PNAS
www.pnas.org

Supplementary Information for

Scaling Up Behavioral Science Interventions in Online Education

René F. Kizilcec[1*], Justin Reich[2*], Michael Yeomans[3*], Christoph Dann[4], Emma Brunskill[5], Glenn Lopez[3], Selen Turkay[6], Joseph J. Williams[7], Dustin Tingley[3]

1 Cornell University
2 Massachusetts Institute of Technology
3 Harvard University
4 Carnegie Mellon University
5 Stanford University
6 Queensland University of Technology
7 University of Toronto

* Co-equal First Authorship

Corresponding Authors: René F. Kizilcec (kizilcec@cornell.edu), Justin Reich (jreich@mit.edu), Michael Yeomans (myeomans@hbs.edu)

This PDF file includes:

Supplementary text
Supplementary Figures S1-S2
Supplementary Tables S1-S6
Intervention Materials

www.pnas.org/cgi/doi/10.1073/pnas.1921417117

Supplementary Information

Contents:

Methods Overview

We report the results of a series of interventions implemented in online courses offered through the EdX (and openEdX) platform by Harvard University, Stanford University, and the Massachusetts Institute of Technology. Courses were selected based on their start dates in the period from September 2016 to June 2019. The implementation of interventions was constant across courses. The interventions were embedded in a course survey given to students at the beginning of the course, presented early in the courseware within the first page introduction (see Figure S1). Most courses already had a start-of-course survey with questions about demographics and prior experience. We standardized the content of this survey across institutions to better compare student characteristics across courses and institutions. The interventions were designed to be scalable so that every new course that launched during this period could participate by copying our survey intervention template.

Adopting best practices from open science, we conducted our study in four pre-registered "waves" of implementation. In each wave, we pre-registered hypotheses and analytic code, collected data, conducted post-hoc investigations of heterogeneous treatment effects, and refined pre-registrations for the subsequent wave. All pre-registrations, analysis code, analytic output, and guidelines for requesting data from the host institutions are available online at https://osf.io/9bacu/. Our study plans were reviewed by the Committee on the Use of Human Subjects at Harvard (MIT ceded review to Harvard) and the Stanford Research Compliance Office.

In the first year (waves 1-2; 9/2016-12/2017), we tested the value-relevance and short- and long-term plan-making interventions, individually and in combination. We collected data in 153 courses from 199,517 students in our focal sample. In the second year (waves 3-4; 1/2018-5/2019), based on the early findings, we updated the survey to be shorter and simplified the value-relevance intervention activity. We also dropped the short-term plan-making, and added mental contrasting with implementation intentions (MCII) and social accountability interventions. We collected data in 94 courses from 69,652 students in our focal sample.

Over the course of the experiment, both the contents of the survey and our planned analyses evolved. We divide the dataset into four "waves", based on which courses were finished at the time we constructed a dataset. Each wave had a distinct analysis plan. Furthermore, waves 1 and 2 had an identical survey and experimental design, which was altered for waves 3 and 4.
In all waves, the randomization to condition was conducted late in the survey, immediately before exposure to the interventions. As expected, a substantial fraction of students assigned to treatment did not engage with the interventions at all - either skipping the writing prompt entirely, or else abandoning the course survey to explore the rest of the course. However, all of our analyses are conducted with intent-to-treat estimators for all students who were randomized into

condition and meet our inclusion criteria, regardless of their level of engagement with the intervention.

Interventions

Waves 1 & 2 Interventions. Assignment to conditions was random, but stratified to balance four self-report variables that were strong predictors of course completion in previous research (1). The four variables used to stratify assignment across courses were: *intentions to complete the course assessments [all or most; few or none (or blank)]*; *intended hours spent on the course per week [0-5 (or blank), 6+]; previous MOOCs completed [0 (or blank),1-3,4+];* and *previous education [graduate degree, bachelor's degree, less than bachelors (or blank)]*. We used stratified assignment to increase the statistical power of our design. Conditional on this stratification, assignment to intervention condition was completely random.

The number of students assigned to each condition was distributed evenly (X^2 = 8.66, P = 0.123) and students in the treatment conditions were representative of those in the control condition in terms of seven covariate measures (intentions to complete the course assessments, intended hours spent on the course per week, previous MOOCs completed, previous education level, English language fluency, geographic location in a low-HDI country, and the number of days from the start of the course until the student took the intervention survey). We calculated the absolute standardized difference (ASD) between each treatment condition and the control condition for each covariate. The maximum ASD out of 42 comparisons (7 variables x 6 treatments) was 0.011 SD and the average was 0.004 SD, an order of magnitude below the commonly used balance threshold of 0.1 SD (1).

These waves included two types of intervention - *value-relevance* and *plan-making* - in a 2x3 design, for a total of five intervention conditions and one control condition. The value-relevance factor had two levels: control (no intervention activity) and value-relevance intervention. The plan-making factor had three levels: control (no intervention activity), short plans, and long plans. In the two conditions with multiple interventions, the order was constant such that value-relevance was first and plan-making was second. We held the order constant because it aligns with common theories of goal pursuit (2) to focus on implementation after building motivation, rather than vice versa.

The value-relevance intervention is designed to reduce a social psychological barrier that can hinder students from achieving their potential in a learning environment (see intervention materials in SI for full text). Students can be concerned about their belonging in the MOOC and worry that others could see them as less capable because of one of their social identities (e.g., nationality, gender, social class). In MOOCs offered by elite universities, students from underprivileged backgrounds may doubt their "fit" into the archetypal role of a high-achieving student at such an institution. The value-relevance intervention draws on two established intervention paradigms: values affirmation based on self-affirmation theory (3) and utility-value interventions, grounded in the expectancy-value theory of motivation, which asks students to foster connections between course materials and their life (4). Students select 2-3 values or qualities from a list that are most important to them. They then write about how taking this online course reflects and serves their cherished values. To support internalization, students then write a note to their future self to remind them of the importance of their values in the course. Previous research (5) tested this intervention in two MOOCs which drew from many different countries, and showed that the course completion rates among people in developing countries approximately doubled, while slightly reducing completion rates in more developed countries. Moreover, treated students were more likely to sign up for future courses, indicating lasting effects to close the global achievement gap. Another experiment in a Chinese MOOC found that a value-relevance intervention, adapted for the interdependent cultural context in China, increased completion rates by 41% among the most identity-threatened subgroup of students in a language course, lower-class men (6).

The plan-making intervention aimed to reduce the logistical psychological barriers to course completion. A lack of strategic planning can impede follow-through on many kinds of long-term goals, often because people tend to ignore the procedural details of goal pursuit (7, 8). In many cases when people forecast their future goal pursuits, they naively think that their strong intentions will sustain over time, while overlooking implementation factors that can often make the difference between follow-through and attrition (9). The planning prompt asks students to consider those details while their intentions are strong, including specific times, locations, and study strategies. Students are also encouraged to write down their self-generated plans as a reminder. Previous research (10) has shown that planning prompts in three MOOCs increased course completion by 29%, and may also increase students' willingness to upgrade to verified certificates.

In this experiment, we included two different plan-making interventions: *long plans*, which asked students to plan their participation over the entire course; and *short plans*, which asked students to plan their participation over the first week of the course. We anticipated that both interventions might work for different reasons and for different lengths of time, as a window into the mechanism of planning prompts. The short plans intervention might allow students to set sub-goals and write more concrete plans, which might then induce habit formation or other long-term behaviors that are consistent with their initial intentions. In fact, the best evidence for plan-making effects in the field were collected from short-term one-shot behaviors (voting, or a doctor's appointment; see 7, 8) Alternatively, the short-term plans may not carry over into the long term, or perhaps give students an illusory feeling of completion (thus reducing long-term follow-through). Given these competing predictions, we decided to compare these two versions of the plan-making prompt.

Waves 3 & 4 Interventions. The results from Waves 1 & 2 suggested some room for improvement in our experimental design. In particular, the intervention uptake rates were lower than we anticipated. The courses in our wide sample took different approaches to sharing the survey link, which affected survey entry rates; mid-survey attrition was also more common than anticipated. We shortened the length of the survey pre-intervention to reduce attrition. Several pre-intervention questions were removed because over time we deemed them unnecessary and likely to cause survey fatigue. We also moved a question about forecasting course completion to before the intervention to use it as a pre-treatment covariate.

The intervention section of the survey underwent multiple changes. Rather than stratified randomization across courses, we used complete random assignment within each survey (implemented by the randomizer block in Qualtrics). Moreover, the study no longer used a factorial design; instead, every student was presented with either one intervention or none. We confirmed that the number of students assigned to each condition was distributed evenly ($X^2 = 0.66$, $P = 0.956$) and students in the treatment conditions were representative of those in the control condition in terms of the same seven covariate measures described above for the previous waves. The maximum ASD out of 35 comparisons (7 variables x 5 treatments) was 0.024 SD and the average was 0.007 SD, which indicates minimal evidence of imbalance.

The set of interventions also changed. We retained modified versions of the value-relevance and long plans interventions (but dropped the short plans intervention, because we lost confidence that it might outperform the long plans condition). We also added two other interventions: mental contrasting with implementation intentions ("MCII" for short) and a social accountability intervention. The MCII intervention was tested in two online courses and found that it increased completion rates for students in individualist countries (such as the United States and Germany) but not collectivist countries, such as India and China (11). The social accountability intervention prompted students to make a plan to ask people to regularly check in about their course progress. This social-regulation intervention was thought to better support students in collectivist countries.

Measures

In the main text, we report the results for the final set of outcome measures: course completion, defined as earning a passing grade in the course, as a long-term outcome measure; and weekly course activity, defined as the log-transformed count of student events in a week, as a short-term outcome measure. The set outcome measures we consider in this study evolved over waves, as described in detail below.

Wave 1 Outcome Measures. The EdX and Open EdX platforms passively log all course activity. This offers an opportunity to directly measure outcomes for all students enrolled in the course, without self-reporting. A potential downside of this fine-grained dataset is that there are a potentially infinite set of outcomes from which to select our metric of intervention effectiveness. Accordingly, our initial pre-registration defined a set of primary and secondary outcomes, which we refined over the course of our research.

Throughout all waves, our primary focus was to test whether the interventions helped students complete the course. We therefore assessed whether a student had achieved a high enough grade in the course to earn a “basic” certificate or statement of accomplishment. Before the start of our intervention, EdX (i.e. the platform for MIT and Harvard courses) had recently changed its certification policy, such that students could not achieve an official designation of course completion without paying for the “verified track” in the course. Before 2016 (when our preliminary research was conducted), students could earn a certificate for free, but these basic certificates no longer existed for courses at MIT and Harvard. In contrast, Stanford courses in this research did not offer paid “verified” certificates, only free statements of accomplishment (equivalent to a “basic” certificate). For consistency across courses on different platforms, we decided to use the course tracking logs for Harvard and MIT courses to determine whether a student had achieved a high enough grade to earn a “verified” certificate, even if the student did not pay. To achieve this, we used the assessment grades for students who were not on the verified track and applied the same calculations and threshold as for students on the verified track (e.g., overall grade above 70%) to determine if a basic certificate was achieved. We confirmed that our results were robust to alternative calculations.

Due to the variation in course structure, we constructed a secondary proxy for course progress that could be measured across all certificate types at all schools. To this end, we computed the percentage of videos in the course that a student at least began to watch, according to the tracking logs. In courses that did not include instructional videos, the percentage of assessments attempted was used.

To quantify the economic effects of our interventions, we additionally assessed the percentage of direct payments from students to enter the “verified” track of a course for an opportunity to earn a verified certificate. This option was only available in Harvard and MIT courses. Critically, students could pay to upgrade to the verified track before the course start (and before our intervention) or anytime during the course. We therefore removed students who upgraded before the course began to assess the rate of post-survey upgrades across conditions.

Finally, we included a tertiary, post-treatment measure of the mechanism for behavior change. It assesses the immediate effects of the interventions on students’ expectations of success. At the end of the course survey, after the intervention activities, students predicted the percentage likelihood that they would complete the course, on a sliding scale from zero to one hundred.

Waves 2-4 Additional Outcome Measures. Based on the results from Wave 1, we discarded some secondary outcome measures that posed unexpected measurement challenges, as we were merging data from courses with substantial variation in course content and assessment policies. We narrowed our focus to two of the original outcome measures that mattered most: course completion, which we define as earning a passing grade in the course (binary) and final grade (continuous 0-100). We also used Wave 1 data to develop early measures of course persistence (i.e. statistical surrogates) and capture any proximal treatment effects.

A challenge of this study context is the long time horizon between the interventions and the outcomes of interest. This allows for numerous unobservable influences on students' long-term goal pursuit (including other demands or distractions, or shifting preferences) to interfere with their progress online. Although long-term outcomes are the ultimate goal of our interventions, they are likely to be (at best) noisy indicators of the direct effects of treatment. To improve the precision of our measurements, we decided to use the raw tracking data from the courses to develop a short-term measure of course engagement, over the first week after a student is exposed to treatment.

Tracking data is high-dimensional and there are many ways to be active within a course (reading, lectures, forums, quizzes, etc.) and many ways to allocate time to complete the course. As there are no official short-term engagement metrics provided by the online course platforms, we developed our own set of metrics. We build on an internal aggregation dataset that tallies each student's activities on every day of the course (days were delineated by UTC time). For each student, the "day" on which they took the survey was demarcated as their day zero (even if they started the survey at 11:59PM on that day). To capture persistence beyond the initial activity after first entering the course, we dropped data from day 1 because it can contain spillover activity from the end of day zero. Figure S2 shows the average event counts for each day after a student enrolls. The general trend reflects a common structure in MOOCs - activity levels tend to wane over time with weekly spikes due to the timed release of new materials in some courses. Early activity is correlated with student intentions to complete the class, thus we use this measure as a proxy for early course persistence. We also found that these early measures correlated well (observationally) with longer-term outcomes - in our focal sample in wave 1, we found that the final course grade correlated strongly with our measures of course activity during each of the first three weeks (week 1: r=.418, week 2: r=.512, week 3: r=.550).

In Wave 2, we pre-registered a single measure for the short-term effects of our treatment conditions, calculating a binary measure of "early activity", that simply indicated whether a student had logged at least two active days during the window of days 2-8 inclusive. In the distribution of daily event counts over this first week, we found a noticeable spike at four student-initiated events per day, so we defined an "active day" as one in which a student had initiated at least four events.

In Waves 3, we added a slightly longer-term outcome. Using the definition of event counts from week one (days 2-8), we applied the same definition to week two (days 9-16). In Wave 4, we decided to drop the "active day" definition and just calculate (log-transformed) user-generated event totals for each of the first two weeks. We confirmed that our results are robust to both definitions.

Participants

In the main text, we report results for our focal sample which is determined using course-level and individual-level exclusions from our Wave 4 pre-registration. We first exclude students who were not assigned to a treatment; then we exclude courses in which the remaining sample is below 100 or the completion rate below 1% (this excluded 15 courses in Year 1 and 20 in Year 2); then we exclude students who were exposed multiple times or shortly before the analysis. For the Wave 3 & 4 focal sample, we additionally exclude students who wrote fewer than 4 characters in an open-ended question about their course goals, presented before the intervention and designed to screen out likely non-compliers. Below we describe all planned exclusions as pre-registered to explain how our sample determination strategy evolved across waves.

Wave 1 Sample Determination. The intervention was added to every course offered by Harvard, and MIT on the EdX platform, and by Stanford on the OpenEdX platform, with a start date after Sept 1, 2016. Only a handful of courses were excluded due to administrative hurdles or because the course format differed substantially from the other courses (e.g., a course that lasts just one week). The transition to the Wave 3 and 4 survey was initiated October 2017, though the uptake was slower for some courses because not all courses updated to the new survey in a timely fashion. Although we could anticipate which courses would be included in our dataset, we could not determine our sample size in advance because students enroll on their own volition. We report sample size statistics in Table S1, after each exclusion criterion and across each of the four waves. The population of online students is highly diverse, and we report descriptive demographic information about the sample of students who entered the survey and more detailed statistics within the focal sample for each wave in Table S2.

The survey was conducted through an embedded Qualtrics link or in-frame window, which created a record for every student who interacted with the survey. Many students do not open the survey, but prior work shows that most of these students also do not engage much in the course: the response rate is much higher among more active and committed students (12). Our intervention activities were located at the end of the survey, and many students did not progress far enough in the survey to be randomized into an intervention group. We exclude un-randomized students from our analyses.

In addition, we excluded students for several other reasons. First, students could take multiple courses simultaneously, which can cause contamination across treatments. We kept a student's first exposure and excluded all later sign-ups. Second, we excluded first exposures from students who signed up for multiple courses within a thirty-day window (and thus could have been exposed to other treatments during their initial course). Third, most courses allow students to enroll at any time, and complete the course survey whenever they want. However, most dropouts occur early, and those who take the survey mid-course may not be affected by the intervention; and students who join late may not be able to finish the course due to structural reasons such as deadlines, grading periods, and closing dates. In summary, all our initial analyses focused on students who:

(a) progressed far enough in the survey to be assigned to conditions and exposed,
(b) were not exposed again within 29 days following their initial exposure,
(c) started the survey within the first hour of accessing any course materials,
(d1) started the course within 14 days of the start of a cohort-based course, OR
(d2) started the course more than 30 days before the end of a self-paced course.

To assess how these exclusions influence the sample characteristics of students in the focal sample, we compare the focal sample to all students who entered the survey, an indicator of engagement with course content. Table S3 provides descriptive statistics for each subgroup in Waves 1 & 2 and Waves 3 & 4. The focal sample closely resembles the broader set of engaged students who entered the survey.

Wave 2 Sample Determination. Between Wave 1 and Wave 2, we decided to revise our sample exclusion rules (and we report results using the Wave 2 analysis plan). Before, we used several filters designed to exclude students who had atypical timing with respect to their engagement with the intervention (items b-d above). For example, we sought to focus on students who got the intervention within the first hour of having entered the course. However, we discovered that this was hard to determine reliably. The primary issue was that some courses sent a link to the survey in an email to students, enabling students to start the survey before entering the course. Additionally, we sought to exclude students who signed up for a cohort-based course after the first two weeks, but it turned out that many cohort-based courses still make efforts to sign up students late and our initial tests did not suggest that this was an important moderator, so we removed this restriction. Finally, we sought to exclude students who signed up within a month of when the data were downloaded, but we decided to remove this restriction for Waves 2-4, and instead used their signup date as a covariate in our model.

In addition to relaxing a number of exclusion rules after analyzing Wave 1 data, we added one new exclusion rule. Some courses in Wave 1 had few students interacting with the interventions (due to low enrollment and/or varying placement of the survey link) or very low certification rates (e.g., when grading was handled outside of the EdX platform). To reduce noise in the dataset, we decided to remove these outlier courses. Specifically, we remove courses in Wave 2 that either (i) did not have at least 100 students that were assigned to a condition; or (ii) had fewer than 1% of students with a passing grade in the intervention sample.

Waves 3 & 4 Sample Determination. We used the exclusion criteria from Wave 2 and added an additional criterion. We observed that many students clicked through the survey but did not engage with the questions, especially skipping open-ended responses, including the open-ended activities that are part of the interventions. For example, 14% of students in the focal sample who saw a planning intervention in Year 1 wrote nothing. To address this issue of noncompliance, we added an open-ended question about students' goals for the course before assignment to condition. We use the responses to this question to exclude students who left it unanswered, or who wrote an answer that was shorter than four characters.

Regression Specifications

In the main text, we report results using the regression specification in our final pre-registration. We estimate treatment effects using a linear OLS model with course fixed effects and heteroskedasticity-robust standard errors. We use a standard set of covariates (intent assessment, planned hours per week, prior MOOCs completed, education level, num. days between starting the course and receiving the intervention). In addition, we add pre-registered covariates in Waves 1 & 2 (random assignment strata, HDI) and in Waves 3 & 4 (self-reported likelihood of completing the course). Below we describe all planned regression specifications as pre-registered to explain how our analytic approach evolved across waves.

Wave 1 Regression Specifications. All of our treatment effect estimation is done using a common framework, following our pre-registered specification of a baseline model for student achievement. We conduct mixed effects regressions, using the *lme4* R package (13), and our pre-registered analytic code and output is available at https://osf.io/9bacu/. All models estimate two separate random effects terms, to identify variation at the course level, and at the level of the strata defined during the stratified random assignment to condition. Additionally, the model includes fixed effects estimates of the four covariates used to create the randomization strata (as continuous variables). Formally, our baseline model of student achievement can be noted as:

$$\boldsymbol{y}_i = \boldsymbol{\beta}_{\text{intentions}} + \boldsymbol{\beta}_{\text{hoursPlanned}} + \boldsymbol{\beta}_{\text{education}} + \boldsymbol{\beta}_{\text{previousMoocs}} + \boldsymbol{\zeta}_{\text{strata}} + \boldsymbol{\zeta}_{\text{course}} + \boldsymbol{e}_{\text{i}}$$

In Table S4, we present our baseline model estimates for all seven of the primary outcomes identified in Waves 1 & 2 (the six initial outcomes in Wave 1 and the week one activity outcome added in Wave 2). Across all seven outcomes, the four fixed effects covariates were frequently significant predictors of eventual student success. This confirms that our baseline specification does capture pre-treatment variation in expected student success, and the effects are in line with theoretical expectations. Recovering these benchmark effects provides support for our model specifications, and our measurement approach.

For all our treatment effect estimates, we simply add a variable indicating treatment level to the baseline model.

$$\boldsymbol{y}_i = \boldsymbol{\tau}_{\text{i}} + \boldsymbol{\beta}_{\text{strataCovariates}} + \boldsymbol{\zeta}_{\text{strata}} + \boldsymbol{\zeta}_{\text{course}} + \boldsymbol{e}_{\text{i}}$$

Waves 2-4 Regression Specifications. The models and estimation procedure in Wave 1 suffered from several flaws. The optimization procedure for the (generalized) linear mixed-effects models (GLMMs) was specified such that it ignored integrating out the random effects when fitting the mixed model and used a faster (less reliable) optimization algorithm. This speeds up the estimation procedure significantly, but it is unclear how reliable the standard errors on the fixed-effects coefficients, including the treatment estimates, would be. In fact, using Laplace approximation for integrating out the random effects (the default method), the optimization procedure did not converge while fitting several of the pre-registered models.

Thus, we significantly revised the models and estimation approach for Wave 2. Instead of fitting potentially unreliable GLMMs, we fit generalized linear models (gaussian for continuous outcomes; logistic with logit link for binary outcomes) with fixed effects for strata and course, and compute cluster-robust standard errors to account for course-level clustering in the error term. We estimated these models using the *estimatr* R package (14).

Due to the difference in experimental design, we also adjusted our treatment effect estimation models in waves 3 & 4. For one, we no longer had to control for strata, as the randomization was done at the course level. Second, we added two new measures (days after the course launch when the student first entered the course; and predicted likelihood of completion) to our set of covariates in the baseline regression model. In Table S5, we present estimates for this baseline model using all four of the outcomes identified in the Waves 3 & 4 pre-registered analysis plans.

Subgroup Analyses

Waves 1 & 2 Subgroup Analysis. Our goal was to estimate average treatment effects of different interventions across a single sample population. However, the two interventions in this study were initially conducted separately, using different sample criteria. Although most variations in analysis were coordinated (as described above), there remained one difference in the subgroups for which the interventions were hypothesized to have the largest treatment effects. We decided to preserve these original subgroup specifications, as the one pre-registered difference across analyses of the two interventions.

In our previous plan-making study, the pre-registered analysis plan stated that we would target students who were fluent in written English, so that their plans could be reported naturally. Most non-fluent students skipped the text questions in the survey, or wrote in another language, and in fact they had a somewhat smaller treatment effect than our focal sample. We had also stated a focus on students who intended to complete the course as benefits were expected to be largest for them (3, 14). Although we found no moderation by intentions in our original study, we maintained these conditions as exclusion criteria. The confirmatory analyses for the plan-making intervention in Wave 1 thus focused on the treatment effects within the subpopulation of fluent English speakers who intend to complete all assessments. We again found no moderation by intentions in the Wave 1 data, so for Wave 2 we removed this exclusion criterion and focused on all students within the focal sample who reported English fluency.

In our previous value-relevance study, we focused on treatment effects for an at-risk group of students, those in less developed countries (HDI < 0.7; in (5)). The intervention was not expected to benefit students not at-risk, those in more developed countries (HDI > 0.7). We fit a model with first- and second-order effects of the value-relevance condition and student subgroup (e.g., with *VR*: 1=yes, 0=no; *highHDI*: 1=yes, 0=no; and a *VR * highHDI* interaction). This enabled us to identify the treatment effect for at-risk students (coefficient on *VR*), the achievement gap between at-risk and not at-risk students (coefficient on *highHDI*), and how the treatment effect differed for not at-risk students (coefficient on the *VR-by-highHDI* interaction). The primary analysis focused on the *VR* main effect, representing the effect of the value-relevance intervention on low-HDI students.

Waves 3 & 4 Subgroup Analysis. For each intervention, we hypothesized that a specific subgroup (within the sample defined by our exclusion criteria above) would be most responsive to treatment, based on the relevant theory of behavior change and prior evidence.
For the value-relevance intervention, we define a new course-level subgroup in Waves 3 and 4 based on the magnitude of the global achievement in the course (in the control condition): the standardized difference in certification rates between students in higher HDI countries (HDI > 0.7) relative to lower HDI countries (HDI < 0.7). In courses with a global gap of 0.2 or more standard deviation units, we hypothesized an effect of the value-relevance intervention among students in less developed countries (HDI < 0.7).

For the plan-making intervention in Wave 3, we used our previously defined subgroup of “students who state they are fluent in English”, which was successful in eliminating many of those students who skipped the open-ended questions in the interventions. However, we had already implemented a more effective strategy for excluding these students in Wave 3, namely by using an open-ended question before the intervention to identify students who tend not to answer such questions. Accordingly, for Wave 4 we dropped the English-fluency requirement, and report results for the entire focal sample defined by the common set of exclusion criteria, with no intervention-specific subgroup.

For the MCII intervention, we follow previous results that found its effect was strongest on students in individualist cultures defined by Hofstede's country-level index of individualism (countries with individualism scores > 0.66 are classified as individualist cultures) (11). The social accountability intervention was developed in response to prior evidence that found the MCII intervention to be ineffective in more collectivist cultures. Accordingly, we test the effect of the social accountability intervention in countries where the individualism score is at most 0.66.

Forecasting Global Achievement Gaps

We build a random forest classifier to predict the occurrence of a global gap (defined as whether the estimated global gap is above 0.2 standard deviation units) in a course. We use as input 21 features of the course and the student population in this course. A full list of the features is below. This model is trained on 153 courses in the first year and was evaluated on 94 courses in the second year. The depth and number of trees in the random forest model were optimized by 5-fold cross validation using grid search (over depth {3, 5, 7, 15} and number of trees {50, 100, 200, 250, 300}) on courses in the first year. To account for the randomization in the training procedure of the prediction model, we built 11 models and report the median accuracy on courses in year 2. This accuracy is 0.54. For comparison, the accuracy of chance predictions is 0.50. These chance predictions ignore all course features and predict the occurrence of a global gap by tossing a coin with bias of the proportion of courses that had a global gap in year 1. This indicates that even a model that can leverage complex non-linear patterns such as random forests are not able to use these course features to predict the occurrence of a global gap significantly more accurately than chance.

Features used by the predictive models:

- Institution
- Wave
- Indicator for whether course was self-paced
- Indicator for whether this is an introductory level course
- Indicator for whether the course has prerequisites
- Course Discipline Coding (topic_4_way),
- Hours per week,
- Course length in weeks
- Indicator for whether course instructor is female
- Indicator for whether course instructor is not white
- Total number of students in the course
- Average intended hours of the students in the course
- Standard deviation of the students' intended hours
- Average education level of the students in the course
- Standard deviation of the students' education level
- Average amount of assessments students intend to complete in the course
- Standard deviation of the students' intended amount of assessments to complete in the course
- Average number of courses finished by the students prior to this course
- Proportion of female students
- Average fluency level of students
- Proportion of students from the US

Adaptive Policy Learning

Instead of intervening on a per-course basis, it may be necessary to deliver interventions on a per-student basis. To explore this, we investigated personalized policies that select an intervention (or none) for a given student based on a set of student features. The goal is to improve overall completion rates. We used current machine learning approaches to train personalized policies on data from students in courses in the first year, and evaluated these personalized policies by estimating the completion rate for each model had we applied it to students in the second year. We here only consider the planning and value-relevance interventions which were administered in both years.
Specifically, we used three approaches to personalized policy learning that have complementary strengths and weaknesses.

Approach 1 (Direct Method): Bias-corrected Reward Imputation. Given as input a student's features and the condition assigned to the student, this approach first learns a predictive regression model of the probability the student will complete the course. Due to the categorical and heterogeneous nature of the student features (see list below), we chose random forests as our regression model. We follow the same procedure as for the global gap forecasting model for fitting hyperparameters (the number of trees and their depth). We compute an adaptive personalized policy by applying a softmax function with temperature parameter T=0.03 to the predicted completion rate for each condition. To reduce modelling bias in this approach, we use this decision policy to compute importance weights for each student in year 1 and re-trained the completion rate regression model using the weighted samples (as suggested by (15)). We repeated this process for 30 iterations to determine a final policy.

Approach 2 (Importance Sampling): Counterfactual risk minimization using importance weighting. In this approach, we do not learn a predictive model for the completion rate but instead specify a parametric model for the adaptive policy directly. Specifically, we use a generalized linear model on the student features *x* and separate parameters for each condition (with softmax as the link function) to model the probability (a | x) of choosing that condition *a* for the student:

$$\pi(a|x) = \frac{\exp(\theta_a^\top x)}{\sum_{a' \in \{\text{control, plan, vr}\}} \exp(\theta_{a'}^\top x)}$$

This model was trained to maximize the structural objective proposed by (16)

$$\hat{R}_{IS}(\pi) - \lambda \sqrt{\frac{\hat{\mathrm{Var}}(\pi)}{n}}$$

where the first term

$$\hat{R}_{IS}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(a_i|s_i)}{\tau(a_i)} y_i$$

is the importance weighting estimator of the average completion rate had we deployed policy π. Here *n* is the number of considered students in the first year, and x_i, a_i, y_i their feature description, assigned condition in the randomized trial and indicator of completion respectively. (a) is the probability of condition *a* in the randomized experiment. The second term penalizes policies that induce high variance, estimated as

$$\hat{\mathrm{Var}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\pi(a_i|x_i)}{\tau(a_i)} y_i - \hat{R}_{IS}(\pi) \right)^2$$

and acts as a data-dependent regularization term to prevent overfitting. The regularization coefficient ƛ is chosen by 5 cross-validation on the training set (year 1 students) from {0.01, 0.05, 0.1, 0.5, 1., 5.} and set to 0.05. We trained this model (which involves finding the best weights theta to minimize the loss function above) using the L-BFGS optimizer.

Approach 3 (Doubly Robust): Policy Optimization using a doubly-robust estimator. The personalized policy construction in Approach 2 uses an importance-sampling based estimators, which typically have low bias but can suffer from high variance. In contrast model-based estimators such as the one used in Approach 1 typically have low variance but may have high bias. A popular approach to mitigate the potential issues of the two types of estimators is to combine them in a doubly-robust estimator (see for example (17)). To that end, we used the predictive random forest model R_{RF} in a doubly robust estimator defined as

$$\hat{R}_{DR}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_i|x_i)}{\tau(a_i)}\left(y_i - \hat{R}_{RF}(a_i, x_i)\right) + \frac{1}{n}\sum_{i=1}^{n}\pi(a|x_i)\hat{R}_{RF}(a, x_i)$$

As a third approach, we learned the same policy as in approach 2 but replaced R_{IS} by R_{DR} in the structural risk objective.

Evaluating personalized policies: Estimating the completion rate. We estimate the completion rate of students in year 2 for each personalized policy optimization approach and baseline using the unbiased importance weighted average given by

$$\frac{1}{m}\sum_{i=1}^{m}\frac{w_i}{v_i}y_i$$

where for all m students in the second year, w_i is the probability that the evaluated policy assigns the ith student to the same condition as they were actually assigned, v_i is the probability under the randomized trial of being assigned to the condition and y_i is the indicator whether they completed the course. Since this estimator and its estimated standard error ignores all covariates, we in addition estimate the completion rate using a parametric approach.

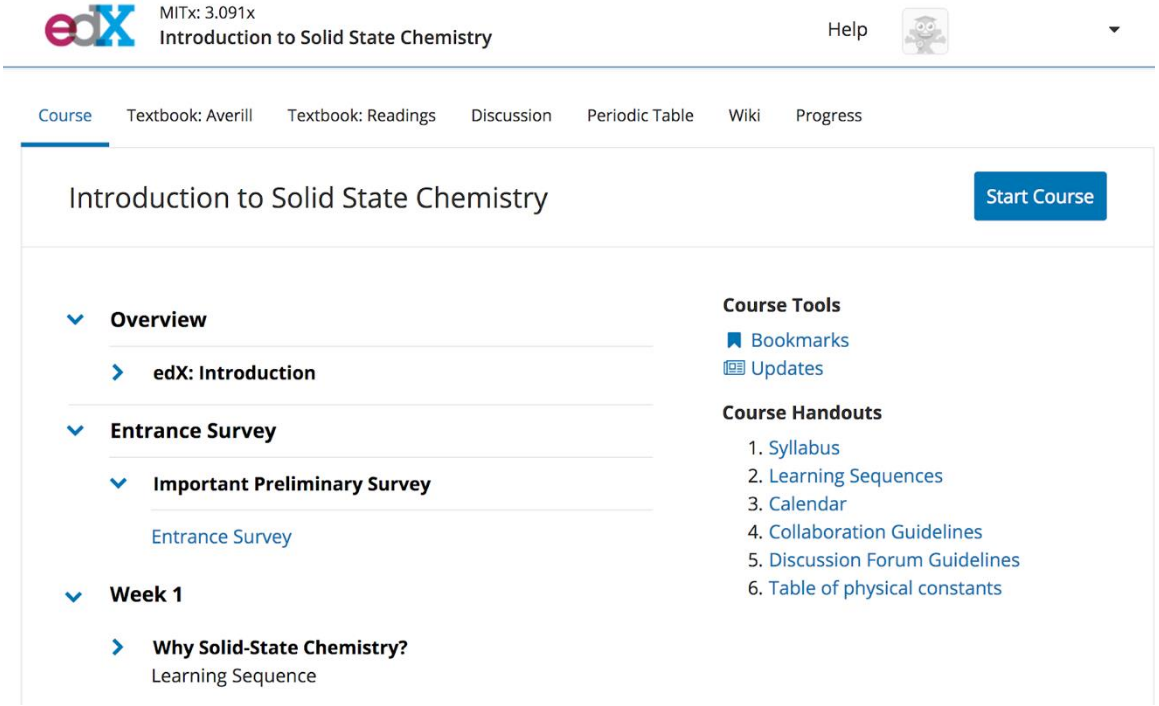
More precisely, we estimate the pre-registered Year 2 baseline model for predicting completion using student-level covariates and course fixed effects (selecting the course with the highest number of students as the reference group), and course-cluster-robust standard errors. By varying the sample weights for different policies, we estimate the regression intercept to indicate the average course completion rate in the reference course in Year 2 data. The sample weights for the ‘No intervention’ policy take 1 if a student’s actual assignment was to the control condition, and 0 otherwise (equivalent for planning and value-relevance). The sample weights for the ‘Random intervention’ policy are ⅓ for students actually assigned to control, value-relevance, or planning, and 0 otherwise. For the optimized policies, sample weights derived from each policy optimization approach are used.

We compare the personalized policies computed by the three approaches above with assigning all students to a single condition (control, planning or vr) or to one of the three conditions uniformly at random. The estimated completion rates are listed in Table S6.

These features were used by the predictive random forest models (Approach 1 + 3): Course Name, Institution, Wave; Indicator for whether course was self-paced; Indicator for whether this is an introductory level course; Indicator for whether the course has prerequisites; Topic_4_way, hours_per_week; Course length in weeks; Indicator for whether course instructor is female; Indicator for whether course instructor is not white; Indicator for whether student is for a country with high HDI; Intent_assess; Hours; Education; Number of courses finished; Days from start; Indicator for whether student is fluent in English; Indicator for whether student identifies as female; Country Name.

These features were used by the softmax-linear policy class (Approach 2 + 3): Wave; Institution (1 hot encoding); Indicator for whether course was self-paced; Indicator for whether this is an introductory level course; Indicator for whether the course has prerequisites; Topic_4_way (1 hot encoding); hours_per_week; Course length in weeks; Indicator for whether course instructor is female; Indicator for whether course instructor is not white; Indicator for whether student is for a country with high HDI; Intent_assess; Hours; Education; Number of courses finished; Days from start; Indicator for whether student is fluent in English, Constant feature 1.

Figure S1: Screenshot showing an example of how the course survey was embedded at the beginning of the course materials. The Section is named “Entrance Survey” (bold black font), the Sub-Section is named “Important Preliminary Survey” (bold black font), and the hyperlink to the page with the actual survey is named “Entrance Survey” (light blue font).



Introduction to Solid State Chemistry

Overview

edX: Introduction

Entrance Survey

Important Preliminary Survey

Entrance Survey

Week 1

Why Solid-State Chemistry?
Learning Sequence

Course Tools

Bookmarks
Updates

Course Handouts

1. Syllabus
2. Learning Sequences
3. Calendar
4. Collaboration Guidelines
5. Discussion Forum Guidelines
6. Table of physical constants

Figure S2: Daily activities logged by students in the first three weeks after their exposure to the interventions. Each point represents a 99%-winsorized average with 95% confidence interval across all 269,169 students in the focal samples, calculated separately by day, by intervention year, and by students' expressed intentions to complete the course assignments. Our pre-registered short-term outcomes focused on the log-transformed sum of activity counts across three consecutive week-long spans, starting on day two.

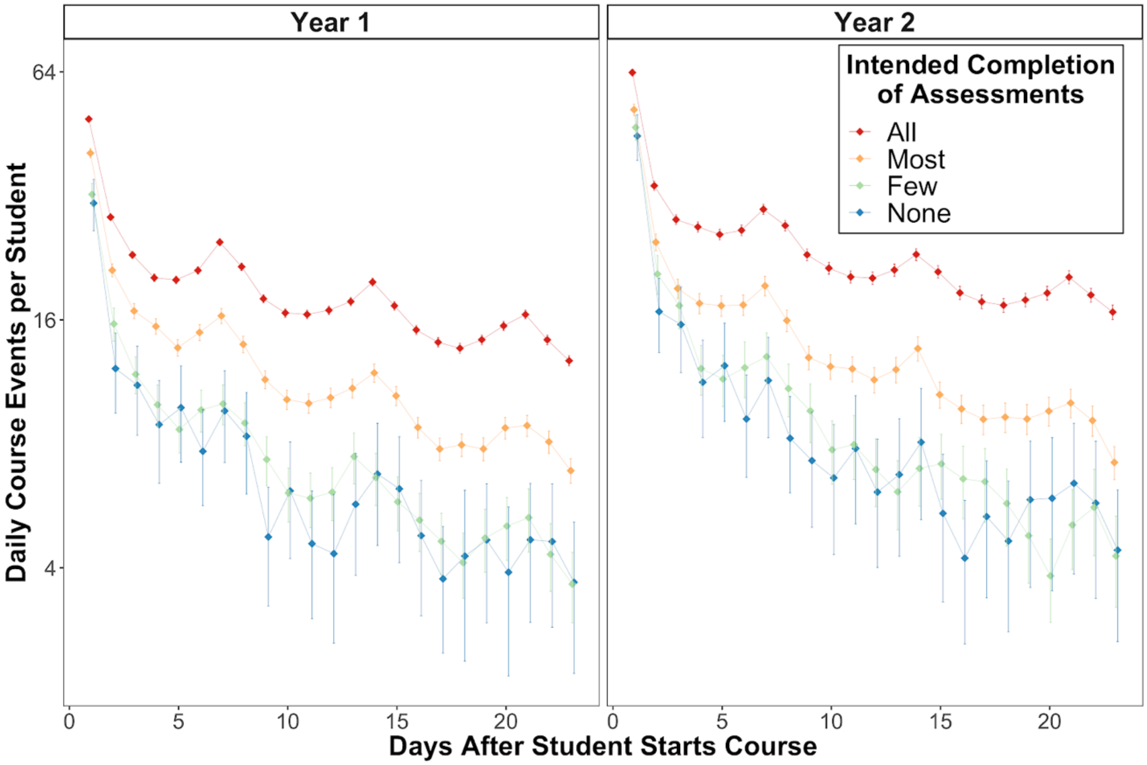


Table S1: Overview of the number of courses and students by study wave and institution at different milestones. Focal Courses are all courses with the intervention that had at least 100 registrants and at least a 1% completion rate among students assigned to treatment. Total Enrolled includes all students who ever registered for the course, and the enrollment in the smallest and largest course. Completed Course is the number of students earning a passing grade, and the lowest and highest course completion rate among students in the focal sample. Entered Survey is the number of students who logged at least one data interaction with the survey. Assigned Treatment includes all students who persisted in the survey until random assignment to treatment condition. The Focal Sample includes students who meet our exclusion criteria as defined in the *Participants* section above.

	Focal Courses	**Total Enrolled [min; max]**	**Completed Course [compl. rate min; max]**	**Entered Survey**	**Assigned Treatment**	**Focal Sample**
Wave 1						
Harvard	7	132,303 [4,286; 61,242]	5,375 [2.1%; 45.3%]	21,695	15,906	13,554
MIT	54	1,060,308 [25,29; 113,228]	25,424 [1.4%; 49.5%]	119,574	83,601	73,725
Stanford	11	118,738 [1,645; 45,277]	5,581 [2.3%; 38.0%]	9,690	9,592	9,269
All	72	1,311,349 [1,645; 113,228]	36,380 [1.4%; 49.5%]	150,959	109,099	96,548
Wave 2						
Harvard	33	695,729 [4,037; 215,456]	20,469 [2.1%; 39.2%]	103,086	74,752	69,075
MIT	39	539,513 [2,046; 75,207]	9,816 [3.3%; 40.8%]	57,806	33,945	27,981
Stanford	9	59,162 [722; 17,860]	3,292 [7.3%; 54.5%]	6,634	6,618	5,913
All	81	1,294,404 [722; 215,456]	33,577 [2.1%; 54.5%]	167,526	115,315	102,969
Wave 3						
Harvard	28	345,490 [808; 54,557]	8,923 [0.7%; 54.5%]	51,856	37,883	31,557
MIT	15	130,275 [1,750; 20,062]	3,331 [2.5%; 49.5%]	14,914	11,844	8,917
Stanford	3	11,392 [3,230; 4,399]	2,188 [7.5%; 48.4%]	1,782	1,726	1,459
All	46	487,157 [808; 54,557]	14,442 [0.7%; 54.5%]	68,552	51,453	41,933
Wave 4						
Harvard	21	248,090 [2,336; 59,864]	8,866 [1.1%; 22.0%]	17,310	12,990	10,792
MIT	22	174,734 [1,320; 40,928]	3,651 [1.1%; 57.0%]	21,290	17,036	11,801
Stanford	5	30,472 [2,160; 9,774]	2,281 [6.2%; 41.9%]	6,527	6,328	5,126
All	48	453,296 [1,320; 59,864]	14,798 [1.1%; 57.0%]	45,127	36,354	27,719
All Waves						
Wave 1&2	153	2,605,753 [722; 215,456]	69,957 [1.4%; 54.5%]	318,485	224,414	199,517
Wave 3&4	94	940,453 [808; 59,864]	29,240 [0.7% ; 57.0%]	113,679	87,807	69,652
All	247	3,546,206 [722; 215,456]	99,197 [1.4%; 57.0%]	432,164	312,221	269,169

Table S2: Descriptive statistics of student demographic characteristics and course performance in the focal samples for each study wave. Showing percentages for binary variables and means (standard deviations) for continuous variables.

	Wave 1	Wave 2	Wave 3	Wave 4
% Course Completion	15.2%	13.4%	13.8%	12.0%
% Enrolled Verified	6.0%	7.7%	5.7%	7.5%
% Verified Upgrade	6.4%	4.5%	8.2%	9.5%
% Verified Certificate	8.2%	7.8%	8.8%	8.2%
% Active in First Week	43.1%	33.3%	35.3%	35.8%
% Active in Second Week	33.5%	22.8%	23.5%	26.9%
Age	33.5 (12.0)	33.0 (12.8)	35.2 (14.1)	32.4 (12.9)
% Female	28.1%	46.0%	40.3%	31.5%
% Low HDI	20.1%	25.3%	25.1%	27.0%
% English Fluency	55.2%	54.2%	55.4%	49.7%
Past MOOCs Completed	2.18 (3.15)	1.64 (2.69)	1.95 (3.74)	1.94 (3.46)
% High School Degree	97.3%	96.8%	98.1%	97.8%
% Bachelor's Degree	73.6%	71.6%	77.5%	77.7%

Table S3: Comparison of population characteristics between students who start the course survey and students included in the focal sample for each pair of waves. The population of students who entered the survey provides a comparison group of engaged students, 62.6% and 61.3% of whom are in the focal sample in Waves 1 & 2 and Waves 3 & 4, respectively. Differences between students who entered the survey and those in the focal sample are minimal in terms of the characteristics shown in the table.

	Wave 1 & 2 Entered Survey	Wave 1 & 2 Focal Sample	Wave 3 & 4 Entered Survey	Wave 3 & 4 Focal Sample
N (Courses)	318,485 (153)	199,517 (153)	113,679 (95)	69,652 (94)
% Course Completion	13.7%	14.3%	11.9%	13.1%
HDI	0.807 (0.127)	0.809 (0.126)	0.802 (0.134)	0.800 (0.135)
% Low HDI Country	23.1%	22.7%	26.0%	25.8%
% Individualist Country	-	-	50.5%	50.2%
% Fluent	54.96%	54.7%	52.8%	53.1%
% Committed	68.24%	68.4%	64.5%	66.9%
Hours Per Week	6.65 (5.91)	6.52 (5.90)	6.26 (5.93)	6.29 (5.88)
Past MOOCs Completed	2.06 (3.07)	1.90 (2.94)	1.96 (3.49)	1.94 (3.63)
% Female	36.0%	37.3%	35.6%	36.8%
Age	33.4 (12.6)	33.2 (12.4)	33.8 (13.6)	34.0 (13.7)

Table S4: Estimates for the baseline model of student outcomes in Waves 1 & 2. The same five regression terms are included in every regression model as fixed-effects strata. For this table, all linear predictors were first transformed into standardized units for comparison (standard errors in parentheses). The coefficients on HDI scores are compared to a reference level of "very high". All coefficient estimates are significant at the P < .002 level.

	(1)	**(2)**	**(3)**	**(4)**
Outcome	Course Completion	Final Grade	Predicted Completion	First Week Activity
Assessment Intentions	.0420 (.0009)	.0396 (.0008)	6.644 (.079)	.2326 (.0077)
Planned Hrs/Week	.0055 (.0010)	.0057 (.0008)	1.535 (.063)	.0866 (.0078)
Highest Formal Degree Earned	.0306 (.0013)	.0290 (.0011)	-0.696 (.089)	.1622 (.0097)
First Day After Course Launch	-.0246 (.0007)	-.0343 (.0006)	-0.573 (.051)	-.1236 (.0068)
Previous MOOCs Completed	.0112 (.0017)	.0116 (.0014)	0.687 (.103)	.0356 (.0115)
High HDI	-.0159 (.0020)	-.0180 (.0017)	3.126 (.125)	-.0538 (.0146)
Medium HDI	-.0466 (.0021)	-.0480 (.0017)	4.442 (.153)	-.3640 (.0167)
Low HDI	-.0407 (.0026)	-.0455 (.0022)	2.185 (.180)	-.1738 (.0215)
Strata FE	YES	YES	YES	YES
Course FE	YES	YES	YES	YES
Pseudo R^2	.101	.156	.211	.117
# of Students	199,517	199,517	121,703	199,517
# of Courses	153	153	153	153

Table S5: Estimates for the baseline model of student outcomes in Waves 3 & 4. The same five regression terms are included in every regression model. For this table, all linear predictors were first translated into standardized units for comparison (standard errors in parentheses). All coefficient estimates are significant at the P < .001 level.

	(1)	**(2)**	**(3)**	**(4)**
Outcome	Course Completion	Final Grade	First Week Activity	Second Week Activity
Assessment Intentions	.0205 (.0011)	.0191 (.0009)	.1560 (.0102)	.1364 (.0095)
Planned Hrs/Week	.0169 (.0014)	.0145 (.0012)	.1734 (.0107)	.1005 (.0099)
Highest Formal Degree Earned	.0146 (.0012)	.0129 (.0010)	.1342 (.0095)	.1542 (.0091)
First Day After Course Launch	-.0158 (.0012)	-.0180 (.0010)	-.0994 (.0108)	-.1871 (.0099)
Predicted Complete	.0303 (.0012)	.0259 (.0010)	.1914 (.0105)	.2128 (.0010)
Previous MOOCs Completed	.0150 (.0013)	.0133 (.0011)	.1141 (.0095)	.1092 (.0095)
Course FE	YES	YES	YES	YES
Pseudo R^2	.118	.183	.186	.187
# of Students	69,652	69,652	69,652	69,652
# of Courses	94	94	94	94

Table S6: Estimated course completion rate of students in Waves 3 & 4 under different intervention policies with estimated standard errors in parentheses.

	Course completion rate estimated by importance weighted average	**Course completion rate** estimated by covariate-adjusted model with course fixed effects (largest course is baseline)
No intervention	12.81% (SE = 0.30%)	26.56% (SE = 0.136%)
Planning intervention	13.28% (SE = 0.30%)	28.53% (SE = 0.141%)
Value-relevance intervention	13.08% (SE = 0.30%)	27.80% (SE = 0.143%)
Random intervention	13.06% (SE = 0.16%)	27.62% (SE = 0.118%)
Optimized policy 1 DM	13.23% (SE = 0.24%)	27.22% (SE = 0.129%)
Optimized policy 2 IS	13.28% (SE = 0.30%)	28.59% (SE = 0.163%)
Optimized policy 3 DR	13.38% (SE = 0.30%)	27.05% (SE = 0.153%)

Intervention Materials

Value-Relevance Intervention Instructions (Waves 1 & 2)

This short activity is designed to help you succeed in the course. Below is a list of characteristics and values, some of which may be important to you, and some of which may be unimportant. Please select the 2 or 3 values that are most important to you.

□ Artistic skills/aesthetic appreciation
□ Sense of humor
□ Relationships with family or friends
□ Spontaneity/living in the moment
□ Learning for the sake of learning
□ Religious/spiritual values
□ Sports and athletics
□ Musical ability/appreciation
□ Physical attractiveness
□ Creativity
□ Business/managerial skills
□ Romantic values

------------------ NEXT PAGE ------------------

Now consider the 2 or 3 values that are most important to you: [piped in from previous page]

How does taking this course reflect and reinforce your most important values?

Please write at least a paragraph. Focus on your thoughts and feelings, and don’t worry about spelling, grammar, or how well written it is.

____
____
____

------------------ NEXT PAGE ------------------

Previous students really appreciated writing something to remind themselves about how their most important values are reinforced by taking this online course.
To give you the chance to remind yourself, we would like you to write a note to your future self about your experience and what you've learned so far. Write about how you can gain strength from the fact that taking this course reinforces your most important values.

We know it can be difficult to write that way, but we believe it will be particularly meaningful for you if you write as though your present self is speaking directly to your future self.

____
____
____

If you like, you can save your writing on your computer or send it as an email to be reminded and motivated over the coming weeks.

Value-Relevance Intervention Instructions (Waves 3 & 4)

As you're starting this course, take a moment to reflect.

What is most important to you? Select one or more below.

□ Relationships with family or friends
□ Compassion and kindness
□ Gaining broad skills and knowledge
□ Spontaneity/living in the moment
□ Meeting people from diverse backgrounds
□ Health and well-being
□ Contributing to society
□ Personal and intellectual growth
□ Creativity and artistic expression
□ Religion and spirituality
□ Athletics and sports
□ Something else: ____________________

------------------- NEXT PAGE -------------------

How does taking this course reflect and serve what's most important to you?

Tip: Write at least a paragraph focusing on your thoughts and feelings, don't worry about how well written it is.

____
____
____

Plan-Making Intervention Instructions (Waves 1 & 2)

Note about instructions: The long-term and short-term plan-making interventions were similar, and all text that differs between them is italicized in brackets below.
Please write down a clear, concrete plan to follow through on your goals in [*the first week of*] the course. Plan-making can be a helpful tool in MOOCs! Successful students in previous courses have made detailed plans for how they will engage [*in the first week of / throughout*] the course.

In the text box below, write out your plans to complete tasks for the course [*this upcoming week*]. Please be as specific as you can!

____
____
____

You might find it helpful to consider these questions when you make your plans:
- When and where do you plan to engage with the course content?
- How much time will you spend studying in the [*first week / course*]?
- What will you do to ensure you complete the required course work?
- How will you overcome potential obstacles in the [*first week / course*]?

Here are some examples to inspire your plan-making (replace them with your own):
“I will watch videos Wednesday night after work, and complete the readings on Saturday morning.”
“If I haven't done the week’s work by Sunday, then I will prioritize the videos to stay on schedule.”
“I will add these times to my calendar so that I don’t forget.”
“If I have trouble understanding the material, I will visit the class discussion forum.”

------------------- NEXT PAGE -------------------

It's great that you have written down your plans. They will be a useful tool for overcoming difficulties and achieving your goals.

Take another look at your plans below. How will you make sure to remember them? For example, take a moment now to: write them down on paper, email them to yourself or a friend, add to a calendar with a reminder, or tell someone about them!

YOUR PLANS FOR THIS WEEK
[response text piped in from previous page]

Plan-Making intervention Instructions (Waves 3 & 4)

We want everyone who signs up to meet their goals in this course. However, while many students who intend to finish the course will complete it, there are others who do not finish as much of the course as they had wanted. We'd like to know your thoughts about why some people do not follow through on their intentions.
Do you think there are some common reasons that explain why some students do not achieve the goals they set for themselves? Are there reasons you might not meet your own goals in this course?
Use the boxes below to describe some of these reasons. (Note: You don't have to fill every box; just use the different boxes to separate distinct reasons).

Reason #1 (1) ____
Reason #2 (2) ____
Reason #3 (3) ____

------------------- NEXT PAGE -------------------

Please write down a clear, concrete plan to follow through on your goals in the course. Plan-making can be a helpful tool in MOOCs! Successful students in previous courses have made detailed plans for how they will engage throughout the course. In the text boxes below, write out your plans to complete your work for the course.
Please be as specific as you can! Write clearly, in full sentences, so that someone else could understand what you mean.

When and where do you plan to engage with the course content?
____
____
____
What specific steps will you take to ensure you complete the required course work?
____
____
____
How will you overcome potential obstacles in the course?
____
____
____

------------------- NEXT PAGE -------------------

Thank you for writing down your plans. Sticking to your plans can help you stay on track and achieve your goals in the course!

Take a moment now to read over your plans below, to make sure you remember them later. For example: write them down on paper, email them to yourself or a friend, add to a calendar with a reminder, or tell someone about them!

YOUR PLANS FOR THIS COURSE
[response text piped in from previous page]

MCII Intervention Instructions

This brief activity is designed to help you achieve your course goal.
Here is the goal for taking the course you specified above. Keep it as is or refine it now:
_[response text piped in from open-ended question on course goals]__________
____
____
The next steps will help you stay committed and plan ahead for how to overcome obstacles that you can anticipate.

------------------- NEXT PAGE -------------------

1. Write down the positive outcomes you expect from achieving this goal:
Tip: After you write them down, think about how it would feel for them to come true.
____
____
____

------------------- NEXT PAGE -------------------

2. Write down one or more obstacles that could get in the way:
Tip: Be specific and focus on predictable obstacles that can occur more than once. Think about how they could interfere.
____
____
____

------------------- NEXT PAGE -------------------

3. For each potential obstacle you identified above, write an if-then plan for how to address it.
Use this format for your plans:
"If *[obstacle]* then I will *[plan to overcome obstacle]*."
____
____
____

Social Accountability Intervention Instructions

Did you know that it can be much harder to stay engaged in an online course than in an in-person class? This is partly because nobody is holding you accountable for making progress towards your goal.

Now is the best time to think about who can hold you accountable.

1. Write down the names of one or more friends, co-workers, family members, or acquaintances who could hold you accountable.
Tip: Pick people who you don't see too often but whose opinion matters to you.

____

____

____

------------------- NEXT PAGE -------------------

2. Now plan for what you are going to tell them about the course and your goal.
Tip: Ask them to regularly check in with you about your progress in the course.

____

____

____

------------------- NEXT PAGE -------------------

3. Finally, write down how and when you will tell them about this.
For example, will you talk in person or on the phone, or send them an email or text message? Be sure to choose a time and place that works.

____

____

____

Supplementary References

1. Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity- score matched samples. *Statistics in medicine, 28*(25), 3083-3107.
2. Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50*(2), 179-211.
3. Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science, 313(5791),* 1307-1310.
4. Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science, 326(5958),* 1410-1412.
5. Kizilcec, R. F., Saltarelli, A.J., Reich, J. & Cohen, G.L. (2017). Closing Global Achievement Gaps in MOOCs. *Science, 355(6322)*, 251-252.
6. Kizilcec, R. F., Davis, G. M., & Cohen, G. L. (2017). Towards equal opportunities in MOOCs: Affirmation reduces gender & social-class achievement gaps in China. In *Proceedings of the Fourth ACM Conference on Learning @ Scale,* 121-130. ACM.
7. Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in experimental social psychology, 38,* 69-119.
8. Rogers, T., Milkman, K. L., John, L. K., & Norton, M. I. (2015). Beyond good intentions: Prompting people to make plans improves follow-through on important tasks. Behavioral Science & Policy, 1(2), 33-41.
9. Koehler, D. J., & Poon, C. S. (2006). Self-predictions overweight strength of current intentions. *Journal of Experimental Social Psychology, 42(4),* 517-524.
10. Yeomans, M. & Reich, J. (2017). Planning to Learn: Planning Prompts Encourage and Forecast Goal Pursuit in Online Education. *Proceedings of the Seventh International Conference on Learning Analytics & Knowledge, 464-473.* ACM.
11. Kizilcec, R. F., & Cohen, G. L. (2017). Eight-minute self-regulation intervention raises educational attainment at scale in individualist but not collectivist cultures. *Proceedings of the National Academy of Sciences, 114(17),* 4348-4353.
12. 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.
13. Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67(1),* 1-48.
14. Blair, G., Cooper, J., Coppock, A., Humphreys, M., & Sonnet, L. (2018). estimatr: Fast Estimators for Design-Based Inference.
15. Wang, L., Bai, Y., Bhalla, A., & Joachims, T. (2019) Batch Learning from Bandit Feedback through Bias Corrected Reward Imputation. *Workshop on Real-World Sequential Decision Making at the International Conference of Machine Learning*
16. Swaminathan, A., & Joachims, T. (2015). Counterfactual risk minimization: Learning from logged bandit feedback. In International Conference on Machine Learning (pp. 814-823).
17. Dudík, Miroslav, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science* 29, no. 4 (2014): 485-511.