Moving Attention but Not Enrollment: Nudging College Information Search at National Scale in Colombia

Eric Bettinger
Stanford University
Michael Kremer
University of Chicago
Felipe Lizarazo
ICFES
Christopher Neilson
Yale University
Christian Posso
EAFIT University
Juan Saavedra
Harvard Kennedy School

Abstract

Students often underuse available information when making high-stakes higher education choices, and a planner rarely knows which option any student should choose—which makes nudging *search* the natural instrument. We randomized informational pop-up messages across 519,000 Colombian students—the full fall national cohort (502,494) and the smaller spring cohort (16,861)—as they logged in to view mandatory exit-exam results, linking each message to a purpose-built portal centralizing earnings, costs, and financial aid. The pop-ups moved attention at scale: 15–25% of treated students clicked through. But message design defied prediction: every belief-targeting variant—earnings levels, earnings ranges, personalized figures, gain or loss frames—underperformed the plain invitation to search by 4–8 percentage points, and loss framing, predicted by prospect theory to dominate, performed worst (13.1% versus 16.7% for gain framing). Enrollment did not move: we rule out average effects above 2.6 percentage points under the conservative inference we report, and above roughly 0.6 points under design-based inference. Students who clicked genuinely searched—nearly half opened the portal's earnings tables or aid pages—yet generic information, even when consulted, did not change what they did. Digital platforms make experimentation cheap; our results show why it remains necessary.

Keywords: Information interventions, higher education, randomized controlled trials, Colombia, digital platforms, scaling
JEL Classification: I23, I24, C93, D83, O15

1 Introduction

Developing countries have expanded access to higher education dramatically, yet many students—particularly from disadvantaged backgrounds—enroll in programs with low economic returns or do not enroll at all (World Bank, 2024). For much of the population this expansion reaches, the college decision is a first in the family: no parent or sibling has navigated program choice, financial aid, or student loans before them. It is among the largest financial decisions of a lifetime, it is made once and early, and—unlike other consequential choices—it cannot be learned through repetition or borrowed from community experience. The stakes of choosing poorly are high: earnings differences across majors are comparable in magnitude to the college wage premium itself, even within the same institution (Altonji, Blom, & Meghir, 2012), students systematically misperceive these returns (Hastings, Neilson, Ramirez, & Zimmerman, 2016; Wiswall & Zafar, 2015), and the potential gains from better-informed choices are large (Hastings, Neilson, & Zimmerman, 2015; Kirkebøen, Leuven, & Mogstad, 2016).

Policy responses to these information failures fall into three families. Demand-side interventions provide information directly or nudge students to search for it (Bettinger, Long, Oreopoulos, & Sanbonmatsu, 2012; Bonilla-Mejía, Bottan, & Ham, 2019; Jensen, 2010). Information infrastructure lowers search costs by measuring, centralizing, and publishing returns, costs, and financial-aid data (Hastings et al., 2016, 2015). And where informed choice fails, supply-side regulation—quality floors, program accreditation, loan-eligibility rules—limits the scope for costly mistakes (Camacho, Messina, & Uribe, 2017). The first two families are attractive because they preserve choice and are cheap to deliver digitally; whether they work at scale is the question this paper addresses experimentally.

There is a deeper reason the first two families take the form they do. In college and career choice, a planner cannot know what any given student *should* do: returns are heterogeneous and students sort on comparative advantage (Kirkebøen et al., 2016), while preferences, aptitudes, and constraints are private information. What the planner can do is supply the public inputs—measured returns, costs, financial aid—and prompt students to combine them with what only they know. This division of labor matters because search itself is underprovided: families search too little, and beliefs about options they have not searched are endogenously biased, distorting the perceived returns to further search (Agte, Allende, Kapor, Neilson, & Ochoa, 2024). Nudging *search*, rather than nudging choices, is therefore the instrument that respects both the planner's ignorance and the student's private information. Which prompt actually gets students to start searching—we observe clicking through to the information portal as the proxy for search initiation—is an empirical question.

Colombia exemplifies both the problem and the response. Despite widespread access to higher education, many students concentrate in programs with poor labor-market prospects (Camacho et al., 2017; Londoño-Vélez, Rodríguez, & Sánchez, 2020). Government agencies published statistics on wages, employment, and financial aid, but the information was fragmented across sources and rarely consulted—a classic search problem despite nominally available data (Stigler, 1961). In response, the Central Bank of Colombia built a portal centralizing returns, employment probabilities, costs, and financial-aid information, and the national testing agency (ICFES)—whose website every high school senior must visit to view mandatory *Saber 11* exit-exam results—delivered randomized pop-up nudges steering students to it. This collaboration between the Central Bank, ICFES, and our research team turned a universal government touchpoint into experimental infrastructure.

Digital government platforms change the economics of addressing this problem, and of learning what works. When an intervention can be delivered through infrastructure the entire population must already use, the marginal cost of reaching an additional student—and of testing an additional design—approaches zero. The binding constraint is no longer delivery cost but uncertainty: even with strong theoretical guidance and evidence from other settings, it is unclear which specific implementation will perform best in a new context. The science of scaling emphasizes precisely this concern: benefit–cost profiles often fail to transport across populations and implementation environments (Al-Ubaydli, List, & Suskind, 2017; DellaVigna & Linos, 2022; Vivalt, 2020), and interventions that succeed in pilots frequently lose "voltage" when deployed through routine systems (Bold, Kimenyi, Mwabu, Ng'ang'a, & Sandefur, 2018; List, 2022). The U.S. college-information literature traces exactly this arc: the documented undermatch of high-achieving, low-income students (C. Hoxby & Avery, 2013) and a celebrated targeted intervention addressing it (C. M. Hoxby & Turner, 2013) were followed by state- and national-scale follow-ups that found precise zeros on enrollment (Bird et al., 2021; Gurantz et al., 2021; Hyman, 2020)—Hyman (2020), like us, finding that message design moved web visits while enrollment stayed flat except among low-income students.

The two experiments proceeded as follows. The first (November 2017) reached 502,494 students—essentially the full national cohort—and tested nine message variants that varied information content, framing, personalization, and statistical presentation. The second experiment (May 2018) reached the 16,861 students of the spring cohort with five refined variants designed using what we learned from the first. Both experiments ran on existing government infrastructure, and outcomes are measured in administrative enrollment records in the years following each exam.

The design tests competing behavioral theories at policy scale. The variants draw on prospect theory, salience, personalization, and information granularity—mechanisms whose implications for message design conflict (Section 2.2), so theory alone could not rank them. The horse race therefore serves a dual purpose: identifying which design works in context, and testing which mechanisms dominate on a government platform in a middle-income country.

Three findings emerge. First, pop-up messages move attention at scale: across both experiments, 15–25% of treated students clicked through to view detailed information about programs, returns, and financial aid. Second, message design defied prediction—theory could not rank the designs ex ante, and neither could we. The plain invitation to search beat every belief-targeting variant—levels, ranges, personalized figures, gain or loss frames—by 4–8 percentage points of engagement. On the clean prospect-theory test, loss-framed messages underperformed gain-framed equivalents ($p < 0.001$), the reverse of the prediction, and personalizing the statistics to the student's own department and score band barely helped. The 2018 round, rebuilt with character counts equalized across content contrasts, confirmed that the penalty is the content, not the length: at fixed length, messages carrying framed returns still lost to neutral text, loss framing again ranked last—and yet the best-performing arm was a *long* message, so even our own lesson from Experiment 1, that simpler is better, failed to fully replicate. This aligns with evidence that behavioral mechanisms have limited external validity across contexts (Meager, 2019; Vivalt, 2020). Third, despite large behavioral responses, effects on tertiary enrollment—each message group compared with the no-message control—are limited: in the 2017 cohort we rule out average enrollment effects larger than 2.6 percentage points under the conservative inference we report, and larger than roughly 0.6 points under design-based inference (Section 3), and the 2018 cohort shows suggestive shifts toward short-cycle technical enrollment (+2.8pp, corrected $p = 0.03$, concentrated among males), which we interpret cautiously in Section 4.4. Nor do the margins a search nudge should move first—program composition, persistence, switching—respond on the subsample we can analyze (Section 4.3). The engagement–enrollment gap echoes recent large-scale findings (Bergman, Denning, & Manoli, 2019; Bird et al., 2021; Hyman, 2020) and suggests that students face multiple binding constraints that information alone cannot address (Bettinger et al., 2012; Dynarski, Libassi, Michelmore, & Owen, 2021).

It is worth stating the counterfactual plainly. Every ingredient argued for deploying this without evaluation—a documented problem, a supportive evidence base, an existing platform, near-zero delivery cost—and standard implementation metrics would have scored it a clear success: hundreds of thousands saw the messages, nearly 80,000 clicked through within two weeks. Only randomization linked to enrollment records revealed that the attention changed nothing downstream, and only the multi-arm design revealed that the mechanism with the strongest prior in the nudge literature—loss aversion—performed worst.

We contribute to three literatures. First, to the science of scaling (Al-Ubaydli et al., 2017; Banerjee et al., 2017; DellaVigna & Linos, 2022; List, 2022; Vivalt, 2020), by treating the choice among intervention variants—and the transportability of benefit–cost profiles—as the object of study in the natural policy environment. Second, to work on information frictions in education in developing countries: earlier studies found substantial effects on beliefs and choices (Bonilla-Mejía et al., 2019; Hastings et al., 2015; Jensen, 2010), recent large-scale implementations find precise zeros on enrollment (Bird et al., 2021; Hyman, 2020), and personalized at-scale designs show promise (Arteaga, Kapor, Neilson, & Zimmerman, 2022; Larroucau, Rios, Fabre, & Neilson, 2025); we help explain this split by separating behavioral transportability failures from engagement–outcome gaps. We add three things to the at-scale design evidence: theory-derived contrasts on an identical message chassis, so each arm isolates one mechanism; individual-level engagement for every arm, so designs are ranked directly at $p < 0.001$ rather than inferred from outcome nulls; and a population where 43% do not enroll, so the extensive-margin null is informative where the U.S. high-achiever nulls were partly mechanical. Third, to digital government platforms as research infrastructure: when implementation is cheap, evaluation can be routine rather than an afterthought.

An Online Appendix collects the full set of supplementary exhibits, verbatim message texts, and additional analyses.

2 Context and Experimental Design

2.1 Higher Education and the ICFES Platform

Colombia's higher education system has grown rapidly: according to Ministry of Education (SNIES) statistics, the gross enrollment rate rose from 37.1% in 2010 to 51.6% in 2020, and the system now comprises roughly 300 institutions—about 70% private—offering several thousand program–institution pairs that vary widely in quality, cost, and returns. The speed of this expansion has raised concerns about the quality of new programs (Camacho et al., 2017). To apply to college, students take a mandatory standardized exit exam, *Saber 11*, administered by ICFES. Unlike the SAT, *Saber 11* is taken by nearly all high school seniors regardless of college plans, and scores are the primary admission criterion at most institutions (Londoño-Vélez et al., 2020). The exam is given twice a year to two graduating cohorts: most public school seniors take it in the fall, most private school seniors in the spring.

Government websites provide statistics on wages, employment, financial aid, and program quality, but the information was fragmented across sources and rarely consulted. Every student, however, must log onto the ICFES website to view exam results. This universal digital touchpoint makes it feasible to deliver randomized informational prompts to the entire national cohort, and to measure downstream outcomes in administrative data.

2.2 Behavioral Mechanisms and Ex-Ante Predictions

The design premise from Section 1 disciplines what the pop-ups try to do: because a planner cannot know which option any given student should choose, the intervention does not recommend choices—it tries to start *search*. The policy question is therefore an attention problem: what should a two-sentence message say to make a seventeen-year-old, midway through retrieving exam scores, decide that searching is worth it? Behavioral economics and the personalization literature offer several distinct answers, each pointing to a different friction, and the experiment was built so that each answer occupies one arm. Every pop-up shares an identical chassis—the same encouragement sentence, portal link, and email option (Figure 1)—and only the middle statistics block varies, so contrasts across arms isolate the attention value of one theoretical ingredient at a time (Online Appendix Table B1 maps each arm to its message content and targeted belief).

The **plain invitation** operationalizes salience and information-gap logic: it asserts qualitatively that higher education is rewarded and offers the link, raising the perceived value of searching while leaving the gap itself open—curiosity does the work (Loewenstein, 1994)—and minimizing cognitive load at a high-pressure moment (Bordalo, Gennaioli, & Shleifer, 2013; Shah, Mullainathan, & Shafir, 2012). The **levels** arms target beliefs about the returns to each education level—university, technological, technical—on the hypothesis that students undervalue postsecondary options; prospect theory motivates their gain-versus-loss variants, which hold the statistics fixed and flip only the frame, so the framing contrast is a clean in-context test of loss aversion (Kahneman & Tversky, 1979; Tversky & Kahneman, 1991). The **range** arms target beliefs about the *heterogeneity* of returns within levels; dispersion is the mechanism closest to the search framing itself, since a wider perceived range of outcomes raises the option value of searching before choosing (Wiswall & Zafar, 2015). The **personalized** arms replace national statistics with figures for the student's own department, benchmarked to graduates in the student's own *Saber 11* score band, on the tailoring hypothesis that generic statistics read as someone else's returns while self-relevant figures are diagnostic of one's own (Kreuter, Farrell, Olevitch, & Brennan, 2000; Lavecchia, Liu, & Oreopoulos, 2016).

Ex ante, these mechanisms disagree about the best design: prospect theory, personalization, and granularity all recommend enriching the message with targeted content, while salience, cognitive load, and the information-gap logic recommend the spare invitation. The factorial structure turns the disagreements into estimable contrasts—framing is identified holding statistics and personalization fixed, personalization holding framing and presentation fixed, and the plain-versus-any-information comparison pits the enrichment theories jointly against the attention theories. One bundle the 2017 design does not break is that adding content adds message length; unbundling the two is precisely what Experiment 2 was built to do (Section 2.4). Theory could not settle ex ante which force dominates; the scaling question is which mechanisms retain their force when implemented through a real government platform for heterogeneous student populations (Online Appendix A expands on each mechanism, and Online Appendix Table A1 confronts each prediction with the realized ranking).[1]

2.3 Experiment 1 (November 2017): Broad Exploration at National Scale

The first experiment covered all 502,494 students in the fall 2017 cohort, who sat *Saber 11* on 27 August 2017 and logged onto the ICFES website to view their scores when results were published on 11 November 2017. Upon login, students were randomly assigned—student by student, at page entry—to a control group (roughly one in ten, which received no message) or to one of nine treatment arms; a student re-entering the page always saw the same message. Treated students saw a pop-up combining a brief encouragement about the value of higher education with the randomized message content described below, and offering a link (and an email option) to a portal built with Colombia's Central Bank. The portal centralized average earnings and the 10th–90th percentile earnings range by degree level (technical, technological, university) for the nation and each department, formal-employment rates, national and municipal scholarship and student-loan programs, and the Ministry of Education's career-orientation tools. (Online Appendix B reproduces treatment pop-ups, the verified message texts, and captures of the destination portal itself.)

The nine arms (Figure 1; Online Appendix Table B1) varied four dimensions: **information content** (a plain prompt—labeled "salient" in our exhibits and "Genérico" in the implementation file—carrying the encouragement and link but no statistics, versus eight variants quantifying returns to college); **framing** (gains from college versus foregone earnings without it); **personalization** (nationwide returns versus returns for the student's

[1] The experiments were approved by the University of Southern California IRB (protocol UP-17-00657, exempt determination) and were not registered in a trial registry; the ex-ante hypotheses summarized in Online Appendix Table A1 are documented in the team's design correspondence of October–November 2017. We flag this so the prediction-versus-performance comparison is read with appropriate caution.

own department, benchmarked to graduates in the student's own *Saber 11* score band); and **statistical presentation** (average returns versus the 90th–10th percentile range). Each pop-up contained at most two sentences beyond the basic encouragement, and randomization was implemented on the platform at login, with roughly 50,000 students per group (56,000 in the control and salient groups, 48,000–49,000 in each information arm) and administrative follow-up for all of them. The implementation itself illustrates what government digital infrastructure makes possible: six weeks elapsed between the technical proposal and national deployment.

Figure 1: What students saw, and what was randomized inside it

1 **Encouragement banner — identical in every treatment arm**
Higher education deepens knowledge in your preferred area, contributes to the country, and is rewarded by the labor market. Present in all arms, so it is differenced out of every contrast.

2 **¿Sabías que? — the randomized statistical content**
The only block that varies. Every quantitative message covered all three degree levels: universitaria, tecnológica, técnica profesional.

3 **Link invitation — identical across treatment arms**
A single button to the Banco de la República portal, carrying a per-student identifier. This is how click-through was measured; the control saw no link, so its click rate is zero by construction.

4 **Email capture — identical across treatment arms**
An optional field to receive the information by email. Source of the second engagement outcome, clicked and provided email.

BLOCK 2, RANDOMIZED · 2017
Arm 0 saw no pop-up; arm 1 the plain invitation, with no statistics.

STATISTICS × PRESENTATION	GAIN FRAME	LOSS FRAME
Nationwide · average	Arm 2	Arm 3
Nationwide · range	Arm 4	Arm 5
Own dept. · average	Arm 6	Arm 7
Own dept. · range	Arm 8	Arm 9

Notes: The capture is the live ICFES results portal on 14 November 2017 (*resultadossaber-icfes2017a.icfes.gov.co*); the arm shown is nationwide × loss × range. The pop-up fired at login, before the student reached their scores. Three of its four blocks were identical in every treatment arm; only the second varied, and in the plain-invitation arm it was left empty. The student's email address has been painted out of the capture. Arm numbers are the *Tipo* field in the ICFES implementation file (`message_catalog_2017.csv`); the verbatim Spanish, the English translations and the full nine-arm listing are in Online Appendix B and Table B1.

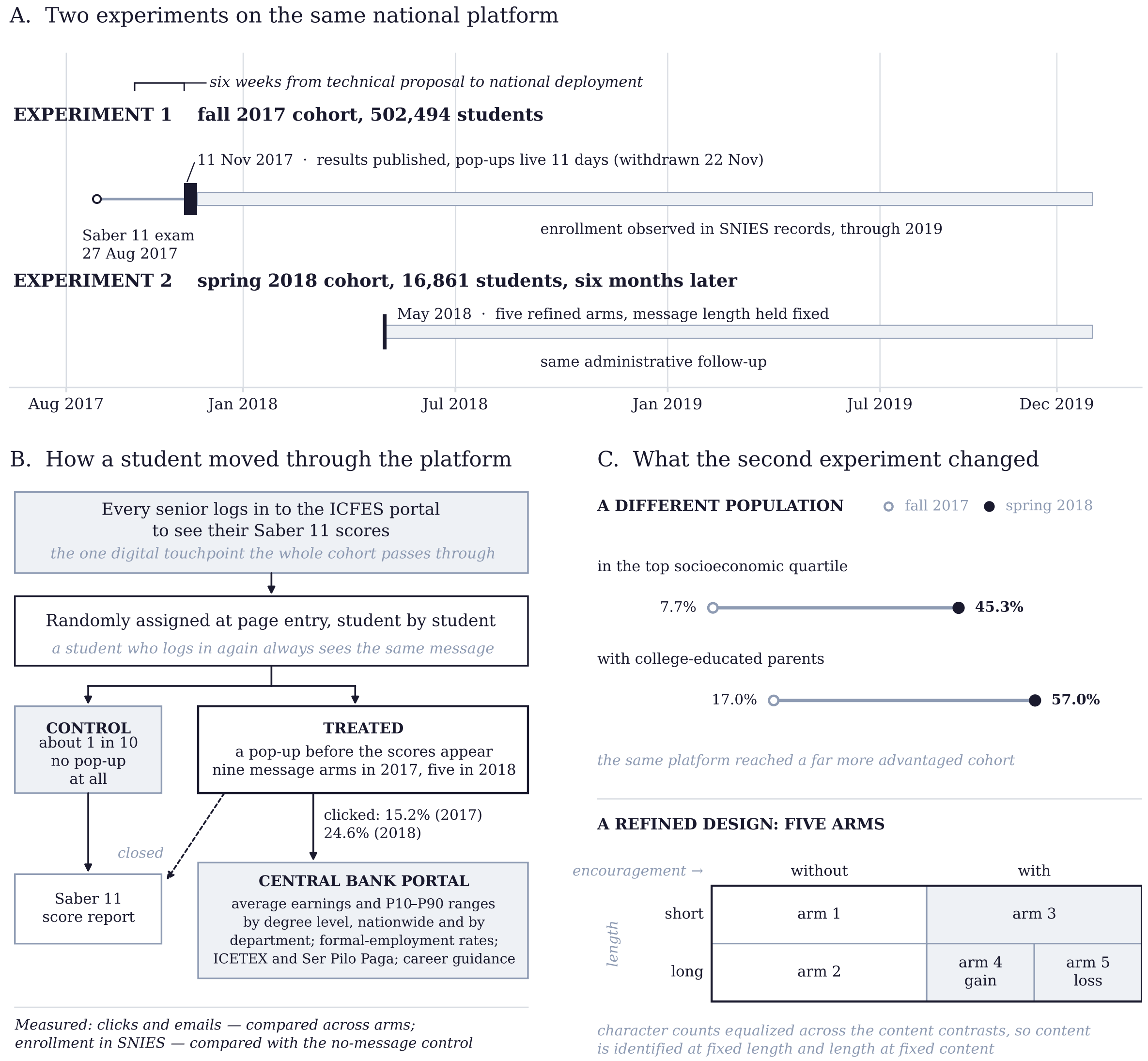
2.4 Experiment 2 (May 2018): Focused Iteration

Experiment 1's engagement results were observable almost immediately, and they taught us two lessons (documented in Section 4): the plain invitation dramatically outperformed the belief-targeting variants, and *which* belief a message targeted mattered far less than whether it targeted one at all. (Enrollment outcomes accumulate over years and were not yet available when the second experiment was designed.) Six months later we implemented a refined experiment on the 16,861 students of the spring 2018 cohort. We simplified message content substantially relative to 2017, and—because the 2017 contrasts bundled content with message length—built the 2018 arms to unbundle them: messages with and without encouragement were calibrated to identical character counts, so content effects are identified at fixed length and length effects at fixed content. The five arms are: short and long messages without encouragement, a short message with encouragement (similar to 2017's "salient" arm), and long messages with encouragement that mention average returns, framed as gains or losses. Each contrast identifies one object: long versus short without encouragement identifies message length at fixed neutral content; adding encouragement at fixed short length identifies the encouragement itself; adding framed returns at fixed long length identifies belief-targeting content—the 2017 headline contrast, now unbundled from length; and the gain-versus-loss pair replays the prospect-theory test on new text and a new population (Online Appendix Table B2 states each arm's role in the design).

The spring cohort differs sharply from the fall cohort: 45.3% fall in the highest socioeconomic quartile versus 7.7% in 2017, and 57.0% have college-educated parents versus 17.0% (Online Appendix Table C1). While not designed as a heterogeneity test, this compositional difference provides evidence on whether information interventions transport across populations within the same country and policy context—exactly the external-validity question that matters for scaling.

Figure 2 summarizes the design just described: panel A the timing of the two experiments, panel B the path each student took through the platform and the outcomes that path generates, panel C how the second cohort differed from the first.

Figure 2: Timing, platform flow, and the two cohorts



Notes: Schematic; the figure reports no estimates. Panel A: exam and results-publication dates are from student score reports and the ICFES deployment correspondence; the 2017 pop-ups were withdrawn on 22 November 2017, while the closing date of the 2018 exposure window is not documented in our records and is therefore shown as a marked date rather than a band. Panel B: the control group saw no pop-up at all, so its click-through rate is zero by construction and the informative engagement comparisons are across treated arms (Section 3); the click-through shares are treated averages (Section 4). The pop-up itself, and what the nine 2017 arms placed inside it, are shown in Figure 1; the full arm listings are in Online Appendix Tables B1 (2017) and B2 (2018). Panel C: cohort composition shares are from Online Appendix Table C1. Portal contents and platform screenshots are reproduced in Online Appendix B.

2.5 Cost Structure

Setting up the intervention—building the information portal, integrating randomization into the ICFES platform, computing returns by program and region, and obtaining approvals—cost approximately $100,000. Operating them costs less than $25,000 per year, since delivery, randomization, click tracking, and outcome measurement are automated on existing infrastructure. Across both experiments the all-in cost was roughly \$0.25 per randomized student. For scale, the mailed packets of C. M. Hoxby and Turner (2013) cost on the order of \$6 per student, so reaching one Colombian cohort by post would cost more than twenty times both experiments together. This is what makes testing many variants at full scale, and iterating, economically feasible (Online Appendix G).

3 Data and Empirical Strategy

We combine three administrative sources. ICFES records provide background characteristics (gender, age, socioeconomic stratum, parental education) and test scores for all *Saber 11* takers in both cohorts, together with the randomization assignment at login. A tracking system on the Central Bank information website records which students clicked the pop-up link or provided an email; the portal's server logs also record aggregate arrivals over the exposure window (the source of the 80,000 figure in the introduction), but the student-level analysis uses the click and email indicators linked to ICFES records. One caveat follows from the logging window. The click log runs from 11 to 24 November 2017, while the definitive ICFES delivery of treated students exceeds the displays recorded in the 24 November interim file by roughly 65,000. Those students fall outside the window—either they consulted after click logging ended or their clicks went unrecorded—so the measured rate understates engagement: among students whose displays the log does cover, portal arrivals imply a rate near 21% rather than the 15.2% we report. We report the lower bound throughout. Because the truncation is by calendar date while assignment was random at login, it falls on all arms alike and leaves the across-arm comparisons that carry our design results unaffected; the enrollment estimates depend on assignment rather than the click log, so they are unaffected for the same reason. Finally, the Ministry of Education's National Higher Education Information System (SNIES) records enrollment status, program, degree type (university versus short-cycle technical/technological), institution, and graduation for every semester from 2017 through 2019. We restrict attention to undergraduate enrollment and classify STEM programs by subject of study. The merged data follow both full cohorts from the exam to the higher education system with essentially no attrition.

For the 2017 cohort, the information arms vary along three cross-cutting dimensions (gain versus loss framing; nationwide versus personalized statewide statistics; average versus range presentation); the fourth contrast—the simple salient message versus any information message—is captured by a separate indicator in every specification. To test the behavioral mechanisms, we estimate intent-to-treat (ITT) effects with grouped treatment indicators pooling the arms on each side of the relevant dimension (e.g., the four gain-framed versus the four loss-framed arms; the full mapping is in Online Appendix Table C2):

$$Y_i = \alpha + \beta_0\, Salient_i + \beta_1\, TT1_i + \beta_2\, TT2_i + X_i'\delta + \varepsilon_i, \tag{1}$$

where Y_i is the outcome (clicking the link for engagement; enrollment for main results), $Salient_i$ indicates the simple message, $(TT1_i, TT2_i)$ are the grouped treatment pair for the relevant dimension, and X_i is a rich vector of student, family, and household controls. Grouped contrasts gain power but are valid only when interactions across dimensions are negligible (Muralidharan, Romero, & Wüthrich, 2025); the headline ranking survives disaggregation—plain beats every variant and gain beats loss at $p < 0.001$, and in Online Appendix Table B3 gain exceeds loss within all four personalization × presentation cells. For the 2017 cohort we report standard errors clustered at the message-by-department cell, the level at which message content varied. Because assignment was individual, this choice is conservative: clustering at an aggregate level can only overstate the variance of the treatment contrasts (Abadie, Athey, Imbens, & Wooldridge, 2023). For the 2018 cohort we include all five treatment indicators in a single regression with heteroskedasticity-robust standard errors. P-values reported in brackets in the 2017 enrollment exhibits are computed directly from the reported coefficients and standard errors; the 2018 exhibit reports multiple-testing-adjusted p-values from the estimation pipeline. Covariate balance across arms is consistent with successful randomization in both cohorts (Online Appendix E reports covariate-by-covariate and joint tests).

The design separates two questions with two different comparisons. *Does the intervention move enrollment?* is identified by comparing each message group to the no-message control—the intent-to-treat contrasts in equation (1). *Which message design captures attention?* is identified by comparisons *across* the nine treated arms: control students saw no pop-up and had no link to click, so the control is uninformative for engagement by construction, and the treated arms serve as one another's counterfactuals.

We focus on ITT estimates because they correspond to the policy-relevant impact of scaling: the average effect on all exposed students. Since 15–25% clicked through, treatment-on-treated effects would be 4–7 times larger if only engagers are affected, but engagement is itself an outcome of message design, so ITT is primary.

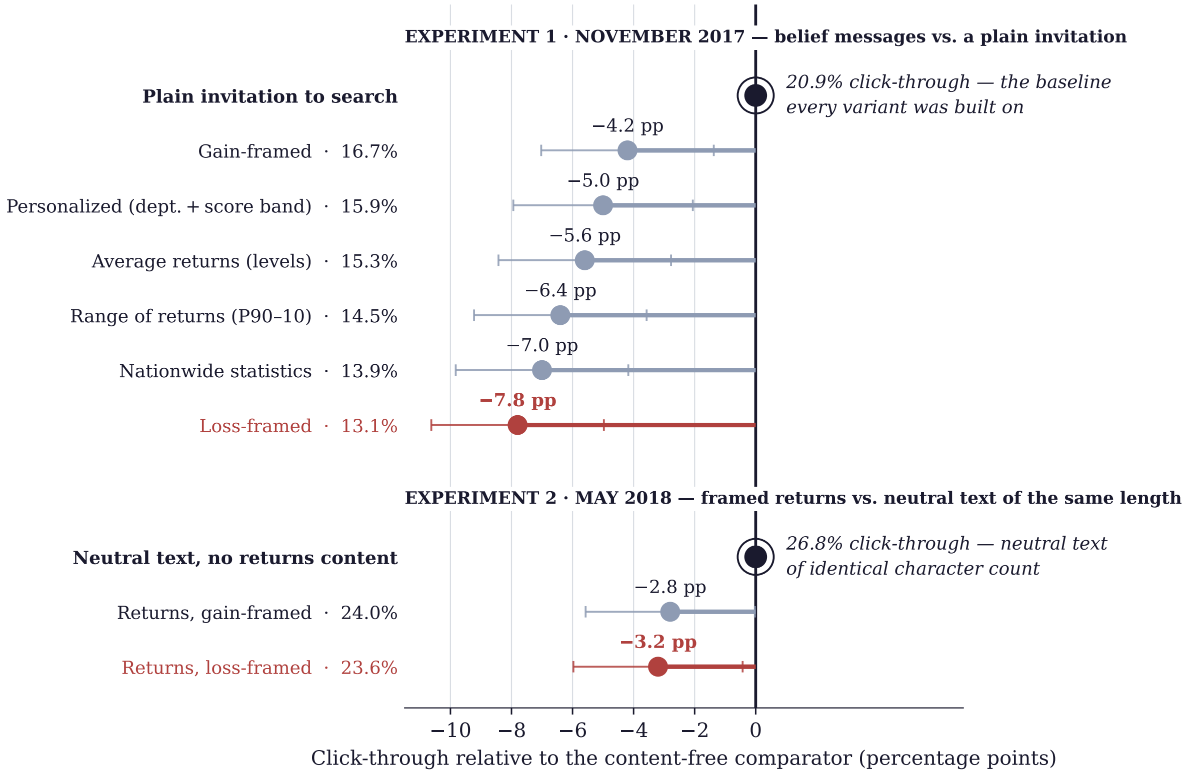
Two power calculations discipline the interpretation of our estimates. Under the conservative clustered inference we report, the realized standard errors for the any-tertiary-enrollment outcome in Experiment 1 (control mean 0.567) imply minimum detectable effects (two-sided, $\alpha = 0.05$, 80% power) of approximately 2.8 percentage points for the grouped treatment indicators (s.e. ≈ 0.010) and 3.6 percentage points for the salient arm (s.e. ≈ 0.013), so the confidence intervals reported below bound average enrollment effects within roughly ± 2–2.6 percentage points. The design itself is far more powerful. Because assignment was individual, the design-based standard error of a treatment–control comparison follows directly from the control mean and group sizes: with roughly 50,000 students per arm it is about 0.3 percentage points, a 95% interval half-width near 0.6 points. The two bracket what the experiment can say—effects above roughly 0.6 points are ruled out under the inference the design licenses, effects above 2.6 points under the most conservative reading. We report the clustered bound throughout because it is what our estimation delivers, but it understates the experiment's precision roughly fourfold. In Experiment 2, the realized robust standard errors on short-cycle enrollment (0.006–0.008 across arms) imply minimum detectable effects of approximately 1.7–2.2 percentage points.

4 Results

4.1 Engagement: Digital Nudges Change Behavior at Scale

In the 2017 cohort, 15.2% of treated students clicked through for more information. Engagement was higher among males, students with higher test scores, college-educated parents, higher income, computer and internet access, and urban locations (Online Appendix Table C3). Because control students saw no pop-up, the informative experimental comparisons are *across* message designs. The headline design result: the plain invitation achieved 20.9% click-through, while every belief-targeting variant landed at 13.1–16.7% ($p < 0.001$; Figure 3, top panel). Messages engineered to move specific beliefs—about the returns to each education level, about their dispersion, about the student's own returns—all underperformed the message that simply raised the value of looking. Within the belief-targeting arms, the clean prospect-theory test fails in reverse: loss-framed messages generated 13.1% click-through versus 16.7% for gain-framed equivalents ($p < 0.001$). Portal logs speak to whether the plain invitation bought shallower attention. Among students whose displays fall inside the click-logging window, 37–48% of each arm's arrivals opened at least one document in the portal; the plain invitation's were among the most likely to (46%), above every loss-framed arm, and three-quarters opened the departmental earnings tables. It attracted more students without attracting less engaged ones. These logs are descriptive—they cover the narrower window of Section 3—but reproduce the arm ranking independently (Online Appendix B). Personalization also underperformed expectations: statistics matched to the student's department and score band achieved 15.9% versus 13.9% for nationwide figures ($p < 0.001$), a gain only a fraction of the plain invitation's advantage, and the range-versus-levels presentation contrast shows the same pattern (the underlying regressions are in Online Appendix Tables C4–C5 and C8).

Figure 3: The belief-targeting penalty: engagement relative to content-free messages (2017 and 2018)



Notes: Each panel plots grouped click-through rates as differences from that experiment's content-free comparator, anchored at zero: in 2017, the plain invitation (20.9% click-through; $N = 502{,}494$); in 2018, a neutral message of identical character count (26.8%; $N = 16{,}861$). Row labels report each group's click-through rate. Whiskers are conservative 95% confidence intervals computed from the reported standard errors treating arm estimates as independent (cluster-robust in 2017, heteroskedasticity-robust in 2018); underlying regressions are in Online Appendix Tables C4–C5 and C8 (2017) and D2 (2018). Every pairwise 2017 contrast rejects equality at $p < 0.001$; in 2018 the framed-returns arms underperform the same-length neutral message ($p = 0.037$ gain, $p = 0.019$ loss).

4.2 Ex-Ante Predictions Versus Empirical Performance

Online Appendix Table A1 summarizes each mechanism's directional prediction against realized performance. Prospect theory is rejected on its own terms—loss framing did strictly worse than gain framing—while personalization delivers only a small gain and the salience/cognitive-load mechanism is the one that survives; the pattern is stable across gender and other subgroups. The point is not that these theories are "wrong," but that their implications for the *best implementation choice* are context-dependent: with mechanisms pointing in opposite directions, the winning design could not have been confidently forecast ex ante, and a one-variant pilot could easily have selected a design that is far from best-in-class.

Four (non-exclusive) explanations merit consideration. First, **information satiation**: the belief-targeting messages partially answer the question inside the pop-up, closing the gap that makes clicking worthwhile (Loewenstein, 1994); the plain invitation leaves it open, so part of the click contrast may reflect a reduced *need* to click. Three facts cap how much work satiation can do: the ranking is unchanged on the costlier email-provision margin—requesting that *more* information be sent, which a satisfied student has little reason to do (13.7% plain versus 9.1% gain and 7.1% loss; Online Appendix Table C8); the plain invitation's arrivals searched the portal at least as intensively (Section 4.1); and the enrollment nulls give no sign that statistics-bearing messages informed decisions through the pop-up alone. Second, **cognitive load in digital environments**: pop-ups appear in brief sessions, often on mobile devices, while students seek exam scores; complexity may impose processing costs that outweigh motivational benefits (Mullainathan & Shafir, 2013; Shah et al., 2012). Third, **trust and reactance on government platforms**: hard-sell framing may trigger skepticism about manipulation (Brehm, 1966), making plain information more credible. Fourth, **decision stage**: our intervention occurs during information search, months before enrollment decisions, whereas loss aversion may matter more for final, irreversible choices. Whichever dominates, the ambiguity is itself informative: if the mechanisms cannot be cleanly adjudicated even ex post, no ex-ante ranking could have been relied upon.

4.3 Null Average Enrollment Effects (2017)

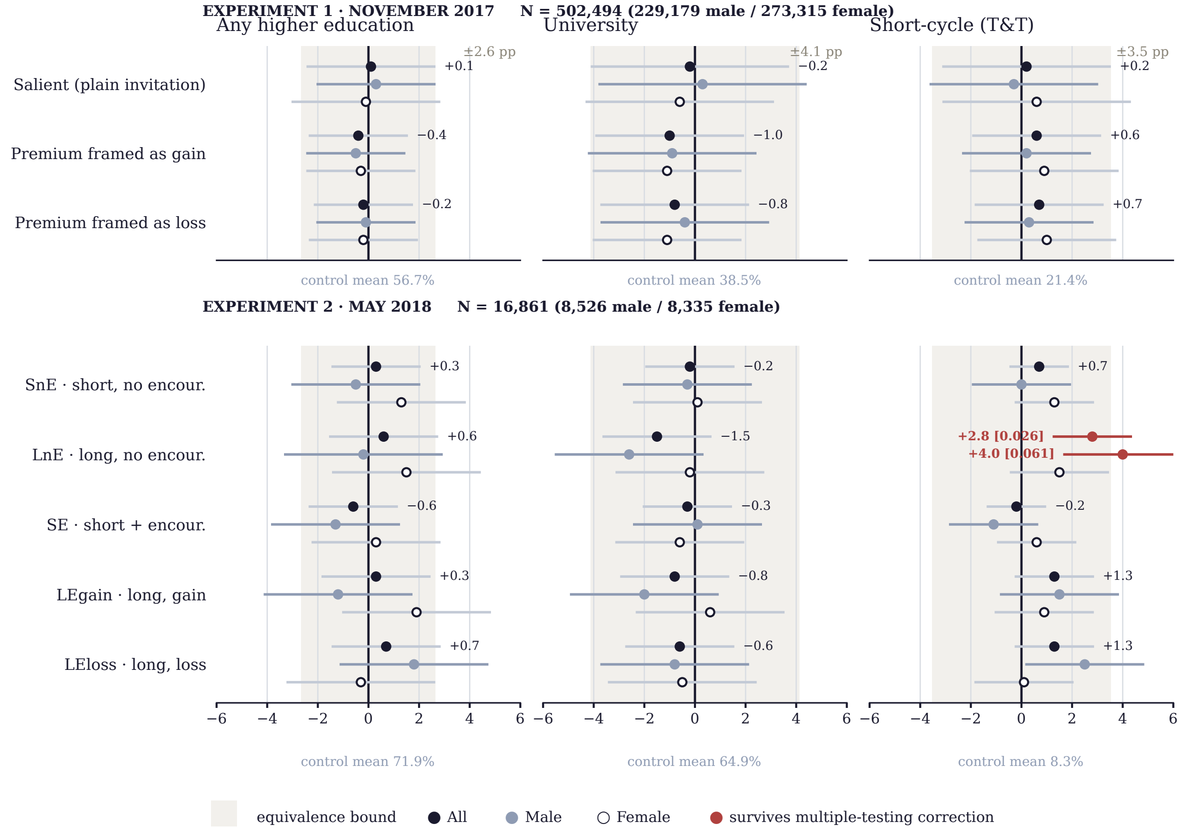
Enrollment is where the no-message control does its work: Figure 4 compares each message group to students who saw nothing at all (results for the other 2017 treatment dimensions, which are very similar, appear in Online Appendix Tables C6–C7; the full coefficient tables, including the male and female columns, are Online Appendix Tables C9 and D3). Despite the large engagement response, enrollment effects are null: the salient message's effect on overall tertiary enrollment is 0.001 (s.e. 0.013), a 95% confidence interval of $[-0.024, 0.026]$, so we can rule out average effects larger than 2.6 percentage points—and the point estimate is less than a tenth of its standard error. That bound is deliberately conservative; the design-based standard error implied by individual assignment is four times smaller, ruling out effects above roughly 0.6 percentage points (Section 3). Point estimates for university and short-cycle enrollment are similarly small—every coefficient lies within 1.3 percentage points of zero—and none is statistically distinguishable from zero under our conservative inference, across all message variants (STEM outcomes, reported by subgroup in Online Appendix Figures F2–F3, show the same pattern). Inducing information search did not translate into enrollment changes. Two considerations discipline the interpretation. First, the extensive margin is the outcome policymakers cite, and the one on which the celebrated enrollment effects in this literature were obtained (Bettinger et al., 2012; Dynarski et al., 2021), so bounding it at national scale is informative regardless of priors—and more so here than in the U.S. high-achiever studies, since 43% of this cohort does not enroll. The search logic of Section 2.2 predicts that information which changes minds should appear first in *which* program a student chooses, or whether she persists—the margin C. M. Hoxby and Turner (2013) actually moved—rather than in whether she enrolls at all. On the analysis subsample we hold, we test those margins directly—composition among enrollers, dropout through 2019, switching—and all three are null, every point estimate within 0.8 percentage points of zero against minimum detectable effects of roughly 3 points for dropout and 7–9 for composition (Online Appendix C). The re-sorting reading is consistent with the data but not supported by it; full-sample estimates would tighten these bounds roughly threefold. Second, at roughly \$0.25 per student, even effects well inside our confidence interval—a few tenths of a percentage point—would be highly cost-effective. The nulls rule out effects of the size celebrated in the intensive-intervention literature, not the possibility that the intervention pays for itself—a possibility the sharper design-based bound narrows but does not close.

4.4 Replication and Refinement: The 2018 Cohort

In the 2018 cohort, overall take-up was 24.6%—higher than 2017's 15.2%, though the comparison bundles population, message-design, and exposure-window differences (engagement patterns by subgroup and arm are in Online Appendix Tables D1–D2). Because the 2018 arms hold character counts fixed across content contrasts, they cleanly separate what 2017 bundled. Three lessons follow. *Content, not length, carries the penalty*: at identical character counts, adding encouragement framing that mentioned returns significantly reduced engagement relative to a neutral message of the same length (26.8% versus 24.0% gain-framed and 23.6% loss-framed; pairwise $p = 0.037$ and $p = 0.019$; Figure 3, bottom panel), while length itself did not matter (short versus long without encouragement, $p = 0.12$) and encouragement without returns content was inert at fixed length ($p = 0.81$). These are the comparisons the equal-length calibration exists to make—prespecified by the design itself rather than selected from the pairwise table. *The framing result is directionally consistent with 2017*: the loss-framed variant again ranked last. *And the refinement delivered a caution against over-learning*: the best-performing 2018 arm was a *long* message, beating both short arms on point estimates—so the lesson we ourselves drew from Experiment 1, that shorter is better, did not survive its own replication. Design rankings resist ex-ante prediction even when the prediction comes from one's own previous experiment.

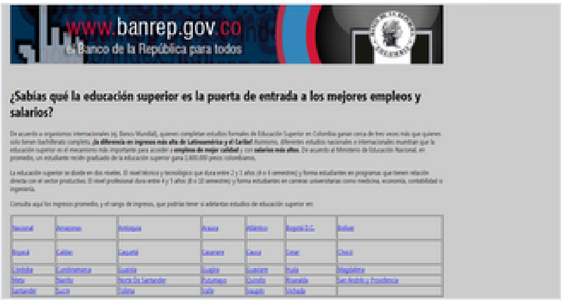
Figure 4 reports enrollment effects for both cohorts. We find suggestive evidence of shifts toward short-cycle technical and technological enrollment: the "long without encouragement" arm increased short-cycle enrollment by 2.8 percentage points overall (corrected $p = 0.03$)—a third of the 8.3% control mean—driven by male students (+4.0pp against a 10.3% male control mean), with weakly negative point estimates for university enrollment, consistent with substitution toward lower-cost, shorter-duration programs rather than an expansion of enrollment—the one margin on which either experiment shows movement. Three cautions apply: the sample is thirty times smaller than in 2017, corrected p-values of 0.03–0.06 across five arms and multiple outcomes leave room for false discoveries, and the effects are concentrated in the relatively advantaged spring population. We therefore interpret these as suggestive—informative about where information interventions may work, not proof that the iteration "succeeded."

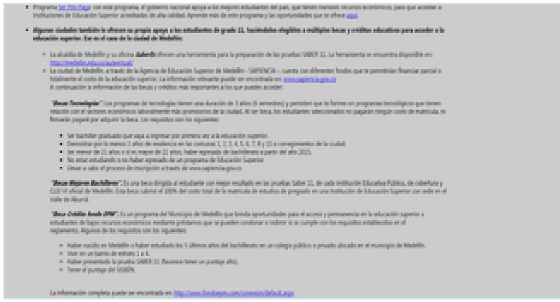
Figure 4: What the enrollment nulls rule out, and the one effect that survives



Notes: Coefficients on each message arm relative to the no-message control, in percentage points, with 95% confidence intervals; within each arm the three estimates are the full sample, males and females. 2017 standard errors are clustered at the message-by-department cell; 2018 standard errors are robust, with p-values adjusted for multiple testing. Labels give the full-sample estimate. The shaded band on each panel is that outcome's equivalence bound—the largest average effect the 2017 data admit, computed as the widest 95% confidence-interval endpoint among the 2017 full-sample arms; it is ±2.6 percentage points for overall tertiary enrollment and looser for the two components, which are estimated less precisely. Two estimates survive multiple-testing correction at the 10% level, both of them the 2018 long-without-encouragement arm on short-cycle enrollment; they are drawn in red with their adjusted p-value. No significance stars are shown: the stars in Online Appendix Tables C9 and D3 derive from unadjusted p-values, while the bracketed p-values there—and the ones reported here—are adjusted. Full coefficient tables, R-squared and control means are in Online Appendix Tables C9 and D3.

Figure 5: The information a click delivered





Section 1 · Returns to higher education

Graduates earn roughly three times what those with only completed secondary earn — the widest such gap in Latin America and the Caribbean. A recent graduate earns COP 1,600,000 on average. Links resolve to average income and income range for the nation and each of the 32 departments.

Section 3 · Financial aid

ICETEX credit, the Ser Pilo Paga scholarship, and municipal programs documented in detail for Medellín: SaberEs, SAPIENCIA, Becas Tecnologías, Becas Mejores Bachilleres and the forgivable Beca-Crédito Fondo EPM, each with its eligibility rules.

EXAMPLE OF THE UNDERLYING STATISTICS · BOGOTÁ D.C.

Level	Average	Range	Formal
Técnica	1,157,091	885,929 – 1,518,110	61%
Tecnológica	1,352,764	1,064,500 – 1,682,190	72%
Universitaria	1,955,908	1,424,066 – 2,806,704	79%

One such page for each department, plus a bar chart of average income by level, in a 34-page annex.

WHY ENGAGEMENT NEED NOT MOVE ENROLLMENT

The portal supplied population statistics, program descriptions and aid rules — public inputs, identical for every student in a department and score band. It did not tell a student which programs they would be admitted to, what their own application would cost, or when to act. A click records attention, not the resolution of a student's own uncertainty.

Notes: The link resolved to a purpose-built page at the Banco de la República carrying a per-student identifier, captured on 14 November 2017 at *www.banrep.gov.co/cgi-bin/educacion/educacion.pl*. Two of its four sections are shown; the other two cover career orientation and the underlying departmental statistics. Income figures are Colombian pesos per month; the formality rate is the share of graduates in formal employment. The Bogotá table is page 6 of the portal's 34-page statistical annex, which carries one such page for every department. Source: Ministerio de Educación Nacional; calculations by Banco de la República.

4.5 Heterogeneity

If population composition drove the cross-cohort contrast, 2017 effects should be strongest among students resembling the 2018 cohort. The evidence is mixed: 2017 engagement is indeed higher among higher-SES, higher-scoring students, but enrollment effects remain null even among advantaged 2017 subgroups (Online Appendix F reports effects by socioeconomic level, score quartile, gender, parental education, and geography for both cohorts). The 2018 patterns likely reflect some combination of composition, the refined designs, and the lower barrier to technical enrollment.

5 Conclusion

We embedded two nationwide randomized experiments in the platform where every Colombian high school senior collects their exit-exam results, nudging students to search a purpose-built portal of earnings, costs, and financial-aid information. The experiments separate three objects that scale differently. *Behavioral engagement scales*: pop-ups reliably moved 15–25% of a national population to seek information, in both cohorts. *Downstream outcomes need not scale with engagement*: enrollment effects are null against the no-message control—bounded at 2.6 percentage points conservatively, and at roughly 0.6 points design-based—consistent with binding financial, preparation, and application constraints (Bettinger et al., 2012; Castleman & Page, 2014; Dynarski et al., 2021)—nor do program composition or persistence respond on the subsample we can analyze. And *the relative performance of designs is itself a scaling object*: every belief-targeting variant lost to a plain invitation to search, loss framing performed worst, and even the lesson we drew from our own first experiment—that simpler is better—did not survive the second. Effects and rankings that look predictable ex ante shrink or shift in the implementation environment (Allcott, 2015; Banerjee et al., 2017; Bold et al., 2018; DellaVigna & Linos, 2022).

Read through the lens of search, the pattern is informative. Students did search—many went on to open the underlying earnings tables—but population statistics do not resolve a student's uncertainty about her own options. Related evidence from Chile sharpens the reading: families are surprisingly unaware of their own choice sets and hold endogenously biased beliefs about unsearched options (Agte et al., 2024); personalized information about students' own planned programs shifted enrollment composition toward higher-return degrees (Hastings et al., 2015); and personalized warnings delivered inside the application window changed behavior substantially (Arteaga et al., 2022). Generic prompts to search arrive too late and say too little; specific, personal, decision-window information forces an update. The null results documented here are part of what redirected this agenda toward dynamic, personalized information policies—designs that, in Chile, roughly doubled matching probabilities for unmatched applicants after scale-up (Larroucau et al., 2025).

For policy, universal generic information campaigns are unlikely to move enrollment by policy-noticeable amounts even when they successfully move attention; more promising are designs that bundle information with financial aid, expand consideration sets toward technical pathways, and arrive at actionable decision points (Castleman & Page, 2015; Dynarski et al., 2021; Page, Kehoe, Castleman, & Sahadewo, 2019). For research design, the study shows what government digital platforms make possible as research infrastructure: the platform is universal, outcomes are administrative, the implementation environment is the policy environment itself (Al-Ubaydli et al., 2017), and the full cycle took six months and roughly $125,000. Our limitations are the usual ones of a single setting: one-time messages at a single touchpoint, enrollment rather than completion as the outcome, and cohort composition confounded with design refinements across experiments. But the central lesson travels. When experimentation is this cheap, the pilot-to-scale sequence inverts—test many variants at scale, learn which retain voltage, iterate. A one-arm pilot could as easily have shipped the 13.1% design as the 20.9% one, and our own first-round lesson failed in the second: cheap deployment does not weaken the case for testing; it removes the excuse for skipping it.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, *138*(1), 1–35.

Agte, P., Allende, C., Kapor, A., Neilson, C. A., & Ochoa, F. (2024). *Search and biased beliefs in education markets* (NBER Working Paper No. 32670). National Bureau of Economic Research. (R&R at Econometrica)

Allcott, H. (2015). Site selection bias in program evaluation. *The Quarterly Journal of Economics*, *130*(3), 1117–1165.

Altonji, J. G., Blom, E., & Meghir, C. (2012). Heterogeneity in human capital investments: High school curriculum, college major, and careers. *Annual Review of Economics*, *4*(1), 185–223.

Al-Ubaydli, O., List, J. A., & Suskind, D. L. (2017). What can we learn from experiments? Understanding the threats to the scalability of experimental results. *American Economic Review*, *107*(5), 282–286.

Arteaga, F., Kapor, A. J., Neilson, C. A., & Zimmerman, S. D. (2022). Smart matching platforms and heterogeneous beliefs in centralized school choice. *The Quarterly Journal of Economics*, *137*(3), 1791–1848.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., ... Walton, M. (2017). From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives*, *31*(4), 73–102.

Bergman, P., Denning, J. T., & Manoli, D. (2019). Broken tax breaks? Evidence from a tax credit information experiment with 1,000,000 students. *Journal of Policy Analysis and Management*, *38*(3), 706–731.

Bettinger, E. P., Long, B. T., Oreopoulos, P., & Sanbonmatsu, L. (2012). The role of application assistance and information in college decisions: Results from the H&R block FAFSA experiment. *The Quarterly Journal of Economics*, *127*(3), 1205–1242.

Bird, K. A., Castleman, B. L., Denning, J. T., Goodman, J., Lamberton, C., & Rosinger, K. O. (2021). Nudging at scale: Experimental evidence from FAFSA completion campaigns. *Journal of Economic Behavior & Organization*, *183*, 105–128.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2018). Experimental evidence on scaling up education reforms in Kenya. *Journal of Public Economics*, *168*, 1–20.

Bonilla-Mejía, L., Bottan, N. L., & Ham, A. (2019). Information policies and higher education choices: Experimental evidence from Colombia. *Journal of Behavioral and Experimental Economics*, *83*.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2013). Salience and consumer choice. *Journal of Political Economy*, *121*(5), 803–843.

Brehm, J. W. (1966). *A theory of psychological reactance*. Academic Press.

Camacho, A., Messina, J., & Uribe, J. P. (2017). *The expansion of higher education in Colombia: Bad students or bad programs?* (Documentos CEDE No. 2017-13). Universidad de los Andes, Facultad de Economía, CEDE.

Castleman, B. L., & Page, L. C. (2014). *Summer melt: Supporting low-income students through the transition to college*. Cambridge, MA: Harvard Education Press.

Castleman, B. L., & Page, L. C. (2015). Summer nudging: Can personalized text messages and peer mentor outreach increase college going among low-income high school graduates? *Journal of Economic Behavior & Organization*, *115*, 144–160.

DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica*, *90*(1), 81–116.

Dynarski, S., Libassi, C. J., Michelmore, K., & Owen, S. (2021). Closing the gap: The effect of reducing complexity and uncertainty in college pricing on the choices of low-income students. *American Economic Review*, *111*(6), 1721–1756.

Gurantz, O., Howell, J., Hurwitz, M., Larson, C., Pender, M., & White, B. (2021). A national-level informational experiment to promote enrollment in selective colleges. *Journal of Policy Analysis and Management*, *40*(2), 453–479.

Hastings, J., Neilson, C. A., Ramirez, A., & Zimmerman, S. (2016). (Un)informed college and major choice: Evidence from linked survey and administrative data. *Economics of Education Review*, *51*, 136–151.

Hastings, J., Neilson, C. A., & Zimmerman, S. (2015). *The effects of earnings disclosure on college enrollment decisions* (NBER Working Paper No. 21300). National Bureau of Economic Research.

Hoxby, C., & Avery, C. (2013). The missing "one-offs": The hidden supply of high-achieving, low-income students. *Brookings Papers on Economic Activity*(1), 1–65.

Hoxby, C. M., & Turner, S. (2013). *Expanding college opportunities for high-achieving, low income students* (SIEPR Discussion Paper No. 12-014). Stanford Institute for Economic Policy Research.

Hyman, J. (2020). Can light-touch college-going interventions make a difference? Evidence from a statewide experiment in Michigan. *Journal of Policy Analysis and Management*, *39*(1), 159–190.

Jensen, R. (2010). The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics*, *125*(2), 515–548.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kirkebøen, L. J., Leuven, E., & Mogstad, M. (2016). Field of study, earnings, and self-selection. *Quarterly Journal of Economics*, *131*(3), 1057–1111.

Kreuter, M. W., Farrell, D. W., Olevitch, L. R., & Brennan, L. K. (2000). *Tailoring health messages: Customizing communication with computer technology*. Mahwah, NJ: Lawrence Erlbaum Associates.

Larroucau, T., Rios, I., Fabre, A., & Neilson, C. (2025). *College application mistakes and the design of information policies at scale* (NBER Working Paper No. 34164). National Bureau of Economic Research. (Conditionally accepted, Journal of Political Economy)

Lavecchia, A. M., Liu, H., & Oreopoulos, P. (2016). The behavioral economics of education: Progress and possibilities. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 5, pp. 1–74). Elsevier.

List, J. A. (2022). *The voltage effect: How to make good ideas great and great ideas scale*. New York, NY: Currency.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75–98.

Londoño-Vélez, J., Rodríguez, C., & Sánchez, F. (2020). Upstream and downstream impacts of college merit-based financial aid for low-income students: Ser pilo paga in Colombia. *American Economic Journal: Economic Policy*, *12*(2), 193–227.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, *11*(1), 57–91.

Mullainathan, S., & Shafir, E. (2013). *Scarcity: Why having too little means so much.* Times Books.

Muralidharan, K., Romero, M., & Wüthrich, K. (2025). Factorial designs, model selection, and (incorrect) inference in randomized experiments. *Review of Economics and Statistics*, *107*(3), 589–604.

Page, L. C., Kehoe, S. S., Castleman, B. L., & Sahadewo, G. A. (2019). More than dollars for scholars: The impact of the dell scholars program on college access, persistence, and degree attainment. *Journal of Human Resources*, *54*(3), 683–725.

Shah, A. K., Mullainathan, S., & Shafir, E. (2012). Some consequences of having too little. *Science*, *338*(6107), 682–685.

Stigler, G. J. (1961). The economics of information. *Journal of Political Economy*, *69*(3), 213–225.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061.

Vivalt, E. (2020). How much can we generalize from impact evaluations? *Journal of the European Economic Association*, *18*(6), 3045–3089.

Wiswall, M., & Zafar, B. (2015). Determinants of college major choice: Identification using an information experiment. *The Review of Economic Studies*, *82*(2), 791–824.

World Bank. (2024). *World development report 2024: The middle-income trap* (Tech. Rep.). Washington, DC: World Bank.