Scaling up behavioral science interventions in online education

René F. Kizilcec[a,1,2], Justin Reich[b,1,2], Michael Yeomans[c,1,2], Christoph Dann[d], Emma Brunskill[e], Glenn Lopez[f], Selen Turkay[g], Joseph Jay Williams[h], and Dustin Tingley[f,i]

[a]Department of Information Science, Cornell University, Ithaca, NY 14850; [b]Comparative Media Studies/Writing, Massachusetts Institute of Technology, Cambridge, MA 02139; [c]Harvard Business School, Harvard University, Cambridge, MA 02138; [d]Machine Learning Department, Carnegie Mellon University, New York, NY 10004; [e]Computer Science Department, Stanford University, Stanford, CA 94305; [f]Office of the Vice Provost for Advances in Learning, Harvard University, Cambridge, MA 02138; [g]School of Computer Science, Queensland University of Technology, Brisbane City, QLD 4000, Australia; [h]Department of Computer Science, University of Toronto, Toronto, M5S 1A1 ON, Canada and [i]Department of Government, Harvard University, Cambridge, MA 02138

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved May 12, 2020 (received for review December 5, 2019)

Online education is rapidly expanding in response to rising demand for higher and continuing education, but many online students struggle to achieve their educational goals. Several behavioral science interventions have shown promise in raising student persistence and completion rates in a handful of courses, but evidence of their effectiveness across diverse educational contexts is limited. In this study, we test a set of established interventions over 2.5 y, with one-quarter million students, from nearly every country, across 247 online courses offered by Harvard, the Massachusetts Institute of Technology, and Stanford. We hypothesized that the interventions would produce medium-to-large effects as in prior studies, but this is not supported by our results. Instead, using an iterative scientific process of cyclically preregistering new hypotheses in between waves of data collection, we identified individual, contextual, and temporal conditions under which the interventions benefit students. Self-regulation interventions raised student engagement in the first few weeks but not final completion rates. Value-relevance interventions raised completion rates in developing countries to close the global achievement gap, but only in courses with a global gap. We found minimal evidence that state-of-the-art machine learning methods can forecast the occurrence of a global gap or learn effective individualized intervention policies. Scaling behavioral science interventions across various online learning contexts can reduce their average effectiveness by an order-of-magnitude. However, iterative scientific investigations can uncover what works where for whom.

behavioral interventions | scale | online learning

Behavioral scientists have argued that it is possible to intervene and modify personal habits, decisions, and thought patterns that contribute to social problems (1). Behavioral science interventions have been developed to promote a variety of prosocial behaviors, such as healthy eating habits, physical activity, getting medical check-ups, voting, and achievement in schools and colleges. While these interventions are usually low-cost—to participants and policy-makers—they are still thought to be effective because they target the psychological mechanisms underlying people's behavior (2). The ubiquity of networked devices has made it even easier to implement these interventions at large scale and to run field experiments that reveal their broader impact.

In this study, we conducted one of the largest global field experiments in higher education, with one-quarter million students across nearly every country, to examine the scalability of several behavioral science interventions that improved outcomes for thousands of students in our own prior research. Online education is rapidly expanding to address problems of educational access and meet the rising economic demands for professional development and retraining. For all this growth, many online students struggle to achieve their goals. Course completion rates are often low: Around 20% in Harvard University, the Massachusetts Institute of Technology (MIT), and Stanford University massive open online courses (MOOCs) among students who intend to complete (3, 4).

Online learning environments are well-suited to test the scalability of behavioral interventions. They have a well-defined outcome (course completion), requiring sustained effort. Student progress is continuously tracked through a common software platform. Improving outcomes in online learning through targeted support holds great promise for human capital development around the world. National education platforms have started using online courses to supplement college STEM (science, technology, engineering, and math) instruction (5) and students who complete MOOCs report benefits ranging from earning credit toward a degree to enhanced skills in a current job or finding a new job (6, 7). Moreover, there is evidence that students can transfer skills learned from MOOCs into real-world settings: They deploy new programming skills into open-source software projects, participate in scholarly activity following a research methods course, and develop new school initiatives after an education leadership course (8–10).

Following the joint Common Guidelines for Educational Research from the National Science Foundation (NSF) and Institute

Significance

Low persistence in educational programs is a major obstacle to social mobility. Scientists have proposed many scalable interventions to support students learning online. In one of the largest international field experiments in education, we iteratively tested established behavioral science interventions and found small benefits depending on individual and contextual characteristics. Forecasting intervention efficacy using state-of-the-art methods yields limited improvements. Online education provides unprecedented access to learning opportunities, as evidenced by its role during the 2020 coronavirus pandemic, but adequately supporting diverse students will require more than a light-touch intervention. Our findings encourage funding agencies and researchers conducting large-scale field trials to consider dynamic investigations to uncover and design for contextual heterogeneity to complement static investigations of overall effects.

Author contributions: R.F.K., J.R., M.Y., C.D., E.B., S.T., J.J.W., and D.T. designed research; R.F.K., J.R., M.Y., C.D., and G.L. performed research; R.F.K., J.R., M.Y., C.D., and G.L. analyzed data; and R.F.K., J.R., M.Y., and C.D. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]R.F.K., J.R., and M.Y. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: kizilcec@cornell.edu, jreich@mit.edu, or myeomans@hbs.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1921417117/-/DCSupplemental.

First published June 15, 2020.

for Education Sciences (IES) (11), we designed this study as a scale-up research impact study to examine "effectiveness in a wide range of populations, contexts, and circumstances, without substantial developer involvement in implementation or evaluation." We took existing interventions and deployed them with minimal ongoing adjustment across a range of courses. The courses included in the study sample spanned a remarkable range of subjects—from poetry to data science—and a diversity of students from nearly every country in the world. We leveraged the unprecedented variation in this platform to advance our understanding of how behavioral science intervention effects vary across students in MOOCs and different institutional and course contexts. The scalable interventions that we tested could be feasibly implemented by administrators or instructors hosting online courses in diverse settings.

Interventions and Prior Results

For 2.5 y, from September 2016 to May 2019, we added a randomly assigned intervention module at the start of nearly all MOOCs offered by Harvard, MIT, and Stanford (n = 269,169 students across 247 courses). We had previously published results from large field experiments for three of the five interventions in this study (12–14). These interventions had each been shown to substantially improve completion rates for targeted groups of students and replicated across courses. We scaled the interventions consistently by embedding a survey early in the materials of every course. After answering typical survey questions about themselves and their goals, students were randomly assigned to receive one of the intervention activities described below, or no activity in the control condition.

The "plan-making" interventions prompted students to concretely describe when and how they will complete required coursework for the entire course. Plan-making interventions target people's reluctance to forecast the procedural details of goal pursuit (15, 16). Previous work showed effects of plan-making on discrete behaviors like voting or doctor's appointments. We had tested plan-making interventions in three online courses on Business, Chemistry, and Political Science (n = 2,053) and found a 29% increase in course completion (from 14 to 18%) among committed English-fluent students (12). We had preregistered the analysis (https://osf.io/wq8m5/) and predicted the effect for this specific subpopulation. We use two versions of a plan-making activity in the present study: A replication of the previous paper (12), focused on long-term plans, and a short-term variant that asks students to plan for the first week only.

The "value-relevance" intervention is a motivational activity that asks students to indicate important values and write about how taking the course reflects and reinforces what is most important to them. The intervention builds on self-affirmation and utility-value intervention research that has been shown to reduce ethno-racial achievement gaps by lifting performance among disadvantaged students (17–19). Previous work tested, preregistered (https://osf.io/g8bu4/), and replicated this intervention in two online courses on Computing and Public Policy (n = 3,451) and found that it closed the global achievement gap between students in more-developed and less-developed countries by raising the completion rate of students in developing countries from 17 to 41% (13). Student's in the developed world remained unaffected in one course but experienced a decline in completion (from 32 to 23%) in the second course.

The "mental contrasting with implementation intentions" (MCII) intervention prompts students to reflect on the benefits and barriers to achieving their goal (20) and then plan ahead for how to overcome obstacles (15). We tested and replicated this intervention in two online courses, on Computing and on Sociology (n = 17,963), and found that it increased the completion rate by 15% (from 26 to 30%) and 32% (from 5.5 to 7.3%) for students in individualist countries (such as the United States and Germany) (14).

The culture-specific effect for the MCII intervention led us to hypothesize that students in less-individualist countries could benefit from a "social accountability" intervention that prompts them to make a plan to ask people to regularly check in about their course progress. This strategy can foster a sense of accountability that strengthens goal motivation (21). In political and education contexts, it has been shown to increase voting and school attendance (22), but unlike the other interventions, we had not previously tested it in online courses.

Results

Our primary hypothesis was that the main results from previous studies—improving course completion rates for targeted subgroups of students—would replicate in a larger sample. Overall, we did not find new evidence for our original large- to medium-sized effects. Specifically, the long-term planning prompts did

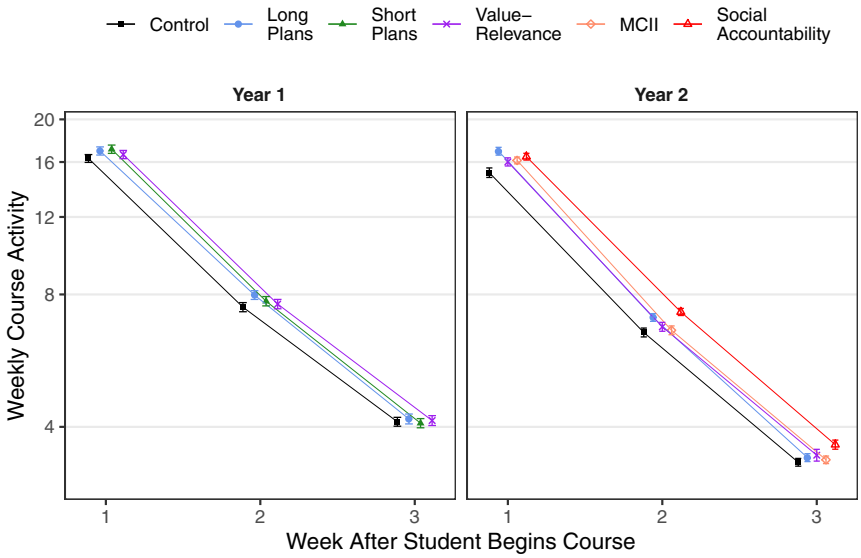

Fig. 1. Average student activity (count of course platform events) in the first 3 wk after exposure to each intervention. Points show covariate-adjusted means on a logarithmic scale (to match the log-transformed outcome in the regression model) with cluster-robust SE bars.

not improve course completion for committed English-fluent students (year 1: β = 0.19 percentage points [pp], 95% CI = [−0.69, 1.07], $t = 0.43$, $P = 0.670$; year 2: β = −0.23 pp, 95% CI = [−1.61, 1.14], $t = -0.33$, $P = 0.741$). The MCII intervention did not improve course completion for students in individualistic countries (year 2: β = 0.25 pp, 95% CI = [−0.87, 1.37], $t = 0.44$, $P = 0.662$), and the social accountability intervention did not improve completion rates for students in nonindividualistic countries (year 2: β = 0.89 pp, 95% CI = [−0.22, 1.99], $t = 1.58$, $P = 0.115$). Finally, the value-relevance intervention did not improve completion rates across all of the courses for students in less-developed countries (year 1: β = 0.14 pp, 95% CI = [−0.753, 1.03], $t = 0.30$, $P = 0.764$; year 2: β = −0.04 pp, 95% CI = [−1.37, 1.30], $t = -0.05$, $P = 0.957$).

A number of recent scale-up studies have failed to find effects of similar magnitude as the original studies, raising questions about whether the original effect estimates generalized beyond the original populations and contexts (23–29). Although our initial results were disappointing, our study was designed for multiple opportunities to preregister, test, explore, and then register and test updated hypotheses. In the process, we found consistent evidence for two benefits from the interventions that were more limited in scope than the original findings.

Planning Supports Short-Term Persistence. We found that the plan-making interventions slightly increased students' engagement during the first few weeks in the course. To measure persistence consistently across diverse course structures, we developed a statistical surrogate for intermediate course progress in our iterative research process (see *SI Appendix* for details). Specifically, we compiled student clickstream logs to measure their follow-up activity in the first weeks after their first day in the course. Fig. 1 shows students' daily activity for 3 wk following their exposure to the intervention, grouped by condition for the first and second year of the study.

Overall, the plan-making interventions increased students' week 1 activity levels in year 1 (short plans: β = 0.0437, 95% CI = [0.0066, 0.0808], $t = 2.31$, $P = 0.021$; long plans: β = 0.0336, 95% CI = [−0.0033, 0.0706], $t = 1.79$, $P = 0.074$) and in year 2 (long plans: β = 0.1057, 95% CI = [0.0496, 0.1619], $t = 3.69$, $P < 0.001$; MCII: β = 0.0603, 95% CI = [0.0039, 0.1167], $t = 2.10$, $P = 0.036$; social accountability: β = 0.0788, 95% CI = [0.0225, 0.1351], $t = 2.74$, $P = 0.006$). However, as Fig. 1 suggests, these treatment effects were short-lived. The effect on week 2 activity was attenuated in year 1 (short plans: β = 0.0257, 95% CI = [−0.0110, 0.0624], $t = 1.37$, $P = 0.169$; long plans: β = 0.0493, 95% CI = [0.0127, 0.0859], $t = 2.64$, $P = 0.008$) and year 2 (long plans: β = 0.0685, 95% CI = [0.0134, 0.1236], $t = 2.44$, $P = 0.015$; MCII: β = 0.0099, 95% CI = [−0.0451, 0.0649], $t = 0.35$, $P = 0.724$), except for a sustained effect of the social accountability intervention (β = 0.0939, 95% CI = [0.0387, 0.1491], $t = 3.34$, $P < 0.001$).

These results suggest that while planning prompts raised engagement in educational goal pursuit, their benefits dissipated over the span of a full course. Short-term effects may be consequential for tasks that require one-time behavior, such as voting or going to the doctor's (30, 31), but other work suggests that planning prompts may be unreliable for more complex goals (32, 33). Online education requires sustained effort toward complex, long-term goals, and the effects of our plan-making interventions attenuated after 1 to 2 wk, and were not detectable in the final course completion rates.

Value-Relevance Intervention Closes the Global Achievement Gap in Courses that Have One. We found that the value-relevance intervention predictably reduces the global achievement gap, insofar as there is a global achievement gap in the course. The gap is defined by the difference in completion rates between students in more-developed versus less-developed countries, as demarcated by 0.7 on the United Nations Human Development Index (HDI) (13). The gap was large on average (as in our previous studies), but it was not uniform across courses and occasionally even reversed. We therefore refined our hypothesis for the value-relevance intervention in the second year to specify that it would only be effective in courses with a significant global gap, defined as a 0.2 SD lower completion rate for students in less-developed than more-developed countries in the control condition.

In courses with a significant global gap, the value-relevance intervention increased the average completion rate among students in less-developed countries by 2.79 pp in the first year (95% CI = [1.30, 4.27], $t = 3.68$, $P < 0.001$) and by 2.74 pp in the second year (95% CI = [0.32, 5.17], $t = 2.22$, $P = 0.026$). The effect of the intervention is significant but an order-of-magnitude smaller than in our prior study (Table 1) (13). In courses without a global gap (or where it was reversed), post hoc analyses indicate that the intervention lowered the average completion rate among students in less-developed countries (year 1: β = −1.62 pp, 95% CI = [−2.73, −0.27], $t = -2.86$, $P = 0.004$; year 2: β = −1.71 pp, 95% CI = [−3.27, −0.16], $t = -2.16$, $P = 0.031$). While our prior study found that the intervention negatively affected students in more developed countries (13), we found no new evidence of this back-firing effect, neither in courses with a global gap (year 1: β = 0.45 pp, 95% CI = [−0.52, 1.43], $t = 0.91$, $P = 0.363$; year 2: β = −0.62 pp, 95% CI = [−2.46, 1.22], $t = -0.66$, $P = 0.509$) nor in courses without a global gap (year 1: β = −0.08 pp, 95% CI = [−0.83, 0.67], $t = -0.21$, $P = 0.835$; year 2: β = 0.94 pp, 95% CI = [−0.01, 1.89], $t = 1.95$, $P = 0.051$). The findings are visualized in Fig. 2.

Consistent with its theoretical underpinnings and prior results, the value-relevance intervention specifically benefits marginalized students in environments where they are at risk for encountering psychological barriers (17–19). The original intervention effect replicates in contexts that most resemble the original courses that featured a global achievement gap. In contrast, in courses without a global gap, it is counterproductive to provide a value-relevance intervention. This result highlights the need to account for contextual variation when scaling an intervention from a few select research sites to a broader set of contexts (34). Table 1 presents a comparison of results from our prior studies and the scaled-up versions in the present research.

Subsequent exploratory analyses revealed that several other interventions significantly improved completion rates for the same population: Students in less-developed countries in courses with a global gap (short plans year 1: β = 2.44 pp, 95% CI = [0.96, 3.92], $t = 3.23$, $P = 0.001$; long plans year 1: β = 2.74 pp, 95% CI = [1.27, 4.22], $t = 3.64$, $P < 0.001$; MCII year 2: β = 2.76 pp, 95% CI = [0.35, 5.18], $t = 2.24$, $P = 0.025$). However, the effect did not replicate in the second year for the plan-making intervention (β = 1.06 pp, 95% CI = [−1.33, 3.44], $t = 0.87$, $P = 0.386$) or the social accountability intervention (β = 1.23 pp, 95% CI = [−1.16, 3.62], $t = 1.01$, $P = 0.314$). The identification of a responsive subpopulation may be evidence of a common underlying mechanism for several interventions (i.e., self-reflective writing) or evidence that these students are receptive to a range of supports. The post hoc determination of which courses present a global gap may also contribute to this unexpected pattern of results.

Forecasting Where and for Whom an Intervention Will Work Is Challenging. Our findings suggest that policymakers and administrators who deploy behavioral science interventions should consider targeting specific students and contexts, such as students in less-developed countries in courses with a global achievement gap. However, the global gap is a characteristic that can only be determined after the course has run, as it depends on differences in completion rates in the control condition. We did

Table 1. Comparison of intervention results from prior research and this research for comparable interventions and subgroups of students

Intervention	Subpopulation	Prior result	Present result
Plan-making (long-term)	Committed English-fluent students	$\beta = 3.9$ pp, $\chi^2_{(1)} = 5.2$, $P = 0.023$, $n = 2{,}053$ (3 courses)	Year 1: $\beta = 0.19$ pp, $t = 0.43$, $P = 0.670$, $n = 26{,}586$ Year 2: $\beta = -0.23$ pp, $t = -0.33$, $P = 0.741$, $n = 10{,}372$
Value-relevance	Students in less-developed countries in courses with a global gap	Study 1: $\beta = 3.4$ course activities, $z = 2.82$, $P = 0.005$, $n = 227$ Study 2: $\beta = 24$ pp, $z = 2.26$, $P = 0.024$, $n = 64$	Year 1: $\beta = 2.79$ pp, $t = 3.68$, $P < 0.001$, $n = 5{,}974$ Year 2: $\beta = 2.74$ pp, $t = 2.22$, $P = 0.026$, $n = 2{,}712$
Mental contrasting with implementation intentions	Students in individualistic countries	Study 1: $\beta = 1.8$ pp, $z = 2.35$, $P = 0.019$, $n = 4{,}628$ Study 2: $\beta = 3.9$ pp, $z = 2.41$, $P = 0.016$, $n = 3{,}248$	Year 2: $\beta = 0.25$ pp, $t = 0.44$, $P = 0.662$, $n = 12{,}879$

Note that there are several differences between the prior and present research in terms of the implementation of intervention instructions and sample exclusion criteria. Effects denote percentage point (pp) increases in course completion except where noted.

not find patterns to forecast when the global gap will occur; it appears uncorrelated with institution (Harvard, MIT, Stanford), subject domain (STEM versus Humanities, or at the level of department or program), or any other features we examined. In fact, among the 79 course offerings that were repeats of courses that had been offered previously, the presence of a global gap in repeat offerings matched the original offering only 60.8% of the time. A predictive model (see *SI Appendix* for details) with 21 course-level features in year 1 could not forecast the occurrence of a global gap in year 2 significantly better than random chance (54.3% accuracy, 95% CI= [44.1, 64.4], compared to 50.0%, 95% CI = [39.8, 60.2]).

One possibility is that behavioral science interventions need to be targeted at a fine-grained individual level, but our analysis suggests that this would have at most a modest impact on course completion. We performed an exploratory analysis using machine-learning algorithms to optimize an individualized policy (*SI Appendix*) using data collected in year 1. We then estimated that the average completion rate of this personalized policy in year 2 is 13.38% (95% CI = [12.79, 13.98]). This is slightly but not significantly higher than the estimated average completion rate of no intervention (12.81%, 95% CI = [12.23, 13.39]) or a randomly assigned intervention (13.08%, 95% CI = [12.74, 13.38]). To realize the potential benefit of personalized policies for students, the field will require more effective interventions and more comprehensive collection of individual- and course-level features to identify which students will benefit from particular supports.

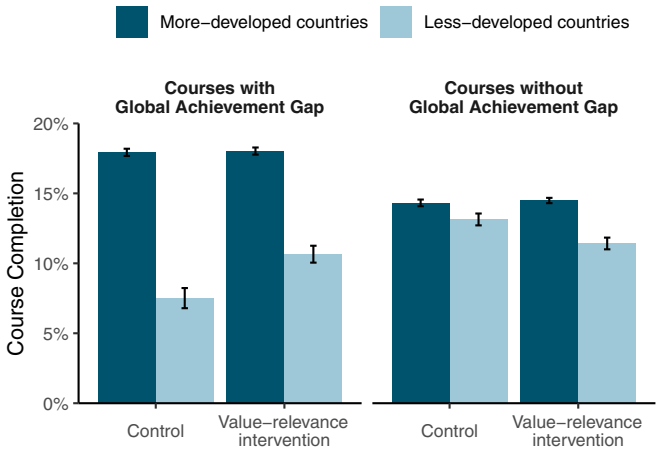

Fig. 2. Average course completion rate in all waves in the value-relevance intervention and control condition by student context (more vs. less developed country) and course context (with vs. without global achievement gap). Bars show covariate-adjusted means with cluster-robust SE bars.

Discussion

Our preregistered analyses demonstrate that a value-relevance intervention improves course completion for students in less-developed countries in courses with a global achievement gap. Post hoc analysis suggests that our other interventions may similarly improve outcomes for these specific students in these specific contexts. Our finding that plan-making interventions have limited benefits provides further evidence that behavioral insights have more promise in encouraging one-time, short-lived actions than more continuous behaviors that require sustained effort and habit change. These conclusions are consistent with a number of recent scale-up studies in other domains that have found diminished scope and magnitude for behavioral interventions (23–29, 32).

In our original studies, we recommended that policymakers and online instructors consider employing the tested behavioral interventions in their own MOOCs (12–14); we now conclude that further research is necessary to predict in advance when these interventions will help populations of students in need. The population of students who complete surveys in MOOCs is exceptionally diverse along certain dimensions, but may not capture aspects of other populations of interest: For example, less engaged students or those in other educational settings. As such, we believe caution is warranted in applying our findings beyond motivated and self-directed students in open-enrollment courses.

The kind of large-scale research that is needed to advance this work is not well-represented in the dominant paradigm of experimental educational research. The NSF/IES Common Guidelines for Education Research define a trajectory for experimental research that proceeds from pilot studies in laboratories, to initial implementations in field sites, to scale-up studies designed to generate "reliable estimates of the ability of a fully-developed intervention or strategy to achieve its intended outcomes" across multiple, diverse, real-world contexts (11). Many large grants available to researchers require that they hold their intervention constant across contexts.

Our present study confirms a principle that is central to social psychology and the learning sciences: Context matters. Alongside large-scale studies that test a single, fully developed intervention across multiple contexts, "scale-up" funding should be available for approaches that assume interventions will need to

be constantly refined and modified to support specific groups of people across diverse contexts. These studies would be designed to respond to concerns of temporal validity, the notion that the effectiveness of interventions can vary as contexts and populations change over time (35). Rather than treating large-scale studies as the conclusion of a research trajectory, scale-up studies should support new research into context-level variation that cannot be explored in small field trials. We encourage greater focus on the characteristics of different contexts that induce variation in the effects of interventions to advance the development of a science of context in education. In a new paradigm, the question of "what works?" would be replaced with "what works, for whom, right here?"

Materials and Methods

Adopting best practices from open science, we conducted our study in four preregistered "waves" of implementation. In each wave, we preregistered hypotheses and analysis code, collected data, conducted post hoc analyses of heterogeneous treatment effects, and refined the preregistration for the subsequent wave. All preregistrations, analysis code, and data are available online at https://osf.io/9bacu/. Our study plans were reviewed and approved by the Institutional Review Boards at Harvard University, Stanford University, and MIT. Participants consented to research participation in agreeing to the terms of service as part of the site registration process for EdX (Harvard, MIT) and Open EdX (Stanford). EdX participants consented to research participation again before taking the survey; Open EdX prominently displayed a message on every course enrollment page notifying participants that they participate in research by using the platform.

In the first year (waves 1 to 2; September 2016 to December 2017), we tested the value-relevance and short- and long-term plan-making interventions, individually and in combination, across 153 courses. In the second year (waves 3 to 4; January 2018 to May 2019), based on early findings, we shortened the overall survey, eliminated the short-term plan-making intervention, simplified the value-relevance intervention, and added MCII and social accountability interventions. We collected data in 94 courses. In total, the 247 focal courses include all courses offered by the three institutions during this time period, unless a course did not implement the survey, had fewer than 100 students assigned to a condition, or less than 1% of students assigned to a condition completed the course. The focal courses span a wide range of subjects (22% humanities, 40% social science, 29% STEM, 9% computer science), sizes (between 102 and 16,645 students assigned to a condition), and completion rates (from 1 to 65% among those assigned to condition). We define the focal sample of students for our main analyses to be everyone assigned to a condition for the first time (*SI Appendix*): 199,517 students in year 1 and 69,652 students in year 2.

Our primary outcome measure, borrowed from our previous research, is course completion operationalized as earning a passing grade in the course. We also developed a surrogate outcome (36) to measure proximate intervention effects on early course engagement in terms of the log-scaled number of actions students performed in the first week (days 2 to 8) and the second week (days 9 to 16). We initially created a continuous measure of progress in the course based on the percentage of videos viewed or assignments completed, but found that the wide heterogeneity of course models prevented reasonable comparisons. Treatment effects were estimated using preregistered regression models with individual- and course-level covariates and course fixed effects. Hypotheses that specify a subgroup effect were tested by fitting the model only for those students (see *SI Appendix* for additional methodological details, including outcomes tested, exclusion criteria, and model specifications).

Data Availability. We provide de-identified datasets to run many of the analyses reported in the paper, hosted on the Harvard Dataverse, which is linked from within our Open Science Framework (OSF) repository (https://osf.io/9bacu/). Some variables have been anonymized (e.g., course names have been hashed), and other variables have been binned to preserve privacy (e.g., number of courses finished or country HDI). Where possible, all of the analyses reported in the paper are conducted on this anonymized dataset, and we confirm that the results are substantively identical when the same analyses are conducted on the raw dataset. However, some analyses reported in the paper cannot be conducted on anonymous data (e.g., tables of demographic descriptives and the adaptive learning policy), so we only include the analysis code and results.

The full raw datasets analyzed in this study are not publicly available to protect the privacy of personally identifiable student information, and restrictions apply to the availability of these data, which were used under license from Harvard, MIT, and Stanford. The datasets are maintained separately by each institution, and contacts and/or guidelines for data requests are available at ir.mit.edu/mitx-data-request-checklist, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/RTVIEM, and https://online.stanford.edu/about-us.

Code Availability. All analysis code, output, and study materials are available at https://osf.io/9bacu/.

ACKNOWLEDGMENTS. This work would not have been possible without Jon Daries, Kathy Mirzaei, Andreas Paepcke, and Yigal Rosen. We thank Yifei Zheng and Tiffany Zhong for assistance coding courses. This material is based upon work supported by National Science Foundation Grant 1646978, a Stanford Interdisciplinary Graduate Fellowship, and a Microsoft Faculty Fellowship.

1. C. R. Sunstein, R. H. Thaler, *Nudge: Improving Deisions About Health, Wealth, and Happiness*, (Yale Univeresity Press, New Haven, 2008).
2. G. M. Walton, T. D. Wilson, Wise interventions: Psychological remedies for social and personal problems. *Psychol. Rev.* **125**, 617–655 (2018).
3. J. Reich, *MOOC Completion and Retention in the Context of Student Intent* (EDUCAUSE Review Online, 2014) https://er.educause.edu/articles/2014/12/mooc-completion-and-retention-in-the-context-of-student-intent.
4. R. F. Kizilcec, S. Halawa, "Attrition and achievement gaps in online learning" in *Proceedings of the Second ACM Conference on Learning @ Scale*, (Association for Computing Machinery, 2015), pp. 57–66.
5. I. Chirikov, T. Semenova, N. Maloshonok, E. Bettinger, R. F. Kizilcec, Online education platforms scale college STEM instruction with equivalent learning outcomes at lower cost. *Sci. Adv.* **6**, eaay5324 (2020).
6. C. Zhenghao *et al*., Who's benefiting from MOOCs, and why. *Harv. Bus. Rev.* **25**, 2–8 (2015).
7. J. Littenberg-Tobias, J. Reich, Evaluating access, quality, and equity in online learning: A case study of a MOOC-based blended professional degree program. *SocArXiv*, 10.31235/osf.io/8nbsz (7 December 2018).
8. G. Chen, D. Davis, C. Hauff, G. J. Houben, "Learning transfer: Does it take place in MOOCs? An investigation into the uptake of functional programming in practice" in *Proceedings of the Third ACM Conference on Learning@ Scale*, (Association for Computing Machinery, 2016), pp. 409–418.
9. Y. Wang, L. Paquette, R. Baker, A longitudinal study on learner career advancement in MOOCs. *J. Learn. Anal.* **1**, 203–206 (2014).
10. A. Napier, E. Huttner-Loan, J. Reich, Evaluating transfer of learning from MOOCs to workplaces: A case study from teacher education and Launching Innovation in Schools. *Revista Iberoamericana de Educación a Distancia* **23** (2), 45–64 (2020).
11. Joint Committee of the Institute of Education Sciences, Department of Education, and the National Science Foundation, *Common Guidelines for Education Research and Development* (IES, DOE, and NSF, Washington, DC, 2013) https://www.nsf.gov/pubs/2013/nsf13126/nsf13126.pdf.
12. M. Yeomans, J. Reich, "Planning to learn: Planning prompts encourage and forecast goal pursuit in online education" in *Proceedings of the Seventh International Conference on Learning Analytics & Knowledge*, (Association for Computing Machinery, 2017), pp. 464–473.
13. R. F. Kizilcec, A. J. Saltarelli, J. Reich, G. L. Cohen, Closing global achievement gaps in MOOCs. *Science* **355**, 251–252 (2017).
14. R. F. Kizilcec, G. L. Cohen, Eight-minute self-regulation intervention raises educational attainment at scale in individualist but not collectivist cultures. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 4348–4353 (2017).
15. P. M. Gollwitzer, P. Sheeran, Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Adv. Exp. Soc. Psychol.* **38**, 69–119 (2006).
16. T. Rogers, K. L. Milkman, L. K. John, M. I. Norton, Beyond good intentions: Prompting people to make plans improves follow-through on important tasks. *Behav. Sci. Policy* **1**, 33–41 (2015).
17. G. L. Cohen, J. Garcia, N. Apfel, A. Master, Reducing the racial achievement gap: A social-psychological intervention. *Science* **313**, 1307–1310 (2006).
18. C. S. Hulleman, J. M. Harackiewicz, Promoting interest and performance in high school science classes. *Science* **326**, 1410–1412 (2009).
19. R. F. Kizilcec, G. M. Davis, G. L. Cohen, "Towards equal opportunities in MOOCs: Affirmation reduces gender & social-class achievement gaps in China" in *Proceedings of the Fourth ACM Conference on Learning @ Scale*, (Association for Computing Machinery, 2017), pp. 121–130.
20. G. Oettingen, Future thought and behaviour change. *Eur. Rev. Soc. Psychol.* **23**, 1–63 (2012).
21. A. T. Hall, D. D. Frink, M. R. Buckley, An accountability account: A review and synthesis of the theoretical and empirical research on felt accountability. *J. Organ. Behav.* **38**, 204–224 (2017).

22. T. Rogers, N. J. Goldstein, C. R. Fox, Social mobilization. *Annu. Rev. Psychol.* **69**, 357–381 (2018).
23. P. Hanselman, C. S. Rozek, J. Grigg, G. D. Borman, New evidence on self-affirmation effects and theorized sources of heterogeneity from large-scale replications. *J. Educ. Psychol.* **109**, 405–424 (2017).
24. D. S. Yeager *et al*., A national experiment reveals where a growth mindset improves achievement. *Nature* **573**, 364–369 (2019).
25. C. R. Dobronyi, P. Oreopoulos, U. Petronijevic, Goal setting, academic reminders, and college success: A large-scale field experiment. *J. Res. Educ. Eff.* **12**, 38–66 (2019).
26. P. Oreopoulos, U. Petronijevic, "The remarkable unresponsiveness of college students to nudging and what we can learn from it (Tech. Rep. No. w26059)" (National Bureau of Economic Research, Cambridge, MA, 2019).
27. H. Lortie-Forgues, M. Inglis, Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educ. Res.* **48**, 158–166 (2019).
28. K. A. Bird *et al*., "Nudging at scale: Experimental evidence from FAFSA completion campaigns. *NBER working paper No. 26158*" (National Bureau of Economic Research, Cambridge, MA, 2019).
29. A. S. Kristal, A. V. Whillans, What we can learn from five naturalistic field experiments that failed to shift commuter behaviour. *Nat. Hum. Behav.* **4**, 1–8 (2019).
30. D. W. Nickerson, T. Rogers, Do you have a voting plan? Implementation intentions, voter turnout, and organic plan making. *Psychol. Sci.* **21**, 194–199 (2010).
31. K. L. Milkman, J. Beshears, J. J. Choi, D. Laibson, B. C. Madrian, Planning prompts as a means of increasing preventive screening rates. *Prev. Med.* **56**, 92–93 (2013).
32. C. Townsend, W. Liu, Is planning good for you? The differential impact of planning on self-regulation. *J. Consum. Res.* **39**, 688–703 (2012).
33. J. Beshears, H. N. Lee, K. L. Milkman, R. Mislavsky, J. Wisdom, Creating Exercise Habits Using Incentives: The Tradeoff between Flexibility and Routinization. *Manag. Sci.*, in press.
34. H. Allcott, Site selection bias in program evaluation. *Q. J. Econ.* **130**, 1117–1165 (2015).
35. K. Munger, The limited value of non-replicable field experiments in contexts with low temporal validity. *Soc. Media Soc.* **5**, 2056305119859294 (2019).
36. R. L. Prentice, Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat. Med.* **8**, 431–440 (1989).