R-squared	0.685		0.686			0.685	
School fixed effects	14,942		14,942			14,942	
Comparison condition(s)	Control condition		Control condition			Conditions without personalized data	
Observations	140,461		140,461			140,461	

Note. This table reports the results of three OLS regressions predicting the average number of Zearn Math lessons completed by teachers' students on the Zearn Math platform during the 4-wk intervention period, weighted proportionally to the total number of students in their Zearn Math classroom(s). In Model 1, we include a single pooled intervention indicator for whether a teacher was assigned to any of 15 intervention conditions (with the reminder-only megastudy control condition as the comparison group). In Model 2, we include 15 different indicators for each of our megastudy's intervention conditions (with the reminder-only megastudy control condition as the comparison group). In Model 3, we include a single indicator for whether a teacher was assigned to any intervention for which emails referenced personalized data specific to a teacher's students either (directly in the email body or indirectly by linking to personalized resources) with the conditions without personalized data, including the reminder-only megastudy control condition, as the comparison group. The superscript "1" denotes an intervention for which emails referenced personalized data; the superscript "2" denotes an intervention for which emails did not reference personalized data. All regressions include the set of control variables and school fixed effects reported in the manuscript's Methods section. R-squared represents the total R-squared, accounting for the effects of all variables and the fixed effects. Standard errors are reported in parentheses. Statistical tests of whether an individual regression coefficient is zero are all two-sided. Adjusted *P* values are calculated using the Benjamini–Hochberg (BH) method to account for multiple comparisons. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$

Post-intervention Effects. To examine the durability of intervention effects after the conclusion of the 4-wk intervention period (i.e., weeks 1 to 4 post-launch, see Fig. 1), we reran our three primary regression models but replaced our dependent variable with the average number of Zearn Math lessons students completed in a given teacher's classroom(s) during weeks 5 to 8 (Models 1 to 3 in *SI Appendix*, Table S12) and during weeks 9 to 12 (Models 4 to 6 in *SI Appendix*, Table S12). As shown in Models 1 and 4 in *SI Appendix*, Table S12, the intervention messages had a significant impact during weeks 5 to 8 (additional lessons completed = 0.03, 1.81% increase; $d = 0.01$; $P = 0.04$) and during weeks 9 to 12 (additional lessons completed = 0.03, 2.42%, increase; $d = 0.02$; $P = 0.01$). As shown in Models 2 and 5 in *SI Appendix*, Table S12, Nudging Weekly Logins showed a directional BH-adjusted impact

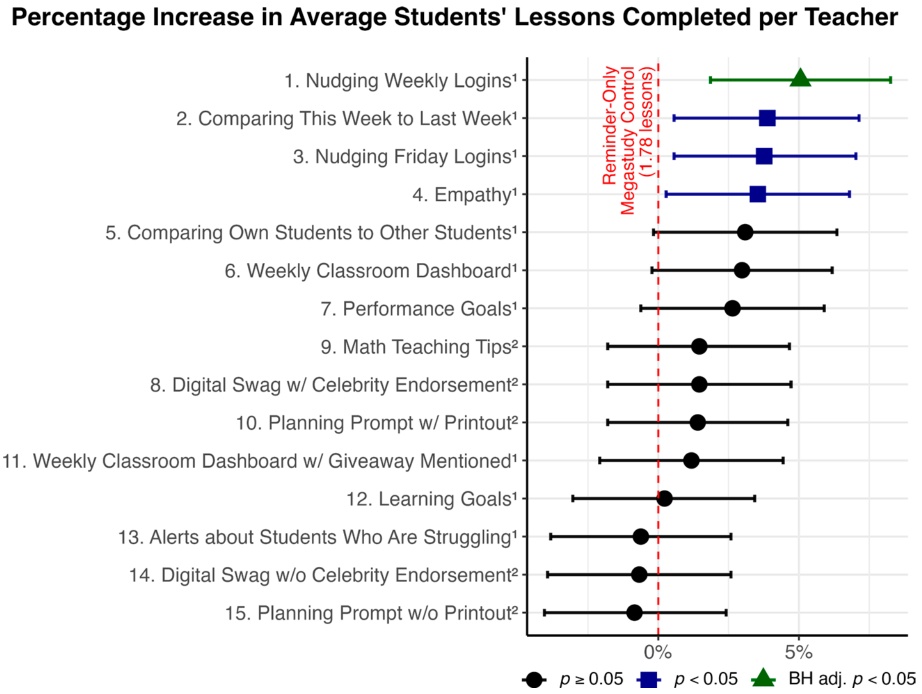
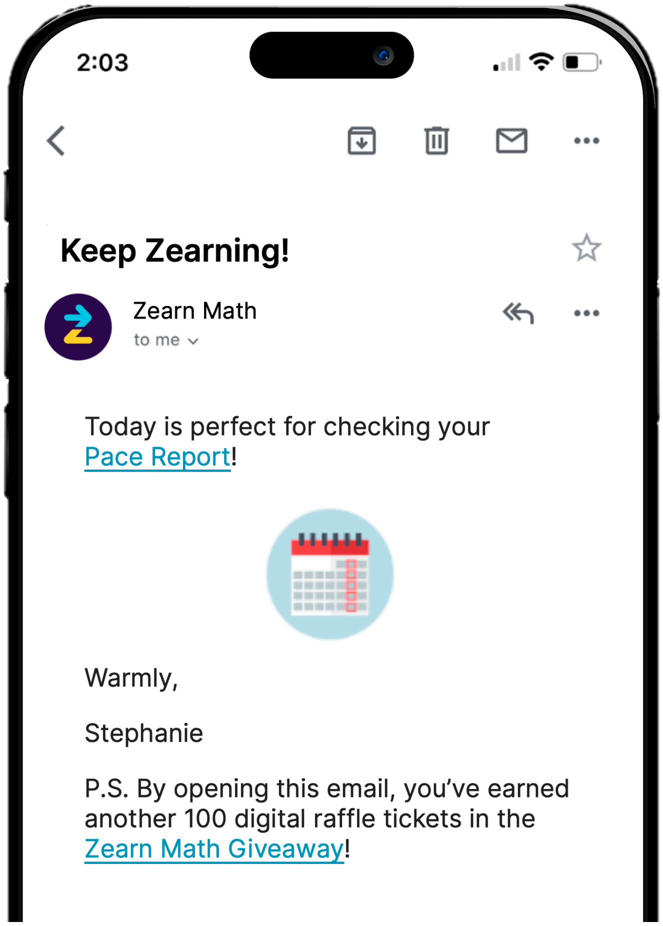


Fig. 2. Regression-estimated increase in the average number of Zearn Math lessons students completed in a given Zearn Math teacher's classroom(s) during the 4-wk intervention period. *Note.* Whiskers depict 95% confidence intervals (CIs) without correction for multiple comparisons. The superscript "1" on a y-axis label denotes an intervention for which emails referenced personalized data specific to a teacher's students; the superscript "2" denotes an intervention for which emails did not reference personalized data.

on student progress during weeks 5 to 8 (additional lessons completed = 0.08, 4.49% increase; d = 0.02; unadjusted P = 0.01, BH-adjusted P = 0.11) and a marginally significant BH-adjusted impact during weeks 9 to 12 (additional lessons completed = 0.07, 5.31% increase; d = 0.02; unadjusted P = 0.004, BH-adjusted P = 0.06). As shown in Models 3 and 6 in *SI Appendix*, Table S12, intervention emails with personalized data improved student progress during weeks 5 to 8 (additional lessons completed = 0.03, 1.60% increase; d = 0.01; P = 0.03) and weeks 9 to 12 (additional lessons completed = 0.02, 1.72% increase; d = 0.01; P = 0.02).

Heterogeneity in Intervention Effects, Analyses of Secondary Dependent Variables, and Robustness Checks. Following our preregistration, we explored how intervention effects differed across nine subgroups (e.g., school type: public, private, Catholic, etc.; proportion of students eligible to receive free or reduced-price meals (median split); classroom grade level). Then, to assess the reliability of observed subgroup differences, we added interaction terms (between subgroup variable(s) and intervention condition indicators) to Model 1, Model 2, and Model 3, respectively. For each of these 27 models, we tested the joint hypothesis that all interaction terms equaled zero. When accounting for multiple comparisons using the Benjamini–Yekutieli correction (19), two of these tests were statistically significant: In Model 3, the effect of personalized (vs. nonpersonalized) interventions was larger for teachers with paid (vs. free) Zearn Math accounts (F = 14.27, unadjusted P < 0.001, BY-adjusted P = 0.02). And in Model 2, the social comparison intervention produced larger gains in student achievement for students whose teachers had paid (vs. free) Zearn Math accounts (F = 2.55, unadjusted P < 0.001, BY-adjusted P = 0.04). While we hesitate to interpret these exploratory results, it is worth noting that schools and districts that purchase Zearn Math accounts for their teachers can be assumed to make this digital platform a more central component of math instruction. Indeed, teachers with paid (vs. free) Zearn Math accounts logged in to their dashboards 155% more frequently and had students who completed 89% more lessons per week and engaged with the Zearn Math platform for 95% more minutes per week (Ps < 0.001, *SI Appendix*, Table S14). See *SI Appendix* for details.



2:03

Keep Zearning!

Zearn Math
to me

Today is perfect for checking your Pace Report!

Warmly,

Stephanie

P.S. By opening this email, you've earned another 100 digital raffle tickets in the Zearn Math Giveaway!

Fig. 3. A sample email message from the Nudging Weekly Logins intervention.

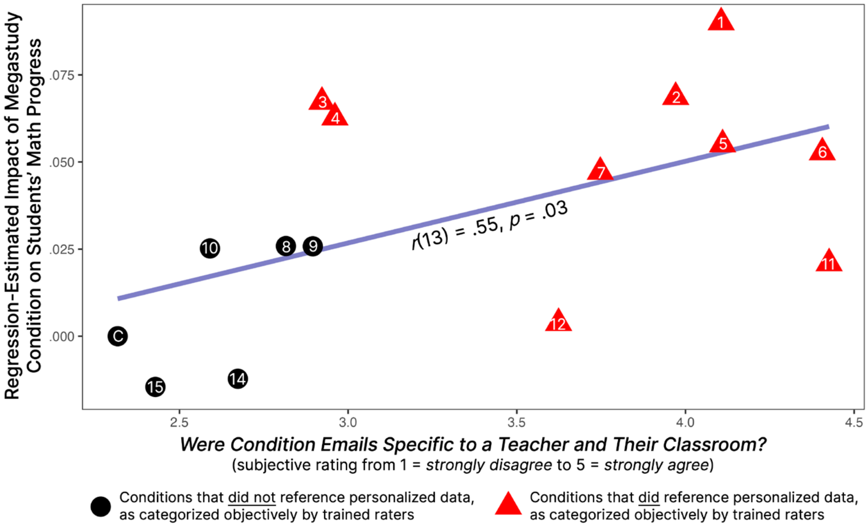


Fig. 4. Scatterplot illustrating the relationship between the degree of specificity to a teacher and their classroom in a condition and that condition's regression-estimated effect on students' math progress. *Note.* This scatterplot displays the relationship between the 15 regression-estimated megastudy condition effects and 1,752 Prolific workers' average ratings of the extent to which intervention emails were specific to a teacher and their classroom. Red triangles indicate the condition was objectively coded by trained raters as referencing personalized data specific to a teacher's students; black circles indicate the condition did not reference such personalized data. 1: Nudging Weekly Logins, 2: Comparing This Week to Last Week, 3: Nudging Friday Logins, 4: Empathy, 5: Comparing Own Students to Other Students, 6: Weekly Classroom Dashboard, 7: Performance Goals, 8: Digital Swag w/Celebrity Endorsement, 9: Math Teaching Tips, 10: Planning Prompt w/Printout, 11: Weekly Classroom Dashboard w/Giveaway Mentioned, 12: Learning Goals, 14: Digital Swag w/o Celebrity Endorsement, 15: Planning Prompt w/o Printout, C: Reminder-Only Megastudy Control.

Following our preregistration, we reran our three primary regression models (from Table 2) focusing on two secondary outcomes that were not the explicit target outcome for the megastudy. For the secondary outcome of student engagement (defined as the average minutes teachers' students spent on the Zearn Math platform during the 4-wk intervention), these models produced results that were similar to those in our primary models predicting student achievement. For the secondary outcome of teacher engagement (defined as the number of times teachers logged in to their Zearn Math dashboards during the 4-wk intervention period), however, none of the three models produced significant results, which may reflect the fact that by design, students can use Zearn Math without their teacher logging in to their dashboard. See *SI Appendix*, Tables S15 to S20.

In robustness checks, we adhered to our preregistration and reran our three primary models using a count data regression framework for student achievement and teacher engagement. Results were largely unchanged. See *SI Appendix*, Tables S21 to S24.

Forecasting of Intervention Efficacy. All forecasts of intervention efficacy collected were overly optimistic. Scientists who collaborated on intervention designs predicted that compared to the reminder-only megastudy control, the 14 interventions they were asked to evaluate would produce an average of 1.07 extra math lessons completed during the intervention period (SD = 1.22, 59.51% increase, median of 0.41 extra lessons), when in reality they produced an average of 0.03 extra lessons completed.[§§] Likewise, Zearn Math staff predicted our average intervention would produce 3.87 additional lessons completed (SD = 2.80, 215.09% increase, median of 2.82 extra lessons). Classroom teachers predicted our average intervention would produce 2.90 additional lessons completed (SD = 1.31, 144.86% increase, median of 2.89 extra lessons).

Whereas forecasts for the average efficacy of interventions were optimistic by a factor of 30 or more, there was remarkable rank-order consistency in how scientists, Zearn Math staff, and classroom teachers predicted individual interventions would perform [scientists and staff: $r(12) = 0.86$, $P < 0.001$; scientists and teachers: $r(12) = 0.71$, $P = 0.005$; staff and teachers: $r(12) = 0.66$, $P = 0.01$]. Given limited degrees of freedom, it is not surprising that these forecasts failed to reliably predict the rank order of intervention efficacy [scientists $r(12) = 0.28$, $P = 0.33$; staff $r(12) = 0.38$, $P = 0.18$; teachers $r(12) = 0.07$, $P = 0.81$]. Nevertheless, behavioral scientists and Zearn Math staff accurately predicted that interventions that referenced personalized data would be more effective [scientists $t(12) = 2.78$, $P = 0.02$; staff $t(12) = 3.97$, $P = 0.002$; teachers $t(12) = 1.36$, $P = 0.20$].

Discussion

In this megastudy, we found that compared to standard email reminders, weekly behaviorally informed email messages to elementary school teachers very slightly improved their students' math progress over the 4-wk intervention period (by 1.89%). The most effective intervention, which increased student math progress by about 5.06% over the 4-wk intervention period (or 3.30% after accounting for the winner's curse), simply prompted teachers to log into Zearn Math weekly for an updated, personalized report on their students' progress. More generally, email reminders referencing personalized data on student progress outperformed email reminders without personalized data, boosting students' math progress during the 4-wk intervention period by about 2.26%. These intervention effects were consistent across contexts—showing comparable benefits regardless of all measured moderators, including the proportion of low-income students in teachers' schools. These small intervention effects remained nearly identical in magnitude and statistical significance in the eight weeks after we sent teachers their last intervention email. Collectively, the behaviorally informed reminders from the megastudy intervention conditions resulted in students completing an

[§§] As noted in Methods, we excluded the intervention Alerts about Students Who Are Struggling in order to avoid endogeneity problems because the receipt of one version of this message would signal to a rater that a student was performing poorly.

estimated 80,424 additional lessons during the 4-wk intervention period and an estimated 156,117 additional lessons during 8-wk follow-up.[¶¶]

Notably, the impact of behaviorally informed reminders in our megastudy was surprisingly small: Students whose teachers received any type of behaviorally informed reminder outperformed students whose teachers received a standard reminder by only $d = 0.01$. This effect was at least 30 times smaller than forecasted by the behavioral scientists who designed interventions for the megastudy, by Zearn Math staff intimately familiar with the platform's operation, or by a sample of US elementary school teachers. Benchmarked against interventions in randomized controlled trials commissioned by the US Department of Education to increase standardized achievement outcomes (20), this effect was slightly below-average in size, falling between the 40th and 50th percentiles of the distribution. Likewise, this improvement (1.89%) falls short of the average (8.0%) effect of nudges carried out by government nudge units across a range of policy domains (21) and in the domain of education in particular (22–24). Further research should explore other, potentially more effective modalities for delivering reminders (e.g., postcards, text messages), other types of nudges (e.g., a default schedule, changes in school-wide social norms), and "wise interventions" that seek to change relevant beliefs. For instance, a 45-min online growth mindset activity for high school math teachers produced a $d = 0.07$ improvement on student course grades (25).

Our findings suggest several additional valuable avenues for future research. First, more random-assignment field experiments are needed to confirm the causal benefits of teacher-targeted nudges. Because such interventions are difficult to implement at scale, we see particular promise for megastudies, which not only enable an apples-to-apples comparison of theoretically distinct approaches but also reduce the marginal cost of conducting field research for individual research teams. Second, future studies should probe the longer-term effects of behaviorally informed interventions. Although we did not observe fadeout effects in the eight weeks following our 4-wk intervention period, many if not most educational interventions diminish in efficacy over time (26, 27), and given a longer follow-up period, we would expect the effects reported here to diminish as well. Further research is also needed to investigate whether short-term gains in performance lead to enduring, long-term gains in learning (28). Finally, in light of a recent study showing that personalized data do not appear to improve the efficacy of proenvironmental interventions (29), additional research is needed to confirm and explain the benefits of referencing personalized data when nudging teachers. If robust, the receptivity of teachers to information about how their own students are performing highlights a motivation that is not directly addressed by traditional policy approaches to teacher accountability. Namely, it may be that capitalizing on teachers' intrinsic motivation to help their students is a distinct and potentially cost-effective approach that can complement other interventions, such as offering performance bonuses and other extrinsic incentives (25).

Data, Materials, and Software Availability. Anonymized csv files of anonymized datasets have been deposited in Open Science Framework (OSF) (https://osf.io/gyhw2/?view_only=0be05c88030e4ab6ae5e6dd78ea5a08e) (30).

ACKNOWLEDGMENTS. Research reported in this article was supported in part by an anonymous donor to Zearn Math. While Zearn Math was involved in data collection from their platform, they had no role in study design, data analysis, the decision to publish, or preparation of the manuscript. Support for this research was also provided in part by the AKO Foundation, J. Alexander, M. J. Leder, W. G. Lichtenstein, and A. Schiffman and J. Schiffman. The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the listed entities. We thank Callie Rosenzweig, Seth Cannon, Jake Rothschild, and Heather Graci for outstanding research assistance.

Author affiliations: [a]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104; [b]Department of Operations, Information and Decisions, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104; [c]Behavior Change for Good Initiative, The Wharton School and the School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19104; [d]Center for Social Norms and Behavioral Dynamics, University of Pennsylvania, Philadelphia, PA 19104; [e]Department of Economics, University of Pittsburgh, Pittsburgh, PA 15260; [f]Department of Marketing, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104; [g]Anderson School of Management, University of California Los Angeles, Los Angeles, CA 90095; [h]Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125; [i]Department of Psychology, Washington State University, Pullman, WA 99164; [j]Behavioural Economics in Action at Rotman, Rotman School of Management, University of Toronto, Toronto, ON M5R 0A3; [k]Harris School of Public Policy, University of Toronto, Toronto, ON 60637; [l]Department of Psychology, Sacred Heart University, Fairfield, CT 06825; [m]School of Business and Social Sciences, Colby-Sawyer College, New London, NH 03257; [n]Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02142; [o]Department of Marketing, Questrom School of Business, Boston University, Boston, MA 02215; [p]Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405; [q]Department of Economics, University of Toronto, Toronto, ON M5S 3G7; and [r]Consultant, Philadelphia, PA 19103

Author contributions: A.L.D., K.L.M., J.S.K., D.M.G., I.B., C.F.C., E.A.C., H.D., M.G., H.E.H., M.D.H., A.K., K.M.K., A.L., B.S.M., N.M., M.M., S.E.M., M.C.M., P.O., S.E.P., R.R., and D.S. designed research; A.L.D., K.L.M., J.S.K., and D.M.G. performed research; R.B. and C.V.d.B. contributed new reagents/analytic tools; A.L.D., A.K., K.L.M., E.D., A.H., Y.J., M.K.P., R.A.S.Z., R.B., and C.V.d.B. analyzed data; and A.L.D., A.K., K.L.M., and E.D. wrote the paper.

The authors declare no competing interest.

[¶¶] The additional lessons completed from the fifteen interventions were calculated by multiplying the pooled intervention effects from Table 2, Model 1 and *SI Appendix*, Table S12, Models 1 and 4 by the number of teachers in the fifteen intervention conditions (n = 110,948) and the average number of students per teacher among these teachers (21.32) and, in the case of the post-intervention period, summing the results of the products from *SI Appendix*, Table S12, Models 1 and 4 to cover the 8-wk post-intervention.

1. OECD, Data from "Mathematics performance (PISA)." https://www.oecd.org/en/data/indicators/mathematics-performance-pisa.html. Accessed 20 November 2023.
2. National Center for Education Statistics, 2022 NAEP mathematics assessment: Highlighted results at grades 4 and 8 for the nation, states, and districts. https://www.nationsreportcard.gov/highlights/mathematics/2022/. Accessed 12 January 2024.
3. E. M. Fahle *et al.*, School district and community factors associated with learning loss during the COVID-19 pandemic. https://cepr.harvard.edu/sites/hwpi.harvard.edu/files/cepr/files/explaining_covid_losses_5.23.pdf. Accessed 12 January 2024.
4. L. Eskreis-Winkler, K. L. Milkman, D. M. Gromet, A. L. Duckworth, A large-scale field experiment shows giving advice improves academic outcomes for the advisor. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 14808–14810 (2019).
5. D. S. Yeager *et al.*, A national experiment reveals where a growth mindset improves achievement. *Nature* **573**, 364–369 (2019).
6. E. A. Hanushek, The economic value of higher teacher quality. *Econ. Educ. Rev.* **30**, 466–479 (2011).
7. D. S. Yeager *et al.*, Teacher mindsets help explain where a growth-mindset intervention does and doesn't work. *Psychol. Sci.* **33**, 18–32 (2022).
8. A. L. Duckworth, K. L. Milkman, A guide to megastudies. *PNAS Nexus* **1**, pgac214 (2022).
9. J. G. Voelkel, J. Y. Chu, M. N. Stagnaro, J. N. Druckman, R. Willer, How to design and conduct a megastudy. *Nat. Hum. Behav.* **8**, 2257–2260 (2024), 10.1038/s41562-024-01998-2.
10. K. L. Milkman *et al.*, A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2115126119 (2022).
11. K. L. Milkman *et al.*, Megastudy shows that reminders boost vaccination but adding free rides does not. *Nature* **631**, 179–188 (2024).
12. K. L. Milkman *et al.*, A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2101165118 (2021).
13. K. L. Milkman *et al.*, Megastudies improve the impact of applied behavioural science. *Nature* **600**, 478–483 (2021).
14. J. G. Voelkel *et al.*, Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity. *Science* **386**, eadh4764 (2024).
15. Zearn Math, Top-rated math program created for teachers, by teachers (2024). Available at: https://about.zearn.org/ [Accessed 10 October 2023].
16. S. Hashim, Measuring the efficacy of Zearn Math in Louisiana. *AERA Open* **10**, 23328584241269825 (2024).
17. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300 (1995).
18. W. James, C. Stein, "Estimation with quadratic loss" in *Breakthroughs in Statistics, Springer Series in Statistics*, S. Kotz, N. L. Johnson, Eds. (Springer, New York, 1992), pp. 443–460.
19. Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).

20. M. A. Kraft, The effect-size benchmark that matters most: Education interventions often fail. *Educ. Res.* **52**, 183–187 (2023).
21. S. DellaVigna, E. Linos, RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica* **90**, 81–116 (2022).
22. J. Beshears, H. Kosowsky, Nudging: Progress to date and future directions. *Organ. Behav. Hum. Decis. Process.* **161**, 3–19 (2020).
23. M. T. Damgaard, H. S. Nielsen, Nudging in education. *Econ. Educ. Rev.* **64**, 313–342 (2018).
24. M. T. Damgaard, H. S. Nielsen, "Behavioral economics and nudging in education: Evidence from the field" in *The Economics of Education*, S. Bradley, C. Green, Eds. (Academic Press, ed. 2, 2020), pp. 21–35.
25. C. A. Hecht, C. J. Bryan, D. S. Yeager, A values-aligned intervention fosters growth mindset–supportive teaching and reduces inequality in educational outcomes. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2210704120 (2023).
26. D. H. Bailey, G. J. Duncan, F. Cunha, B. R. Foorman, D. S. Yeager, Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychol. Sci. Public Interest* **21**, 55–97 (2020).
27. J. Protzko, The environment in raising early intelligence: A meta-analysis of the fadeout effect. *Intelligence* **53**, 202–210 (2015).
28. N. C. Soderstrom, R. A. Bjork, Learning versus performance: An integrative review. *Perspect. Psychol. Sci.* **10**, 176–199 (2015).
29. V. Beermann, J. M. Enkmann, M. Maier, F. Bartoš, "How effective are digital green nudges? A publication bias-adjusted meta-analysis" in *Forty-Fifth International Conference on Information Systems, Bangkok, Thailand* (Association for Information Systems, 2024), pp. 1–5.
30. A. L. Duckworth *et al.*, Data from "A national megastudy shows that email nudges to elementary school teachers boost student math achievement, particularly when personalized." Open Science Framework. https://osf.io/gyhw2/?view_only=0be05c88030e4ab6ae5e6dd78ea5a08e. Deposited 5 September 2024.