American Journal of Interdisciplinary Studies

Volume: 14; **Issue:** 6
Pages: 01-40
eISSN: 3067-5146

A Quantitative Evaluation of Behavioral Design Patterns on User Engagement and Activation in Large-Scale Campus Software Products

Gazi Yeasin Arifin[1];

[1]. *Product Designer, Associated Students, Inc. (ASI);*
California State University, East Bay, Hayward, California, USA;
Bachelor of Science in Software Engineering, American International University-Bangladesh;
Email: arifin.yeasin@gmail.com

Doi: 10.63125/x3d5k517
Received: *12 March 2024;* ***Revised:*** *09 April 2024;* ***Accepted:*** *01 May 2024;* ***Published:*** *02 June 2024;*

Abstract

Campus software products borrow behavioural design patterns (checklists, defaults, reminders, social proof, goals, streaks, loss framing, friction reduction, personalised timing and urgency cues) to raise activation and engagement, yet the evidence in higher education is thin, field effects are small, and harms (opt-outs, complaints, loss of trust, crowding-out of motivation) are rarely measured. This article presents a quantitative evaluation framework for such patterns and applies it on a transparent synthetic testbed: an agent-based model of 40,000 students in five campuses over a 16-week term with latent motivation, habits, notification tolerance and peer influence through course sections, on which ten patterns are implemented with stylised mechanisms from the literature and the true effect of each is known. Activation is moved most by friction reduction and an onboarding checklist (about 12 and 11 percentage points), sustained engagement in the final term week by gated social proof, timed reminders and personalised timing (1.4 to 2.1 points, building over the term), and short-run engagement by an urgency countdown, goals and streaks, whose effects decay to near zero. Default opt-in raises opt-outs by 11 points and the urgency countdown produces 11 complaints per 100 students; three patterns raise engagement while lowering opt-outs. The ranking of patterns inverts between activation and sustained engagement and between week 2 and week 16. Standard methods are scored against the truth: a megastudy on 40,000 students cannot resolve the persistent 1 to 2-point effects, user-level randomisation understates social-proof effects by a third or more under peer interference, weekly peeking raises the A/A false-positive rate from 2.5% to 27.5%, small samples exaggerate significant effects several-fold, CUPED cuts standard errors by 10 to 37%, and uplift targeting finds heterogeneity that is not there. A Monte Carlo analysis over fourteen assumed parameters shows which conclusions survive. The article contributes a reference architecture, a behavioural state model, an experiment lifecycle workflow, a regulatory and ethical mapping and a governance roadmap. All data are synthetic; all results are properties of stated assumptions.

Keywords

Behavioral Design; Nudges; User Engagement; Activation; A/B Testing; Campus Software; Learning Management Systems; Interference; Dark Patterns; Simulation

Introduction

Universities now run their relationship with students through software. Learning management systems carry assignments and grades, student portals and mobile applications carry deadlines, advising, registration and notifications, and the same products are expected to keep students engaged, to help them meet deadlines and to flag those who are drifting away (Dahlstrom et al., 2014; Henrie et al., 2015; EDUCAUSE, 2023). The teams that build these products have borrowed from consumer software the toolkit of behavioural design: onboarding checklists with progress bars, default opt-ins, timed reminders, social proof, goal setting, streaks and badges, loss-framed messages, friction reduction, personalised timing and urgency cues. The toolkit rests on a large behavioural science literature, from defaults that move organ donation and retirement saving (Johnson & Goldstein, 2003; Madrian & Shea, 2001) to social norms that move energy use (Allcott, 2011; Schultz et al., 2007), reminders that move saving and vaccination (Karlan et al., 2016; Milkman et al., 2011) and the general programme of choice architecture (Thaler & Sunstein, 2008; Kahneman, 2011).

Two things complicate the transfer to campus products. First, the effects are smaller and less reliable than the early literature suggested. Meta-analyses of nudging report median effects that shrink as study quality rises (Hummel & Maedche, 2019; Mertens et al., 2022), a re-analysis that adjusts for publication bias finds little evidence of an average effect (Maier et al., 2022), and the two largest administrative nudge units found effects about a sixth the size of those in academic papers (DellaVigna & Linos, 2022). In higher education specifically, large field experiments on goal setting, reminders and planning found small or null effects on grades and persistence (Dobronyi et al., 2019; Oreopoulos & Petronijevic, 2019; Clark et al., 2020), financial-aid nudges that worked at a university scale did not work at a national scale (Bird et al., 2021; Castleman & Page, 2016), and interventions that worked in one online course did not transfer to others (Kizilcec et al., 2020). Second, campus products are not consumer products. Their users are a captive population, often young adults, whose attention the institution has a duty to protect; their data are protected by federal law; and experiments on them raise questions that the Facebook emotional-contagion study made public (Kramer et al., 2014; Meyer et al., 2019; Slade & Prinsloo, 2013). Some of the toolkit shades into what regulators now call dark patterns (Gray et al., 2018; Mathur et al., 2019; Federal Trade Commission, 2022).

The product teams' answer is to measure: run controlled experiments on each pattern and keep the ones that work. The measurement is harder than it looks. Engagement metrics are many and easily gamed (Dmitriev et al., 2017); novelty inflates short tests (Kohavi et al., 2012; Hohnhold et al., 2015); social patterns violate the independence that randomisation by user assumes, because students share course sections (Hudgens & Halloran, 2008; Eckles et al., 2017); ten patterns times six metrics is sixty hypotheses (Benjamini & Hochberg, 1995); peeking at results inflates false positives (Johari et al., 2017); small samples produce estimates of the wrong sign and size (Gelman & Carlin, 2014); and the effects that matter, retention and on-time work at the end of the term, arrive months after the decision has to be made (Athey et al., 2019). A campus team that evaluates patterns without attending to these problems will ship patterns that do not work and miss patterns that do.

This article asks what a quantitative evaluation of behavioural design patterns in a large campus product should look like, and what it would find. Because no campus can run two hundred experiments on the same term, and because the true effect of a pattern on a real population is never observed, the article builds a transparent synthetic testbed: an agent-based model of 40,000 students in five campuses using a student hub application over a 16-week term, with latent motivation, habits, notification tolerance and social connections through course sections, and ten patterns implemented as interventions whose mechanisms and stylised effect sizes are taken from the literature. On this testbed the true effect of every pattern is known, so the evaluation methods themselves can be evaluated. It poses four research questions. RQ1: how large are the effects of the ten patterns on activation and on engagement, how do they evolve over the term and how do they differ across students? RQ2: what harms do the patterns produce (opt-outs, complaints, idle use, loss of trust, crowding-out of motivation), and how do harms trade against engagement? RQ3: how well do standard experimentation methods recover the true effects: which randomisation unit, estimator, sample size, duration and correction, and what biases remain? RQ4: how stable are the answers when the assumed effect mechanisms vary?

The contributions are a reference architecture for an experimentation platform inside a campus product with its safeguards (Figure 1), a behavioural state model that places each pattern on the transition it acts on (Figure 2), an experiment lifecycle workflow with ethics and student review built in (Figure 3), a synthetic testbed and evaluation framework that are released in full (Figure 4), a quantified comparison of the ten patterns with their time paths, heterogeneity and harms, an evaluation of the evaluation methods, a Monte Carlo analysis of robustness and a governance roadmap. The study is explicit about what it is not: it does not use data from any campus or product, and the effects it reports are properties of assumptions, not measurements. Its value is in the structure: which patterns are worth testing first, which harms must be measured alongside engagement, and which measurement errors will mislead a team that does not design against them.

Background and related work

Behavioural design patterns and the evidence behind them

The patterns a product team can deploy map onto behaviour change techniques with named mechanisms (Michie et al., 2011; Michie et al., 2013; Fogg, 2009; Oinas-Kukkonen & Harjumaa, 2009). Defaults exploit inertia: when the default option is to participate, participation rises and stays high, as it did for organ donation and retirement plans (Johnson & Goldstein, 2003; Madrian & Shea, 2001). Reminders act on attention and memory: a text message before a deadline raises saving and vaccination and reduces summer melt among college-bound students (Karlan et al., 2016; Milkman et al., 2011; Castleman & Page, 2015). Social proof acts on descriptive norms: telling hotel guests that most others reuse towels raises reuse, and telling households how their energy use compares with neighbours lowers it, with the well-known boomerang for those already below the norm unless an injunctive cue is added (Goldstein et al., 2008; Schultz et al., 2007; Allcott, 2011). Goal setting and implementation intentions act on planning (Gollwitzer, 1999); the goal-gradient effect, in which effort rises as a goal nears, underlies progress bars and checklists (Kivetz et al., 2006); temporal landmarks motivate fresh starts (Dai et al., 2014); loss framing draws on loss aversion (Tversky & Kahneman, 1991). Friction reduction is the least glamorous and often the largest lever: simplifying a notice raised take-up of a tax credit more than any behavioural framing (Bhargava & Manoli, 2015), and the removal of sludge is now a policy agenda of its own (Sunstein, 2022). Gamification, the use of game elements in non-game contexts (Deterding et al., 2011), has a mixed record (Aminul Islam, 2024; Ashik, 2024; Kynatun & Saba, 2024): reviews find positive effects that depend on context and decay (Hamari et al., 2014; Sailer & Homner, 2020), a longitudinal classroom study found lower intrinsic motivation and grades (Hanus & Fox, 2015), and the meta-analytic finding that extrinsic rewards can crowd out intrinsic motivation is the mechanism to fear (Deci et al., 1999; Ryan & Deci, 2000). Habit formation gives the time scale: habits form over weeks of repetition and plateau, and context cues sustain them (Lally et al., 2010; Wood & Neal, 2007).

The evidence on effect sizes has matured. Across domains, meta-analyses find median relative effects around a fifth, with wide variance and a strong dependence on study quality (Hummel & Maedche, 2019; Mertens et al., 2022); adjusting for publication bias removes most of the average (Maier et al., 2022); and administrative nudge units, which register every trial, find average effects of about 1.4 percentage points against 8.7 in the academic literature (DellaVigna & Linos, 2022). Persistence is the other concern: energy-conservation effects decayed when the intervention stopped and persisted only with repetition (Allcott & Rogers, 2014). Megastudies, which test many interventions in one population at the same time (Md Maksudul, 2024; Md Shahid, 2024; Md Tazul Islam, 2024), were introduced to address the comparability problem that separate experiments leave open, and they found that the ranking of interventions differed from experts' predictions and that most effects were small (Milkman et al., 2021a; Milkman et al., 2021b).

Nudging and engagement in higher education

Education has its own nudge literature, reviewed by Damgaard and Nielsen Damgaard and Nielsen (2018). Text-message nudges reduced summer melt and increased financial-aid renewal in targeted populations (Castleman & Page, 2015; Castleman & Page, 2016), but the same approach at national scale did not move FAFSA completion (Bird et al., 2021). In large field experiments at a Canadian university, goal setting, academic reminders and coaching by text produced small or no effects on grades and persistence (Dobronyi et al., 2019; Oreopoulos & Petronijevic, 2019), and goal setting in a British

university moved task completion but not performance (Clark et al., 2020). Brief psychological interventions have had effects that depend strongly on context: a national growth-mindset experiment found effects concentrated in lower-achieving students in supportive schools (Yeager et al., 2019), a social-belonging intervention improved outcomes for minority students (Walton & Cohen, 2011), and behavioural interventions that worked in one online course did not transfer across courses and populations (Kizilcec et al., 2020). Student engagement itself is a multidimensional construct with behavioural, emotional and cognitive components (Fredricks et al., 2004), and the engagement that software can measure (logins, clicks, submissions) is a proxy for it whose validity has to be argued, not assumed (Henrie et al., 2015). Learning analytics has used that proxy to build early-warning systems from LMS activity (Macfadyen & Dawson, 2010; Arnold & Pistilli, 2012; Siemens, 2013), with ethical debates about what the institution owes students once it knows (Slade & Prinsloo, 2013; Prinsloo & Slade, 2017; Ifenthaler & Schumacher, 2016). Persistence theory places engagement in a social context (Tinto, 1993; Kuh, 2008), which is why course sections matter in the testbed below.

Measuring engagement and running experiments

User engagement with technology has been defined and measured in human-computer interaction as a multi-attribute construct (O'Brien & Toms, 2008; O'Brien et al., 2018); product analytics reduces it to behavioural metrics (activation, daily and weekly active use, retention), and customer-base models describe retention as a stochastic process (Fader et al., 2005; Lemon & Verhoef, 2016). The practice of online controlled experiments is well documented: the survey and practical guide by Kohavi and colleagues, the infrastructure that runs overlapping experiments at scale, the challenges summit of the large platforms and the pitfalls of metric interpretation (Kohavi et al., 2009; Kohavi et al., 2020; Tang et al., 2010; Gupta et al., 2019; Dmitriev et al., 2017; Bakshy et al., 2014; Xu et al., 2015). Specific threats are known: novelty and learning effects that make short experiments misleading (Kohavi et al., 2012; Hohnhold et al., 2015); peeking, which inflates false positives unless sequential methods are used (Johari et al., 2017); invalid experiments that automatic diagnostics can catch (Chen et al., 2019); and the loss of sensitivity that pre-experiment covariates recover (Deng et al., 2013). Causal inference supplies the framework (Rubin, 1974; Holland, 1986; Imbens & Rubin, 2015; Athey & Imbens, 2017), including heterogeneous effects estimated with forests and meta-learners (Wager & Athey, 2018; Künzel et al., 2019), uplift modelling for targeting (Gutierrez & Gérardy, 2017), bandits for adaptive allocation (Li et al., 2010; Lattimore & Szepesvári, 2020), surrogate indices for long-term outcomes (Athey et al., 2019) and the statistics of generalisation across sites (Tipton, 2014; Bloom et al., 2007; Raudenbush & Bryk, 2002). Interference is the threat specific to social patterns: when a student's outcome depends on classmates' treatment, randomising by user biases the estimate, and cluster or graph-cluster randomisation reduces the bias at a cost in power (Hudgens & Halloran, 2008; Aronow & Samii, 2017; Ugander et al., 2013; Eckles et al., 2017). Multiple testing and the credibility of findings add the last layer: false discovery control (Benjamini & Hochberg, 1995), Type S and Type M errors in underpowered studies (Gelman & Carlin, 2014), researcher degrees of freedom (Simmons et al., 2011) and the reproducibility record (Ioannidis, 2005; Open Science Collaboration, 2015). Meta-analysis across sites uses random-effects models and heterogeneity statistics (DerSimonian & Laird, 1986; Higgins & Thompson, 2002; Cohen, 1988).

Ethics, dark patterns and the regulatory context

The ethics of nudging has been argued at length, with the consensus that nudges must preserve choice, be transparent and serve the nudged person's own ends (Sunstein, 2015; Thaler & Sunstein, 2008). Dark patterns are the designs that fail those tests: interfaces that trick, pressure or obstruct users (Gray et al., 2018; Mathur et al., 2021), found at scale in commerce (Mathur et al., 2019) and now the subject of U.S. enforcement under the Federal Trade Commission Act and of the European Digital Services Act's prohibition of deceptive interface design (Federal Trade Commission, 2022; U.S. Congress, 2023; European Parliament and Council, 2022). Notifications are the medium through which most campus patterns act, and the human-computer interaction literature documents their cost (Sumaiya, 2024; Taher et al., 2024): interruptions, overload and the receptivity that depends on timing and content (Pielot et al., 2014; Sahami Shirazi et al., 2014; Mehrotra et al., 2016; Monge Roffarello & De Russis, 2019). For students, two further frameworks apply: the Family Educational Rights and Privacy Act governs education records and the uses to which they may be put (U.S. Department of Education, n.d.), and the

Common Rule governs research with human subjects, so that an experiment on students may require review even when a product change would not (U.S. Department of Health and Human Services, 2018; Kramer et al., 2014; Meyer et al., 2019). Accessibility guidelines constrain the surfaces themselves (World Wide Web Consortium, 2023), and the EU's data-protection and AI regulations apply where the product reaches European students (European Parliament and Council, 2016; European Parliament and Council, 2024). The testbed treats harms as metrics with the same status as engagement, which is the practical form of these constraints.

Agent-based simulation as an evaluation instrument

Agent-based models generate population behaviour from individual rules and are used to study interventions whose effects run through heterogeneity and interaction (Epstein, 1999; Macal & North, 2010); their verification and validation follow the general discipline of simulation (Law, 2015; Sargent, 2013). The present testbed is used not to predict what a pattern will do on a real campus, which it cannot, but to provide a population in which every effect is known so that evaluation methods can be scored, in the way that synthetic benchmarks are used elsewhere. Table 1 summarises the strands and the gap this article addresses.

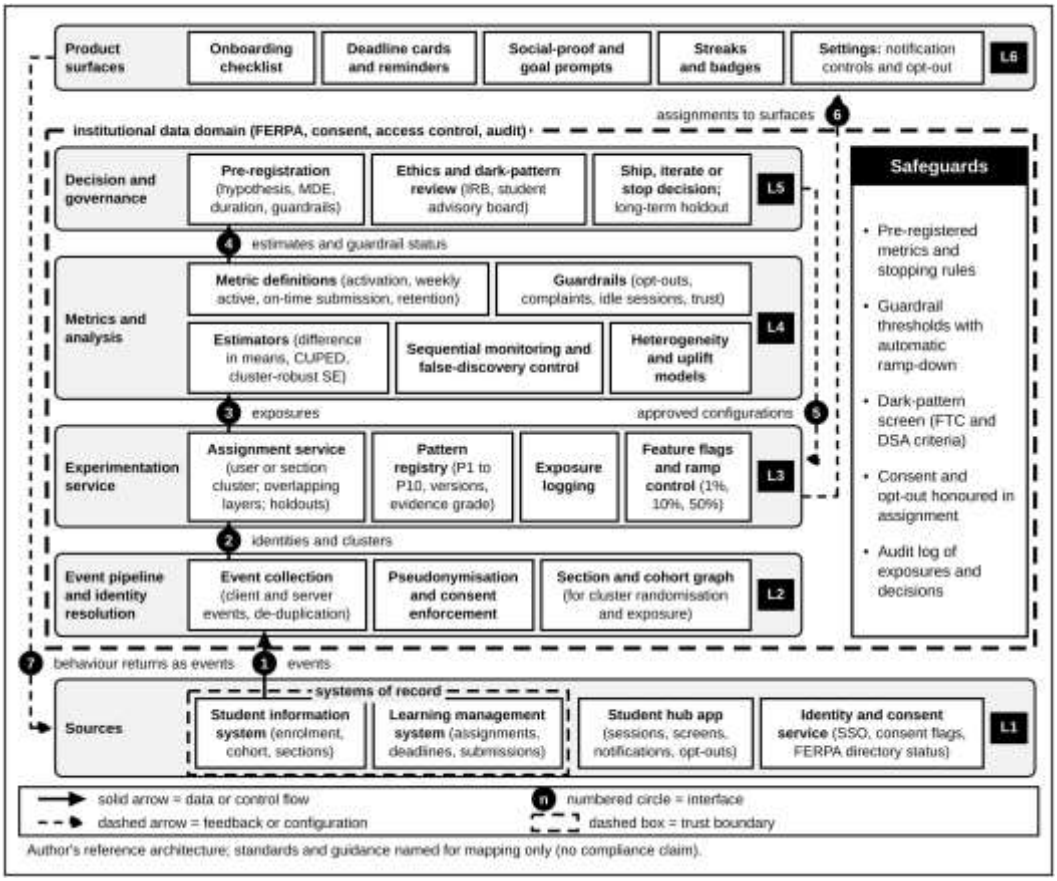
Table 1: Literature strands, what they establish, representative sources and the gap addressed in this article

Strand	What the literature establishes	Representative sources	Gap addressed here
Behavioural economics and choice architecture	Defaults, reminders, social norms, planning prompts and framing move behaviour; effects are real but smaller and less durable than early reports	Thaler & Sunstein (2008); Johnson & Goldstein (2003); Madrian & Shea (2001); Allcott (2011); Karlan et al. (2016); Allcott & Rogers (2014); DellaVigna & Linos (2022); Maier et al. (2022)	The same mechanisms implemented as product patterns and compared on one population with known truth
Nudging in education	Text nudges help in targeted populations; large field experiments on goals and reminders in college find small or null effects; effects depend on context	Castleman & Page (2015, 2016); Dobronyi et al. (2019); Oreopoulos & Petronijevic (2019); Bird et al. (2021); Kizilcec et al. (2020); Damgaard & Nielsen (2018)	Why small effects are hard to measure in campus products and what design recovers them
Engagement, habits and gamification	Engagement is multidimensional and software measures a proxy; habits form over weeks; game elements can help or crowd out motivation	Fredricks et al. (2004); Henrie et al. (2015); Lally et al. (2010); Deterding et al. (2011); Sailer & Homner (2020); Hanus & Fox (2015); Deci et al. (1999)	Short-run and long-run effects separated; crowding-out and harm metrics measured alongside engagement
Online controlled experiments	Trustworthy experiments need pre-registration, guardrails, variance reduction, sequential methods and attention to novelty and metric pitfalls	Kohavi et al. (2009, 2012, 2020); Deng et al. (2013); Hohnhold et al. (2015); Johari et al. (2017); Dmitriev et al. (2017); Gupta et al. (2019)	The methods scored against known truth for the specific case of campus products
Causal inference, interference and heterogeneity	Randomisation by user is biased under interference; cluster designs reduce bias; heterogeneous effects and uplift targeting are estimable with machine learning	Rubin (1974); Imbens & Rubin (2015); Hudgens & Halloran (2008); Eckles et al. (2017); Ugander et al. (2013); Wager & Athey (2018); Künzel et al. (2019); Athey et al. (2019)	Interference through course sections quantified; a targeting policy evaluated against treating everyone
Credibility of findings	Multiple testing, peeking and low power produce false and inflated findings	Benjamini & Hochberg (1995); Gelman & Carlin (2014); Simmons et al. (2011); Ioannidis (2005); Open Science Collaboration (2015)	False-discovery and Type S/M errors shown for ten patterns and six metrics at campus sample sizes
Ethics, dark patterns and regulation	Nudges must preserve choice and transparency; dark patterns are enforceable; student data and experiments have their own rules	Sunstein (2015); Gray et al. (2018); Mathur et al. (2019, 2021); Federal Trade Commission (2022); Kramer et al. (2014); Meyer et al. (2019); Slade & Prinsloo (2013)	Harms as guardrail metrics; a workflow with ethics and student review; a governance roadmap

Evaluation framework

Figure 1 shows the experimentation platform as six layers inside a campus product. The sources are the systems of record (student information system, learning management system), the student hub application itself and the identity and consent service that carries single sign-on, consent flags and FERPA directory status. The event pipeline collects and de-duplicates client and server events, pseudonymises them, enforces consent, and maintains the section and cohort graph that cluster randomisation and exposure analysis need (Asheke & Maksudul, 2022; Asheke & Badrul, 2022; Akter & Sumaiya, 2022). The experimentation service assigns units to arms (by user or by section cluster, in overlapping layers, with holdouts), keeps a registry of patterns with versions and evidence grades, logs exposures and controls ramps (Imran & Jannatul, 2022; Maksudul & Islam, 2022; Morshedul & Islam, 2022). The metrics and analysis layer defines the metrics and guardrails, runs the estimators (difference in means, CUPED, cluster-robust standard errors), monitors sequentially with false-discovery control, and fits heterogeneity and uplift models (Stabak & Shahid, 2022; Feroz & Pervaz, 2022; Feroz & Shabbir, 2022). The decision and governance layer holds pre-registration, the ethics and dark-pattern review with the student advisory board, and the ship, iterate or stop decision with a long-term holdout (Islam & Md, 2023; Asheke, 2023; Shahid, 2023). The product surfaces (onboarding checklist, deadline cards and reminders, social-proof and goal prompts, streaks and badges, and the settings that let a student turn notifications off) are where the patterns live (Adnan, 2024; Islam & Ahmed, 2023; Sadia & Mosheur, 2023). Two trust boundaries and a safeguards panel make the ethics operational: pre-registered metrics and stopping rules, guardrail thresholds with automatic ramp-down, a dark-pattern screen using the FTC and DSA criteria, consent and opt-out honoured in assignment, and an audit log of exposures and decisions (Kohavi et al., 2020; Tang et al., 2010; Federal Trade Commission, 2022; European Parliament and Council, 2022; U.S. Department of Education, n.d.).

Figure 1: Experimentation platform reference architecture for a campus software product: six layers with numbered interfaces, trust boundaries and safeguards



Behavioural state model and pattern touchpoints

Figure 2 places the patterns on a state model of the student. A student who has not logged in moves to a first login under enrolment email or deadline pressure, then to an activating state in which the onboarding steps are completed (profile, calendar sync, notifications, first in-app submission, advising booking); completing the three key steps within 14 days is activation. From there the student is either in habitual use, in which the app is opened most days and habit strength grows, or dormant, opening only under deadline pressure while habit decays; a dormant student who does not open for three weeks has churned, and a churned student rarely returns. A side state, opted out of notifications, is entered when accumulated annoyance exceeds the student's tolerance, and it removes the reminder channel for the rest of the term. Each pattern acts on specific transitions: the checklist and friction reduction on first login to activating; defaults, timed reminders, loss framing and personalised timing on the reminder transitions and on annoyance; social proof and urgency on dormant to habitual; goals on activating to habitual; streaks on the daily self-loop of habitual use. Table 2 gives the catalogue with the mechanism, the behaviour change technique, the evidence and the testbed implementation of each pattern.

Figure 2: Behavioural state model of a student user with the ten pattern touchpoints: UML-style state machine with guards, effects and the opted-out side state

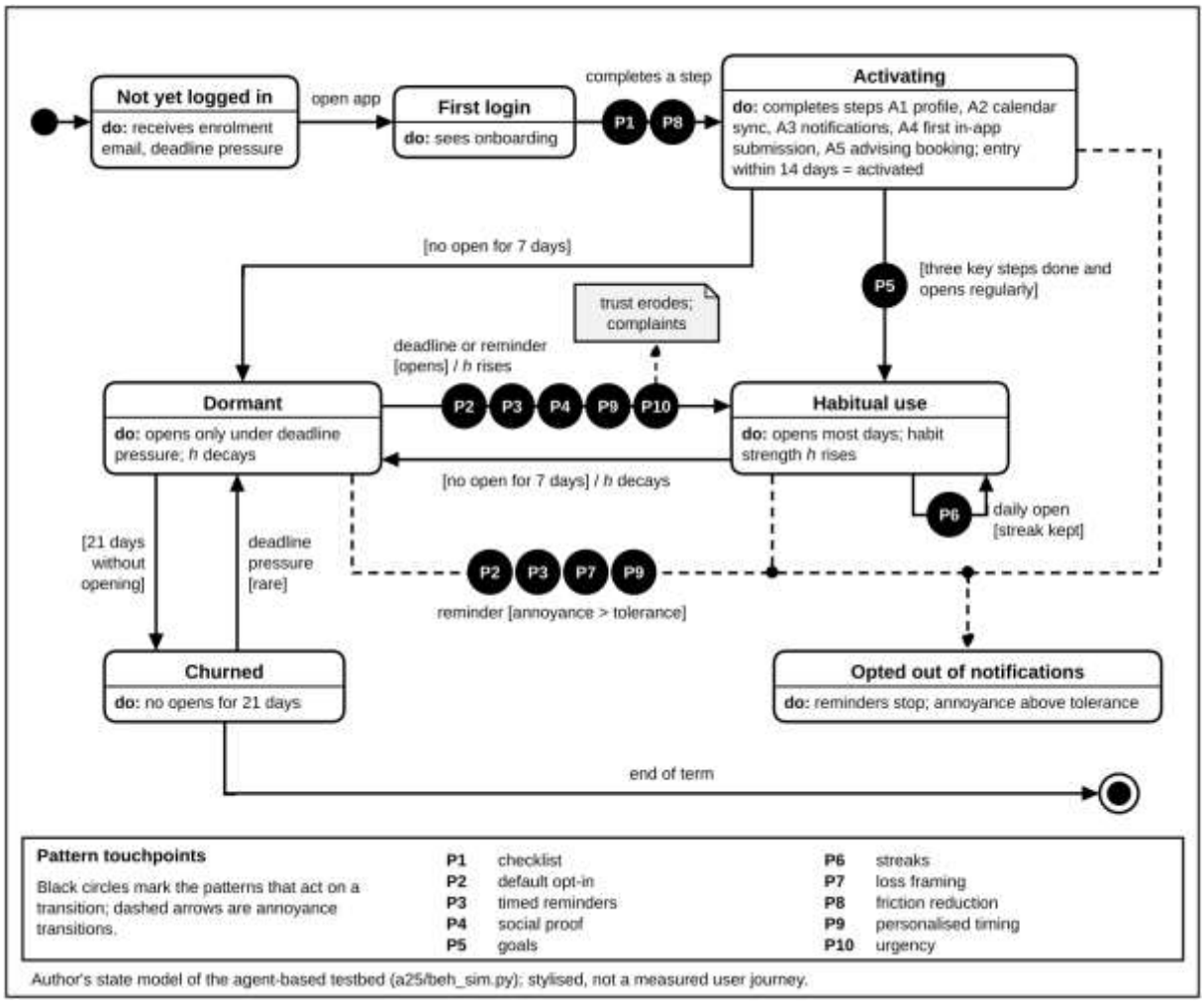


Table 2: Pattern catalogue: mechanism, behaviour change technique, evidence from the literature and implementation in the testbed

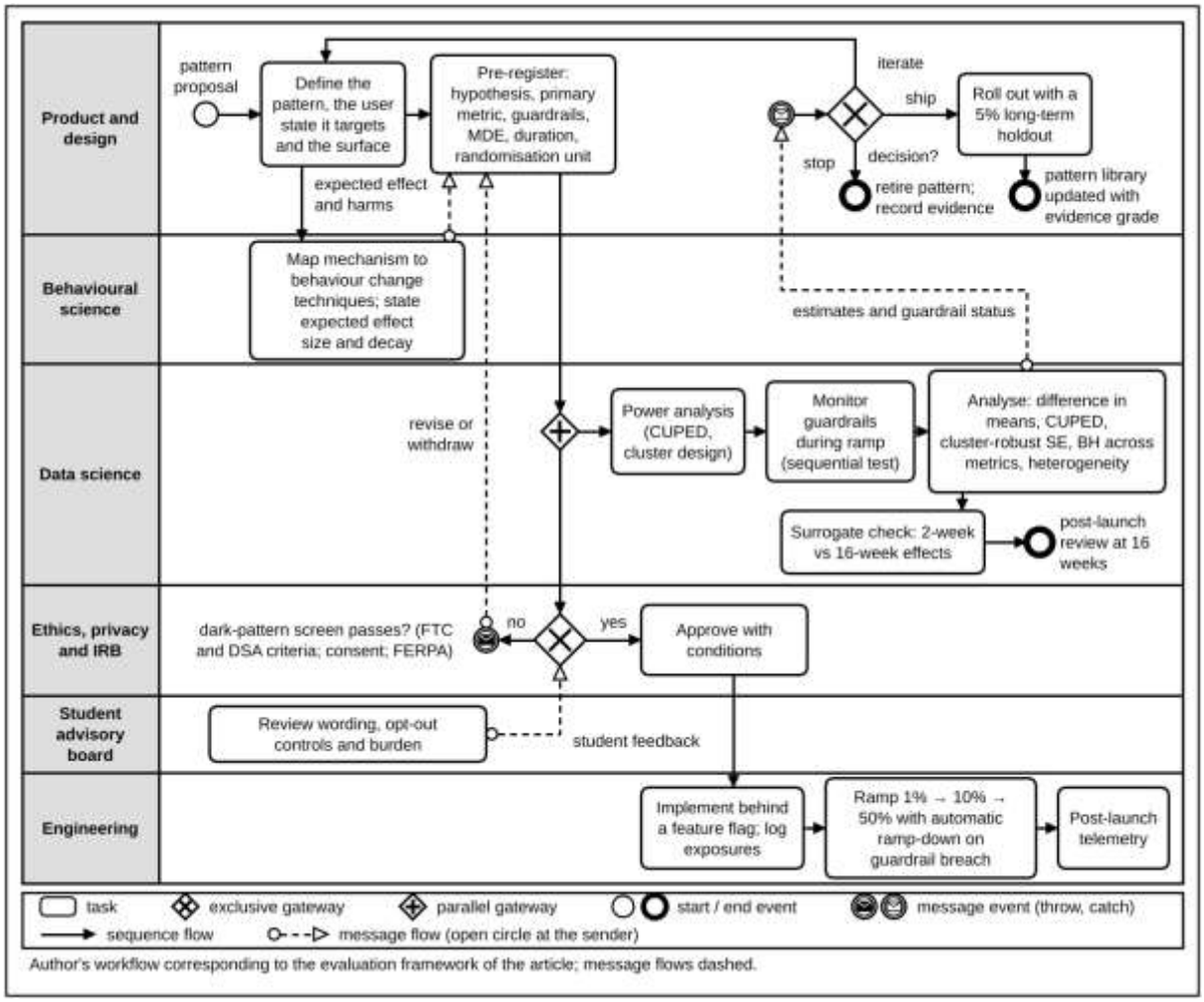
Pattern	Mechanism and technique	Evidence from the literature	Implementation in the testbed
P1 Onboarding checklist with progress bar	Goal gradient; feedback on progress (BCT: self-monitoring, goal setting)	Goal-gradient effects in loyalty programmes (Kivetz et al., 2006); no large field trials in campus software	Perceived friction of every step cut by 15%; completion probability rises by 4 points per step done
P2 Default opt-in to deadline reminders	Inertia and defaults (BCT: restructuring the environment)	Defaults move organ donation and saving (Johnson & Goldstein, 2003; Madrian & Shea, 2001)	Reminder coverage 90% instead of 35%; opt-out hazard 1.3x for the unasked
P3 Timed deadline reminders	Attention and memory; salience at the right time (BCT: prompts and cues)	Reminders raise saving and vaccination (Karlan et al., 2016; Milkman et al., 2011); receptivity depends on timing (Mehrotra et al., 2016)	Reminder effect on opening 1.5x; annoyance per reminder 0.9x
P4 Social proof on the deadline card	Descriptive norms (BCT: social comparison)	Norms move reuse and energy use; boomerang below the norm (Goldstein et al., 2008; Schultz et al., 2007; Allcott, 2011)	Peer-activity norm amplified 1.6x for those shown; shown only when at least 50% of the section is active; backfires below 35%
P5 Goal setting and planning prompt	Implementation intentions (BCT: action planning)	Strong in some domains (Gollwitzer, 1999; Milkman et al., 2011); small or null for college grades (Dobronyi et al., 2019; Clark et al., 2020)	Intent-driven action boosted by 0.05 for three weeks
P6 Streaks and badges	Loss aversion over a streak; rewards (BCT: rewards, feedback)	Mixed; decay and crowding-out of intrinsic motivation (Hamari et al., 2014; Sailer & Homner, 2020; Hanus & Fox, 2015; Deci et al., 1999)	Streak utility up to 0.6; novelty 1.5x; from week 6 high-intent students lose 10% of excess motivation
P7 Loss-framed reminders	Loss aversion (BCT: framing)	Loss aversion is robust (Tversky & Kahneman, 1991); framing effects in megastudies small (Milkman et al., 2021a)	Reminder effect 1.15x; annoyance 1.15x
P8 Friction reduction (SSO, one-tap sync)	Removing sludge (BCT: restructuring the environment)	Simplification outperforms framing (Bhargava & Manoli, 2015; Sunstein, 2022)	Friction of the first two steps cut by 60%
P9 Personalised timing (uplift targeting)	Targeting by predicted response; receptivity (BCT: tailoring)	Uplift modelling and bandits (Gutierrez & Gérardy, 2017; Li et al., 2010); heterogeneity of nudge effects (Yeager et al., 2019)	60% of weak-habit students receive reminders with 1.3x effect and 0.6x annoyance
P10 Urgency countdown with manipulative framing	Scarcity and pressure (dark-pattern adjacent)	Urgency and pressure are dark patterns (Mathur et al., 2019, 2021; Federal Trade Commission, 2022)	Opening bonus 0.6 near deadlines for four weeks, scaled by trust; trust falls 0.03 per exposure; complaints when trust is low

Experiment lifecycle

Figure 3 shows the lifecycle as swimlanes. Product and design define the pattern, the state it targets and the surface, and pre-register the hypothesis, the primary metric, the guardrails, the minimum detectable effect, the duration and the randomisation unit. Behavioural science maps the mechanism to

techniques and states the expected effect and its decay. Data science does the power analysis, monitors guardrails during the ramp, analyses with the pre-registered estimators and corrections, and checks short-term against long-term effects. Ethics, privacy and IRB screen for dark patterns, consent and FERPA and approve with conditions; the student advisory board reviews wording, opt-out controls and burden. Engineering implements behind a feature flag, logs exposures, ramps from 1% to 10% to 50% with automatic ramp-down on a guardrail breach, and keeps post-launch telemetry. The decision gate ships with a 5% long-term holdout, iterates or stops, and the pattern library records the evidence grade. The workflow is the operational form of the guidance on trustworthy experiments and on the ethics of nudging (Kohavi et al., 2020; Sunstein, 2015; Meyer et al., 2019).

Figure 3: Experiment lifecycle workflow: BPMN-style swimlanes for product and design, behavioural science, data science, ethics and privacy, the student advisory board and engineering

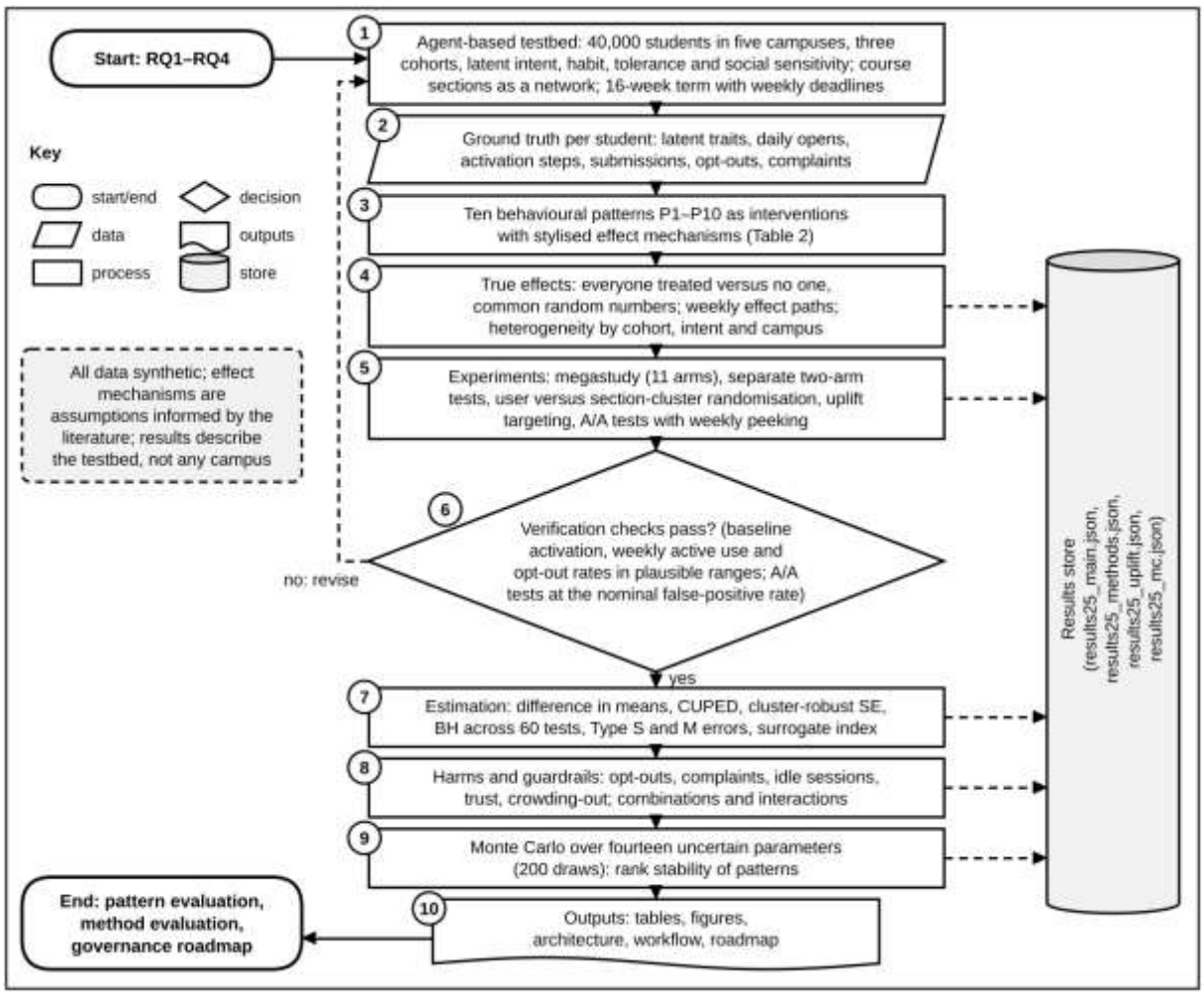


Methods

Study design

Figure 4 shows the study as a flowchart. An agent-based testbed generates the daily behaviour of 40,000 students over a term. Ten patterns are implemented as interventions with stylised mechanisms. True effects are computed by treating everyone and no one with common random numbers, which no real experiment can do, and are decomposed by week, cohort, intent and campus. Experiments are then run on the same population in the ways a product team would run them (a megastudy, separate two-arm tests, user and section-cluster randomisation, uplift targeting, A/A tests with weekly peeking), and the estimates are scored against the truth. Harms are measured as metrics with the same status as engagement, combinations are tested for interaction, and a Monte Carlo analysis varies fourteen assumed parameters jointly.

Figure 4: Study design: testbed, patterns, true effects, experiments, verification, estimation, harms, Monte Carlo and outputs, drawn with standard flowchart symbols



All data are synthetic. No student, campus or product data were used. The behavioural parameters are stylised values chosen so that the baseline behaves plausibly (activation, weekly active use and opt-out rates in ranges a campus product might see) and so that each pattern's mechanism and rough effect size follow the literature in Table 2; they are assumptions, and the Monte Carlo analysis varies them. Results describe the testbed.

Testbed

Table 3 gives the testbed. Students have a campus (five, with campus-level offsets), a cohort (first-year, continuing undergraduate, graduate) and latent traits: intent (motivation) on a logistic scale, an initial habit strength, a notification tolerance, a social sensitivity and a pre-period covariate (opening days per week in the prior term, drawn as a count from the student's prior-term opening probability). Each student belongs to four course sections of about thirty students in the same campus; sections post one assignment per week during the 16-week term, due on a section-specific weekday. Each day a student opens the app with a probability that is a logistic function of intent, habit, deadline proximity, reminders received, the activity of section peers (the interference channel), a novelty term that decays over the first weeks, streak and goal terms when the corresponding patterns are active, and accumulated annoyance. On opening, the student completes onboarding steps with probabilities that fall with each step's friction and rise with intent, submits pending assignments (with a higher probability near the deadline), or idles. A student who has completed the calendar sync sees deadlines more saliently (the deadline term is multiplied by 1.5), which is the only channel in the testbed through which activation changes later engagement; it is an assumption, and the Monte Carlo analysis varies it from no link to a doubling. Habit strength rises on days with an open and decays otherwise; a streak

counts consecutive opening days; a student who has not opened for 21 days is churned. Reminders are sent the day before a deadline to students who have opted in; each adds annoyance that decays over time, and a student whose annoyance exceeds tolerance opts out with a daily hazard, which removes reminders for the rest of the term; a student far above tolerance may complain. Trust falls under manipulative urgency cues and recovers slowly.

Table 3: Testbed: population, traits, sections and calendar, daily behaviour, activation, reminders and annoyance, and the metrics recorded

Element	Setting in the testbed
Population	40,000 students in five campuses (campus offsets on intent, sd 0.3); cohorts 30% first-year, 50% continuing undergraduate, 20% graduate
Latent traits	Intent (logistic of a normal with sd 1 plus cohort and campus offsets); initial habit 0.15 + 0.25 x intent + noise; notification tolerance mean 1.0 (sd 0.4); social sensitivity mean 1.0 (sd 0.4); pre-period covariate = opening days per week in a 16-week prior term, a Poisson count from the prior-term opening probability
Sections and calendar	Four sections of about 30 students per student within a campus (5,324 sections); 126 days: two weeks before term and a 16-week term; one assignment per section per week, due Tuesday to Sunday
Daily opening	Logistic in: intercept -4.2, 2.4 x intent, 2.2 x habit, 0.9 if a deadline is within two days (x 1.5 once the calendar is synced, the only link from activation to later engagement), 1.0 per reminder, 0.5 x social sensitivity x (peer activity - 0.4), novelty 0.9 x exp(-day/7), streak and goal terms, -0.6 x annoyance, -1.0 if churned
Onboarding steps	Profile, calendar sync, notifications (requires opt-in), first in-app submission (requires a pending assignment), advising booking (from week 2); per-open completion probability (0.25 + 0.45 x intent) x (1 - friction), friction 0.35, 0.55, 0.45, 0.30, 0.65
Submissions	On opening, a pending assignment is submitted with probability 0.6 + 0.3 x intent when due within two days (earlier with probability 0.3 x intent); unsubmitted assignments are missed at the deadline
Habit, streaks, churn	Habit rises by 0.05 x (1 - habit) on opening days and decays 3% otherwise; streak counts consecutive opening days; no open for 21 days = churned
Reminders and annoyance	Reminder the day before a deadline to opted-in students (35% opt in when asked); annoyance +0.12 per reminder, decays 8% per day; opt-out hazard 0.5 per day while annoyance exceeds tolerance; complaint hazard 0.1 per day while annoyance exceeds 1.6 x tolerance
Metrics recorded	Activation within 14 days, weekly active use by week, opens per term, on-time in-app submissions, retention in the last two weeks, opt-outs, complaints, idle opens, trust, first-login day
Seeds and scale	Five population seeds for the main comparison; common random numbers for true effects; 10,000 students per Monte Carlo draw

Patterns

Table 2 lists the ten patterns and Table 4 their assumed parameters. P1, the onboarding checklist with a progress bar, lowers the perceived friction of every onboarding step by 15% and adds a goal-gradient term that raises completion probability with each step done. P2, default opt-in to deadline reminders, raises reminder coverage from 35% (the share who opt in when asked) to 90%, with a higher opt-out hazard because many more students receive reminders they did not ask for. P3, timed deadline reminders (the day before, in the evening), multiplies the effect of a reminder on opening by 1.5 with slightly lower annoyance per reminder. P4, social proof on the deadline card ("x% of your section checked their deadlines this week"), amplifies the peer-activity norm for students who see it, mostly moves students who have been inactive, is shown only when the norm is favourable (at least half of the section active that week), and backfires when it is shown against an unfavourable norm; a naive always-on variant is tested as an ablation. P5, goal setting with a planning prompt in the first week, adds a

small boost to intent-driven action for three weeks. P6, streaks and badges, adds a utility for keeping a streak, multiplies the novelty term, and from week six crowds out part of the intrinsic motivation of high-intent students. P7, loss-framed reminders, raises the reminder effect and the annoyance per reminder by 15%. P8, friction reduction (single sign-on and one-tap calendar sync), cuts the friction of the first two onboarding steps by 60%. P9, personalised timing, applies to 60% of students with weak habits (the output of an uplift model, stylised) a 30% larger reminder effect with 40% less annoyance. P10, an urgency countdown with manipulative framing, adds a large opening bonus near deadlines for four weeks that is scaled by trust, erodes trust on every exposure, and raises complaints when trust is low. Every pattern's additive and multiplicative effects are scaled by a common factor in the Monte Carlo analysis.

Table 4: Assumed effect parameters of the ten patterns in the testbed, their literature anchors and the Monte Carlo ranges

Parameter	Default	Monte Carlo range	Literature anchor (stylised)
P1 friction multiplier; goal-gradient increment	0.85; 0.04	0.6 to 1.0	Goal gradient (Kivetz et al., 2006)
P2 reminder coverage; opt-out hazard multiplier	0.90; 1.3	0.6 to 0.95	Defaults (Johnson & Goldstein, 2003)
P3 reminder effect multiplier; annoyance multiplier	1.5; 0.9	1.0 to 2.0	Reminders and timing (Karlan et al., 2016; Mehrotra et al., 2016)
P4 norm amplification; backfire threshold; gating threshold	1.6; 0.35; 0.50	1.0 to 2.5	Norms and boomerang (Schultz et al., 2007)
P5 intent boost; weeks	0.05; 3	0 to 0.12	Small, transient college effects (Dobronyi et al., 2019)
P6 streak utility; novelty multiplier; crowding-out share	0.6; 1.5; 0.10	0.2 to 1.0	Gamification and crowding-out (Sailer & Homner, 2020; Deci et al., 1999)
P7 reminder effect and annoyance multipliers	1.15; 1.15	fixed	Loss framing small in megastudies (Milkman et al., 2021a)
P8 friction multiplier on steps 1 and 2	0.40	0.2 to 0.8	Simplification (Bhargava & Manoli, 2015)
P9 effect and annoyance multipliers; targeted share	1.3; 0.6; 0.6	1.0 to 1.6	Tailoring and uplift (Gutierrez & Gérardy, 2017)
P10 opening bonus; trust erosion; weeks	0.6; 0.03; 4	0.2 to 1.0	Dark patterns (Mathur et al., 2019)
Annoyance per reminder; tolerance mean	0.12; 1.0	0.06 to 0.2; 0.6 to 1.4	Notification overload (Pielot et al., 2014; Sahami Shirazi et al., 2014)
Interference strength; novelty	1.0; 0.9	0 to 2; 0.3 to 1.5	Peer effects and novelty (Eckles et al., 2017; Kohavi et al., 2012)
Activation link (deadline salience after calendar sync)	0.5	0 to 1.0	Activation as a leading indicator (Fader & Hardie, 2005; Lemon & Verhoef, 2016)

Metrics

Table 5 defines the metrics. Activation is the completion of the profile, calendar sync and first in-app submission within 14 days of first login (or of the start of term, for students who log in before it). Weekly active use is the share of students with at least one open in a given term week; no assignment is posted in the final week, so week-16 activity reflects habitual rather than deadline-driven use; opens per term is the count of opening days; on-time submission is the share of assignments submitted through the app before the deadline; retention is activity in the last two weeks of term. Harms are the opt-out rate, complaints per student, idle opens (opens with no deadline and no action), and end-of-term trust.

Effects are reported in percentage points for rates and in units for counts. The true effect of a pattern is the difference in population means between everyone treated and no one treated, simulated with common random numbers so that the only difference is the pattern.

Table 5: Metrics: engagement, activation, academic proxy, retention and harm metrics, their definitions and their role as primary, secondary or guardrail

Metric	Definition	Role
Activation (14 days)	Profile, calendar sync and first in-app submission completed within 14 days of first login (or of the start of term)	Primary activation metric
Weekly active use (WAU, week w)	Share of students with at least one open in term week w; reported for weeks 2, 4, 8 and 16	Primary engagement metric; its time path shows novelty and decay
Opens per term	Number of opening days during the term	Secondary engagement metric
On-time submission	Share of assignments submitted through the app before the deadline	Academic proxy
Retention	At least one open in the last two weeks of term	Long-term engagement
Opt-out rate	Share of students who turned notifications off during the term	Guardrail (harm)
Complaints	Complaints per 100 students (annoyance far above tolerance, or low trust under urgency cues)	Guardrail (harm)
Idle opens	Opens with no deadline within two days and no action taken, per student	Guardrail (unproductive engagement)
Trust	End-of-term trust index (1 = intact)	Guardrail (harm)
Crowding-out	Change in weekly active use among the top intent quartile after week 6	Guardrail (motivation)

Experiments and estimators

Table 6 lists the experiment designs and Table 7 the estimators and checks. The megastudy assigns each student to one of eleven arms (control and ten patterns, about 3,660 per arm) in the same term and compares each pattern with the shared control. Separate two-arm experiments assign half the population to each pattern in turn. For the two social patterns (P4 and P6), user-level randomisation is compared with randomisation by course section at three interference strengths, against the true effect, with cluster-robust standard errors for the cluster design. Uplift targeting trains a T-learner (two gradient-boosted regressors) on half of a two-arm experiment using noisy proxies of the latent traits, targets the half of the other half with the highest predicted uplift, and compares the realised effect of the policy with treating everyone and with treating a random half. A/A tests (no treatment) with weekly peeking measure the false-positive rate of naive repeated testing. Power, Type S and Type M errors are computed by the retrodesign method of Gelman and Carlin (2014) from the true effect of each pattern and the outcome standard deviation and pre-period correlation observed in the baseline term, at sample sizes from 250 to 16,000 per arm, with and without CUPED. Multiple testing is examined by drawing estimates around the true effects of the ten patterns on six metrics at 2,000 per arm (200 replications) and applying the Benjamini-Hochberg procedure Gelman and Carlin (2014). Combinations of patterns are simulated to measure interactions, and the true effects are decomposed by campus to compute a random-effects meta-analysis with the I-squared statistic. The Monte Carlo analysis draws 200 combinations of fourteen parameters (the pattern effect parameters, annoyance per reminder, tolerance, interference strength, novelty and the activation link) from uniform ranges, on populations of 10,000, and records the true effects of every pattern on seven metrics.

Table 6: Experiment designs: unit, arms, sample sizes, duration and what each design is used to show

Design	Unit and arms	Sample	Duration	What it shows
True effects	Everyone treated vs no one, common random numbers	40,000 x 5 seeds	Whole term	The target of estimation; weekly paths; heterogeneity by cohort, intent and campus
Megastudy	User; 11 arms (control + P1-P10)	About 3,660 per arm	Whole term	Comparable estimates of all patterns in one term; precision at shared-control size
Separate two-arm tests	User; 2 arms per pattern	20,000 per arm	Whole term	Precision at large scale; CUPED gains; source for subsampling
Interference designs	User vs section cluster; P4 and P6; interference 0.5, 1, 2	20,000 per arm; about 2,660 clusters	Whole term	Bias of user-level randomisation under spillovers; cost of cluster design in precision
Uplift targeting	User; T-learner trained on half; policy evaluated on the other half (P2, P3, P9)	20,000 per arm; 3 seeds	Whole term	Value of targeting against treating everyone and a random half
A/A tests with peeking	User; no treatment; 40 replications; weekly looks	20,000 per arm	14 weekly looks	False-positive rate of naive repeated testing
Retrodesign	True effect, outcome standard deviation and pre-period correlation from the baseline term; 250 to 16,000 per arm	20,000 draws per case	Whole term	Power, Type S and Type M errors with and without CUPED
Multiple testing	10 patterns x 6 metrics at 2,000 per arm; estimates drawn around the true effects; 200 replications	Derived	Whole term	Raw vs Benjamini-Hochberg significance against known truth
Combinations	Everyone treated with two or three patterns	40,000 x 5 seeds	Whole term	Interactions (combined effect minus the sum of single effects)
Multi-campus meta-analysis	True effects by campus; random-effects pooling	5 campuses x 5 seeds	Whole term	Between-campus heterogeneity (I-squared)
Monte Carlo	True effects under 200 draws of 14 parameters	10,000 per draw	Whole term	Stability of the pattern ranking; harms under uncertainty

Table 7: Estimators and checks: difference in means, CUPED, cluster-robust standard errors, sequential monitoring, false-discovery control, Type S and M errors, surrogates, heterogeneity and uplift

Estimator or check	Definition	Use
Difference in means	Treated mean minus control mean with a Welch standard error	All user-level designs
CUPED	Outcome adjusted by its regression on the pre-period covariate (opens per week in the prior term) before the difference in means (Deng et al., 2013)	Variance reduction; reported as the ratio of standard errors
Cluster-robust standard error	Between-cluster variance of cluster means, weighted by cluster size, with a G/(G - 1) correction	Section-cluster randomisation
Sequential monitoring	Weekly looks at the running estimate; naive significance at any look versus a single pre-registered look	A/A false-positive rate; motivates always-valid methods (Johari et al., 2017)
Benjamini-Hochberg	Step-up procedure at q = 0.05 over 60 p-values (10 patterns x 6 metrics)	False discovery control
Type S and Type M errors	Among significant estimates: share with the wrong sign; mean absolute estimate over the true absolute effect; computed by retrodesign from the true effect and the standard error at each sample size (Gelman & Carlin, 2014)	Small-sample credibility
Surrogate check	Correlation across patterns between the week-2 or week-4 effect and the week-16 effect	Whether early metrics predict long-term effects (Athey et al., 2019)
Heterogeneity	True effects by cohort, intent quartile and campus; DerSimonian-Laird pooling with I-squared across campuses	Generalisation and targeting
Uplift (T-learner)	Two gradient-boosted regressors (treated and control) on noisy trait proxies; predicted uplift = difference; policy treats the top half	Personalisation value (Künzel et al., 2019)
Interaction	Combined-pattern effect minus the sum of the single-pattern effects	Whether patterns add, overlap or conflict

Verification and reproducibility

Three checks were applied before the experiments. In the baseline term, 38% of students activated within 14 days, weekly active use was 72% at mid-term and 45% in the final week (in which no assignment is due), 46% of assignments were submitted on time through the app, 7% of students opted out of notifications, 12% churned and 2% never logged in, which are within the ranges a campus product might report and were taken as plausible rather than calibrated to any source. A/A tests with a single look at the end of term produced significant results at or below the nominal rate (1 of 40). The megastudy estimates with about 3,660 students per arm bracketed the true activation effects of the checklist and of friction reduction within their 95% intervals in every seed, and the true effects of every pattern on the three primary metrics in at least four of five seeds. The simulation, the runner and the result files are in the supplementary archive; fixed seeds reproduce every number (Law, 2015; Sargent, 2013).

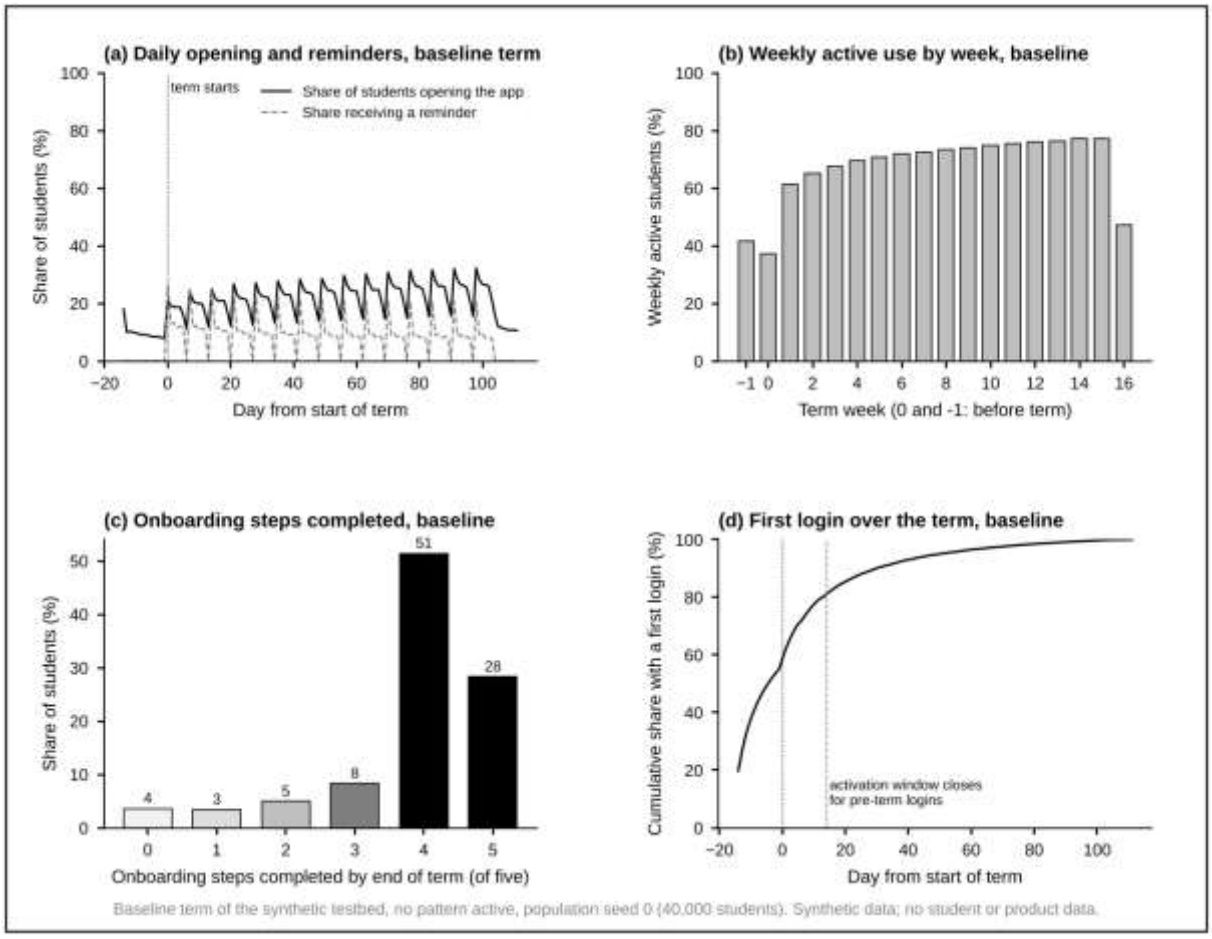
Results

Testbed behaviour

Figure 5 shows the baseline term for population seed 0 with no pattern active. Daily opening runs at about 10% of students in the two weeks before term and about 22% during the term, with peaks of up to 32% on days before deadlines and troughs of 11% between them; reminders reach up to a quarter of students on the day before a section's deadline. Weekly active use rises from 61% in the first term week to 77% in week 15 as habits form and the early non-users log in, and falls to 47% in week 16, the week

with no assignments due, which is why week 16 is used as the measure of habitual rather than deadline-driven use. By the end of term 51% of students have completed four of the five onboarding steps and 28% all five; 12% have completed two or fewer. A fifth of students log in on the first day the app is available, 59% by the start of term, 81% within two weeks of it and 98% by the end; 2% never log in. The baseline metrics in the first row of Table 8 (activation 38%, weekly active use 68% in week 4 and 45% in week 16, 46% of assignments submitted on time through the app, 80% retained, 7% opted out, 3.2 idle opens per student) are the levels against which the pattern effects are measured.

Figure 5: Baseline behaviour of the testbed: daily opening and reminders, weekly active use, onboarding steps completed and first logins over the term (seed 0, no pattern active)



True effects of the ten patterns

Table 8 gives the true effect of each pattern on every metric, averaged over five population seeds, and Figure 6 shows the four primary metrics with the experimental estimates overlaid. Three groups emerge. The onboarding patterns move activation and little else: friction reduction (P8) raises activation by 11.9 percentage points and the checklist (P1) by 11.2 points, both with standard deviations across seeds of 0.2 points or less, while their effects on later engagement run through the assumed salience link only, at 0.5 to 2.1 points on weekly active use, 0.9 to 1.5 points on on-time submission and 1.1 to 1.6 points on retention. The deadline and habit patterns move engagement and academic behaviour: gated social proof (P4) has the largest effects on on-time submission (6.1 points), retention (4.7 points) and week-16 active use (2.1 points) despite a negligible effect on activation (1.0 point); timed reminders (P3) add 3.8 points to on-time submission, 3.0 to retention and 1.8 to week-16 use; personalised timing (P9) adds 2.2, 2.0 and 1.4 points; default opt-in (P2) adds 5.1 points to on-time submission and 3.1 to retention but lowers week-16 use by 0.9 points, because the students it annoys into opting out lose the reminder channel and some of their habit. The novelty patterns move early engagement: the urgency countdown (P10) raises week-2 active use by 13.1 points and activation by 8.6 points, goals (P5) and

streaks (P6) raise week-2 use by 4.7 and 3.5 points and activation by 3.3 and 3.1 points, but their week-16 effects are 1.3, 0.5 and 0.2 points. Loss framing (P7) does almost nothing on any engagement metric (0.4 points on activation, 0.5 on week-4 use) while raising opt-outs by 2.4 points. Opens per term rise by 3.1 under urgency, 2.0 to 2.1 under the two reminder patterns and 1.7 under social proof, and the patterns that raise opens also raise idle opens (0.2 to 0.4 per student), most under streaks, which reward opening regardless of purpose. The ranking of patterns differs by metric: on activation it is P8, P1, P10, P2; on week-16 use it is P4, P3, P9, P10; on on-time submission it is P4, P10, P2, P3; and the rank correlation between the activation ranking and the week-16 ranking across the ten patterns is slightly negative.

Figure 6: True effects of the ten patterns on activation, week-16 active use, on-time submission and retention, with the megastudy and separate two-arm estimates

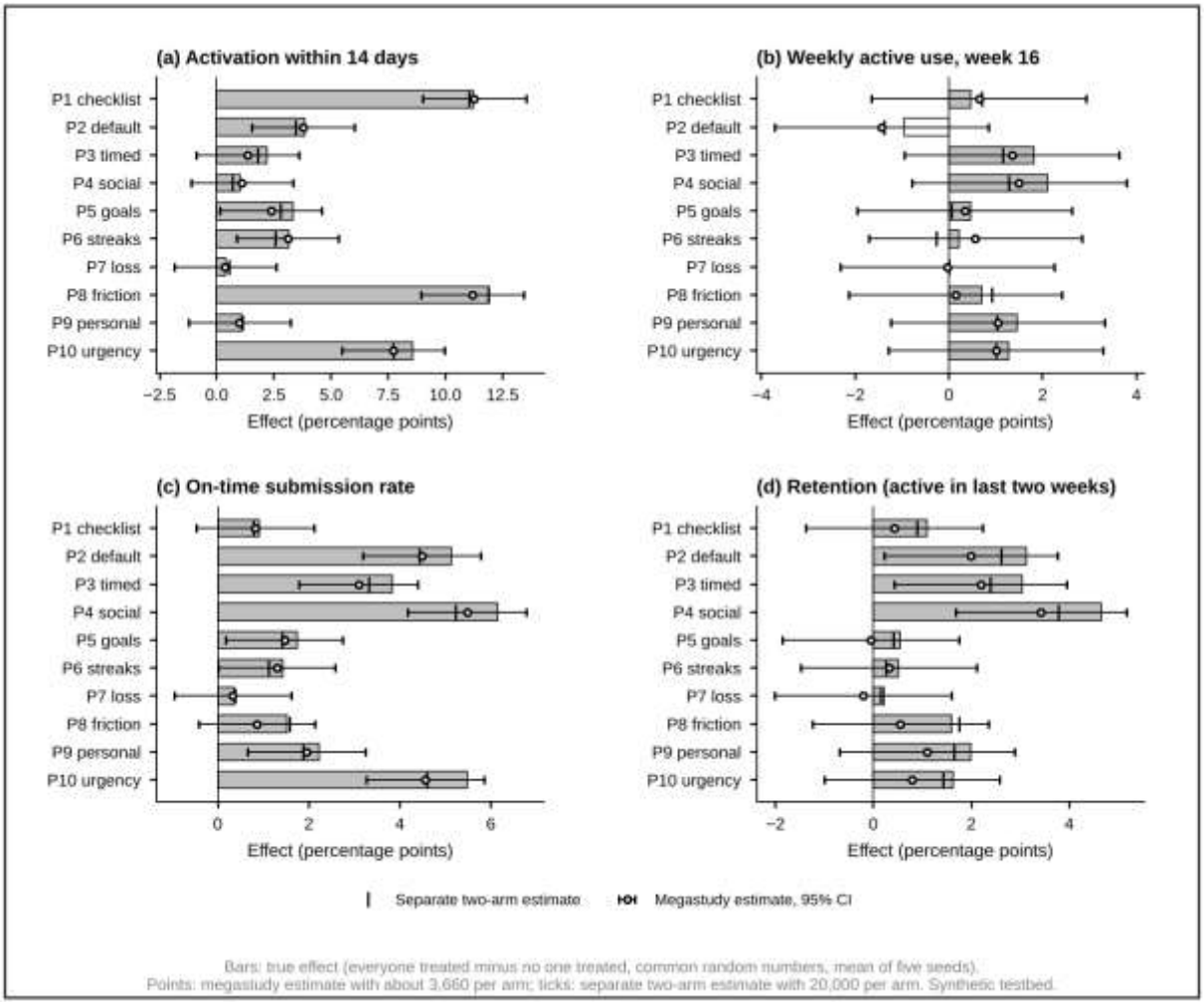


Table 8: True effects of the ten patterns on engagement, activation, academic proxy, retention and harm metrics (mean of five seeds, everyone treated against no one treated)

Pattern	Activation (pp)	WAU week 4 (pp)	WAU week 16 (pp)	Opens per term	On-time (pp)	Retained (pp)	Opt-out (pp)	Complaints per 100	Idle opens
Baseline (no pattern), level	37.6	67.8	45.3	23.3	45.6	80.1	7.2	0.0	3.18
P1 Onboarding checklist	+11.2	+0.9	+0.5	+0.50	+0.9	+1.1	0.0	+0.0	+0.04
P2 Default opt-in reminders	+3.9	+5.3	-0.9	+2.03	+5.1	+3.1	+11.3	-0.0	-0.00
P3 Timed deadline reminders	+2.2	+3.7	+1.8	+2.12	+3.8	+3.0	-2.0	-0.0	+0.20
P4 Social proof (gated)	+1.0	+3.2	+2.1	+1.70	+6.1	+4.7	-0.4	-0.0	+0.31
P5 Goal setting and planning	+3.3	+2.5	+0.5	+1.10	+1.7	+0.5	-0.1	-0.0	+0.27
P6 Streaks and badges	+3.1	+2.2	+0.2	+1.03	+1.4	+0.5	-0.1	-0.0	+0.42
P7 Loss framing	+0.4	+0.5	0.0	+0.15	+0.4	+0.2	+2.4	+0.0	+0.00
P8 Friction reduction	+11.9	+1.8	+0.7	+0.83	+1.5	+1.6	-0.1	+0.0	+0.07
P9 Personalised timing	+1.1	+2.1	+1.4	+1.30	+2.2	+2.0	-2.9	-0.0	+0.15
P10 Urgency countdown	+8.6	+11.4	+1.3	+3.13	+5.5	+1.6	-0.2	+11.1	+0.33

Recovery of the true effects by experiments

Table 9 compares the megastudy and the separate two-arm experiments with the truth. The megastudy, with about 3,660 students per arm, estimates every activation effect within 1 point of the truth on average and has standard errors of 1.1 to 1.2 points on activation and week-16 use and 0.7 points on on-time submission; its 95% intervals cover the true effect in at least four of five seeds for every pattern and metric, and in 141 of the 150 pattern-metric-seed cases. It declares the activation effects of P1, P8, P10, P2 and P6 significant in every seed, P5 in three, and finds the week-16 effects of P3, P4 and P9 (1.4 to 2.1 points) significant in at most two seeds of five. The separate experiments, with 20,000 students per arm, halve the standard errors (0.5 points on activation and week-16 use, 0.3 on on-time submission), and CUPED with the prior-term covariate reduces them further to 0.4 and 0.2 points. Their precision exposes a bias that the megastudy's noise hides: the estimates of on-time submission effects are 0.7 to 0.9 points below the truth for P2, P4 and P10 (4.4 against 5.1, 5.2 against 6.1 and 4.6 against 5.5), and the 95% intervals for these three cover the truth in at most one seed. The bias is interference: a treated student who opens the app more raises the recent-activity norm of the section, which raises the opening probability of control students in the same section, so the control arm is partly treated and the contrast understates the effect of treating everyone. The same mechanism lowers the separate-experiment estimates of week-16 use for P3 and P4 by 0.7 to 0.8 points. The bias is relative to the total effect, the quantity a product team wants when it decides whether to ship to everyone.

Table 9: Megastudy and separate two-arm estimates against the true effects: estimates, standard errors, coverage of the 95% interval and significance across five seeds

Pattern	True activation (pp)	Mega-study (SE)	Separate (SE)	True WAU16 (pp)	Mega-study (SE)	Separate (SE)	CI covers truth, of 15 (mega / sep.)	WAU16 signif., of 5 (mega / sep.)
P1 Onboarding checklist	+11.2	+11.3 (1.2)	+11.0 (0.5)	+0.5	+0.6 (1.2)	+0.7 (0.5)	15 / 15	0 / 2
P2 Default opt-in reminders	+3.9	+3.8 (1.1)	+3.4 (0.5)	-0.9	-1.4 (1.2)	-1.4 (0.5)	14 / 10	1 / 4
P3 Timed deadline reminders	+2.2	+1.4 (1.1)	+1.8 (0.5)	+1.8	+1.3 (1.2)	+1.1 (0.5)	12 / 11	2 / 2
P4 Social proof (gated)	+1.0	+1.1 (1.1)	+0.7 (0.5)	+2.1	+1.5 (1.2)	+1.3 (0.5)	14 / 8	0 / 4
P5 Goal setting and planning	+3.3	+2.4 (1.1)	+2.8 (0.5)	+0.5	+0.3 (1.2)	+0.1 (0.5)	14 / 13	0 / 0
P6 Streaks and badges	+3.1	+3.1 (1.1)	+2.5 (0.5)	+0.2	+0.6 (1.2)	-0.3 (0.5)	15 / 14	0 / 1
P7 Loss framing	+0.4	+0.4 (1.1)	+0.5 (0.5)	0.0	0.0 (1.2)	0.0 (0.5)	15 / 15	0 / 0
P8 Friction reduction	+11.9	+11.2 (1.2)	+11.9 (0.5)	+0.7	+0.1 (1.2)	+0.9 (0.5)	14 / 15	0 / 3
P9 Personalised timing	+1.1	+1.0 (1.1)	+1.1 (0.5)	+1.4	+1.0 (1.2)	+1.0 (0.5)	15 / 12	1 / 3
P10 Urgency countdown	+8.6	+7.7 (1.2)	+7.7 (0.5)	+1.3	+1.0 (1.2)	+1.0 (0.5)	13 / 6	1 / 3

Time paths, novelty and surrogates

Figure 7 and Table 10 give the effect of each pattern on weekly active use by term week. The reminder and social patterns build: timed reminders rise from 2.8 points in week 1 to 3.7 by week 4 and hold at 3.5 to 3.8 through week 15; social proof rises from 0.4 points in week 1 to 5.7 in week 15 as the favourable norm it displays becomes more common and more students have a deadline card to see it on; personalised timing rises from 1.1 to 2.3 points; the onboarding patterns rise from near zero to 1.3 (P1) and 2.1 (P8) points as the students they activated sync their calendars. The novelty patterns decay: the urgency countdown peaks at 13.1 points in week 2, falls to 6.0 in week 5 when the four-week countdown ends and to 1.3 by week 16; goals hold at 4.5 to 4.8 points for the three weeks of the planning prompt and then decay to 0.5; streaks fall from 5.0 points in week 1 to 0.2 in week 16 as the novelty multiplier fades and crowding-out sets in for high-intent students. Default opt-in starts at 6.4 points, declines to 4.1 as opt-outs accumulate, and turns negative (0.9 points below control) in the no-deadline week 16. Across the ten patterns, the rank correlation between the week-2 effect and the week-16 effect is close to zero (Spearman rank correlation of about 0.1 below zero) and between the week-4 and week-16 effects about 0.24; the week-2 ranking puts urgency, default opt-in, goals and streaks first, the week-16 ranking puts social proof, timed reminders and personalised timing first. The persistence ratio (week-16 effect over week-4 effect) is 0.11 for urgency, 0.09 for streaks and 0.18 for goals against 0.48 for timed reminders, 0.65 for social proof and 0.70 for personalised timing. On-time submission, which is observed every week, ranks the patterns more like week-16 use (rank correlation 0.54) than week-2 use does, and is the better early surrogate in this testbed.

Figure 7: Weekly time paths of the effects on weekly active use, the week-4 effect as a surrogate for the week-16 effect, and the persistence ratio by pattern

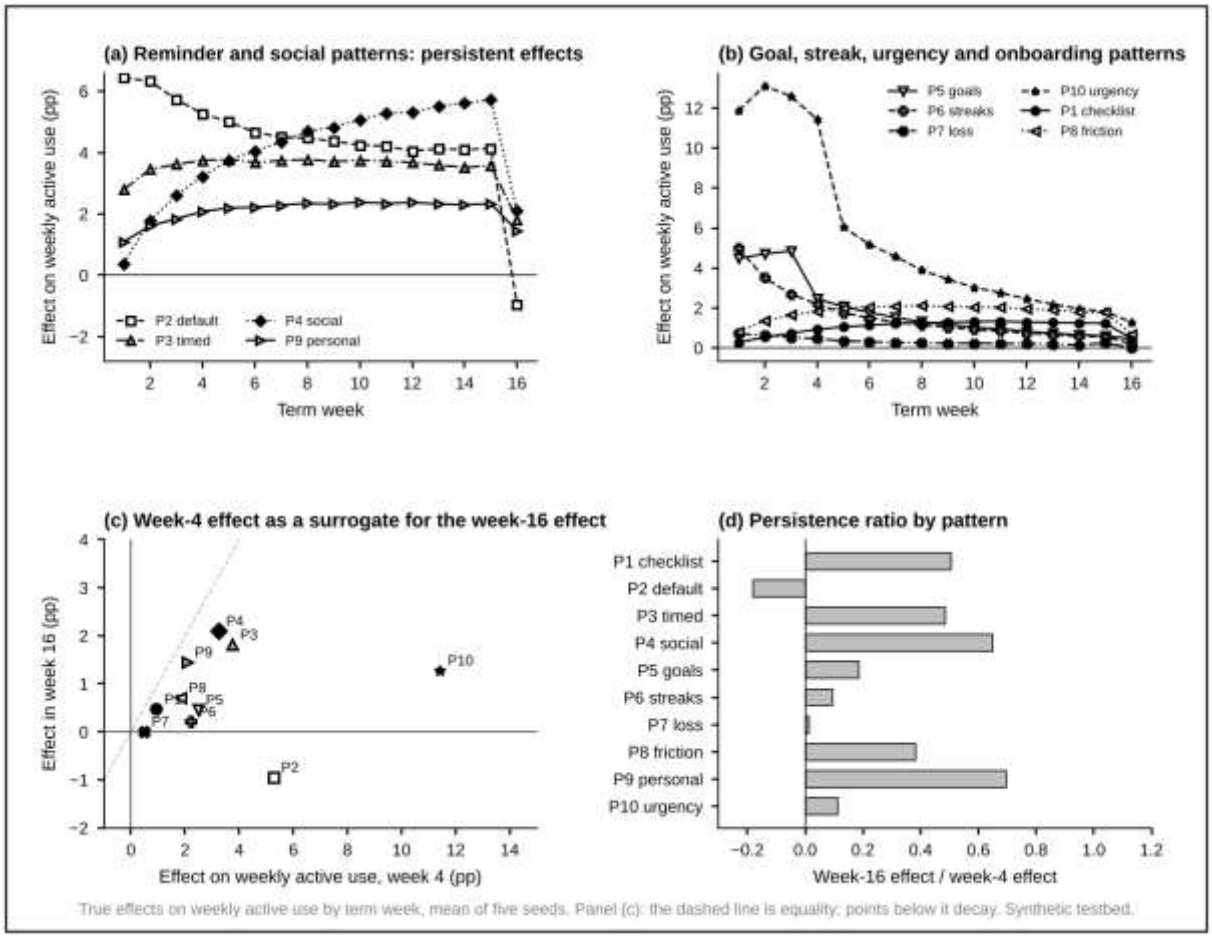


Table 10: Effects on weekly active use by term week, peak week, persistence ratio and the shape of the time path for each pattern

Pattern	Week 1	Week 2	Week 4	Week 8	Week 12	Week 16	Peak week	Week 16 / week 4	Time path
P1 Onboarding checklist	+0.3	+0.5	+0.9	+1.3	+1.3	+0.5	11	0.51	Partly persistent
P2 Default opt-in reminders	+6.4	+6.3	+5.3	+4.5	+4.1	-0.9	1	-0.18	Reverses by week 16
P3 Timed deadline reminders	+2.8	+3.4	+3.7	+3.8	+3.7	+1.8	5	0.48	Partly persistent
P4 Social proof (gated)	+0.4	+1.8	+3.2	+4.7	+5.3	+2.1	15	0.65	Partly persistent
P5 Goal setting and planning	+4.5	+4.7	+2.5	+1.3	+0.8	+0.5	3	0.18	Decays
P6 Streaks and badges	+5.0	+3.5	+2.2	+1.2	+0.8	+0.2	1	0.09	Decays
P7 Loss framing	+0.7	+0.6	+0.5	+0.3	+0.2	0.0	1	0.01	Decays
P8 Friction reduction	+0.8	+1.4	+1.8	+2.1	+1.9	+0.7	8	0.38	Partly persistent
P9 Personalised timing	+1.1	+1.6	+2.1	+2.3	+2.4	+1.4	10	0.70	Partly persistent
P10 Urgency countdown	+11.8	+13.1	+11.4	+3.9	+2.5	+1.3	2	0.11	Decays

Heterogeneity and targeting

Table 11 and Figure 8 decompose the true effects by cohort, intent quartile and campus, and report the uplift targeting experiment. Cohort differences are small (within 1 point for every pattern and metric), because cohorts differ in the testbed only through an offset to intent. Intent quartiles matter more. The onboarding patterns help the middle quartiles most: the checklist raises activation by 5.9 points in the lowest intent quartile, 11.8 and 15.0 in the middle two and 12.1 in the highest, because the lowest quartile rarely opens the app often enough to complete the steps and the highest completes them without help; friction reduction follows the same shape (6.6, 13.5, 16.3 and 11.2 points). Default opt-in's harm concentrates where its benefit is smallest: opt-outs rise by 14.0 points in the lowest intent quartile against 7.9 in the highest, and the week-16 effect is 0.2 points below control in the lowest quartile but 1.7 points below in the highest, whose students were already active and are mainly annoyed. Social proof's week-16 effect is 2.0, 3.1, 2.4 and 0.9 points across the quartiles, personalised timing's 1.3, 1.8, 1.7 and 1.1, and streaks' week-16 effect is positive in the three lower quartiles and 0.7 points negative in the highest, where crowding-out of intrinsic motivation outweighs the streak utility. Campus ranges of the true effects are within 2 points for every pattern and metric except social proof on on-time submission (3.7 to 7.9 points across campuses), because the campus intent offset changes how often the favourable norm is displayed.

The uplift experiment (Figure 8c and 8d) trains a T-learner on half of a two-arm experiment and targets the half of the held-out students with the highest predicted uplift. The predicted uplift on week-16 use spans about 10 points below zero to 13 above across quintiles for the default gradient-boosting learner (standard deviation 8 points across students) and 3 below to 6 above for a strongly regularised learner (standard deviation 3 points), while the true effect by predicted-uplift quintile is flat: 1.7 to 2.0 points for timed reminders and 1.3 to 1.6 for personalised timing. The targeting policy therefore does no better than treating a random half: for timed reminders, treating everyone raises week-16 use by 1.9 points on the held-out half, the targeted half by 0.9 (0.9 with the regularised learner) and a random half by 1.0; for personalised timing 1.5, 0.7 (0.7) and 0.8; for default opt-in, where the goal would be to avoid the students it harms, the targeted half shows 0.4 points below control against 0.5 for a random half, and opt-outs of 5.8 against 5.7 points. The true heterogeneity in Table 11 (about 1 point between intent quartiles) is small relative to the noise of a binary outcome in 10,000 training students per arm, so the learners fit noise, and a product team that trusted the predicted uplift would believe in a 20-point spread of effects that does not exist. Two ablations complete the picture. Applying personalised timing to every student rather than to the 60% with weak habits raises its week-16 effect from 1.4 to 2.4 points, its on-time effect from 2.2 to 3.7 and its opt-out reduction from 2.9 to 4.9 points, because in the testbed the mechanism does not depend on habit, so targeting forgoes benefit without saving harm. Showing social proof always rather than only when the section norm is favourable lowers its activation effect from 1.0 to 0.5 points and its on-time effect from 6.1 to 5.6, a modest cost because unfavourable norms are rare in a testbed where most sections are more than half active.

Figure 8: Heterogeneity of the true effects by intent quartile and the uplift targeting experiment: predicted and true uplift by quintile and the realised effect of the targeting policy against treating everyone

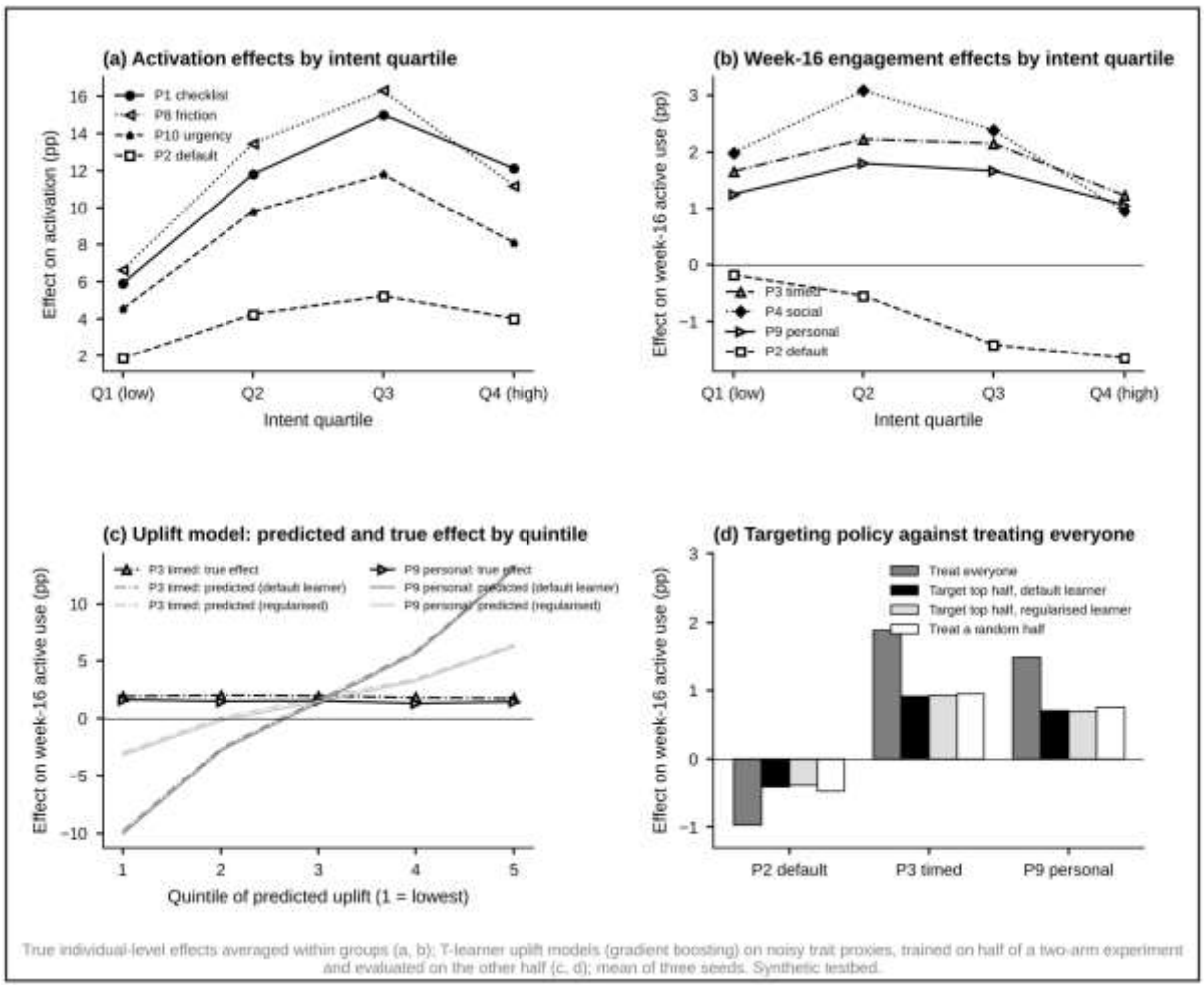


Table 11: Heterogeneity of the true effects by cohort, intent quartile and campus for activation, week-16 active use and opt-out

Pattern	Activation by cohort: first-year / continuing / graduate (pp)	Activation, intent Q1 / Q4 (pp)	WAU week 16 by cohort (pp)	WAU week 16, intent Q1 / Q4 (pp)	Opt-out, intent Q1 / Q4 (pp)	Campus range: activation / WAU16 (pp)
P1 Onboarding checklist	+10.8 / +11.3 / +11.7	+5.9 / +12.1	+0.5 / +0.5 / +0.4	+0.5 / +0.2	0.0 / 0.0	1.8 / 0.1
P2 Default opt-in reminders	+3.8 / +3.8 / +4.0	+1.9 / +4.0	-0.8 / -1.0 / -1.1	-0.2 / -1.7	+14.0 / +7.9	0.5 / 0.3
P3 Timed deadline reminders	+2.2 / +2.2 / +2.2	+1.6 / +1.6	+1.8 / +1.8 / +1.8	+1.7 / +1.2	-2.3 / -1.5	0.3 / 0.1
P4 Social proof (gated)	+1.1 / +1.0 / +1.0	+0.7 / +0.7	+2.2 / +2.1 / +1.9	+2.0 / +0.9	-0.2 / -0.5	0.8 / 1.1
P5 Goal setting and planning	+3.1 / +3.4 / +3.5	+1.3 / +4.1	+0.5 / +0.5 / +0.4	+0.3 / +0.4	0.0 / -0.1	0.7 / 0.1
P6 Streaks and badges	+2.9 / +3.2 / +3.3	+0.9 / +4.5	+0.3 / +0.2 / +0.1	+0.2 / -0.7	0.0 / -0.1	1.0 / 0.1
P7 Loss framing	+0.4 / +0.4 / +0.4	+0.3 / +0.3	0.0 / 0.0 / 0.0	+0.1 / -0.1	+3.0 / +1.7	0.1 / 0.1
P8 Friction reduction	+11.6 / +12.0 / +12.1	+6.6 / +11.2	+0.7 / +0.7 / +0.6	+0.8 / +0.3	0.0 / 0.0	1.5 / 0.1
P9 Personalised timing	+1.1 / +1.1 / +1.1	+0.8 / +0.8	+1.4 / +1.4 / +1.5	+1.3 / +1.1	-3.7 / -1.8	0.2 / 0.1
P10 Urgency countdown	+8.4 / +8.5 / +8.8	+4.5 / +8.1	+1.3 / +1.3 / +1.3	+0.9 / +0.9	-0.1 / -0.3	1.2 / 0.2

Interference and the randomisation unit

Table 12 and Figure 9 show what the randomisation unit does under interference for the two patterns with social mechanisms. For social proof at the default interference strength, the true effect on week-16 use is 2.1 points; user-level randomisation estimates 1.4 (standard error 0.5), a downward bias of a third, and randomisation by course section estimates 1.7 (standard error 0.7). At double the interference strength the true effect rises to 2.9 points while the user-level estimate falls to 1.1, a bias of 1.8 points or 62%, and the section estimate is 1.8. At half strength the biases are 0.5 and 0.2 points. On on-time submission the user-level estimates are 0.9 to 1.7 points below the truth (5.9 to 6.6 points) and the section estimates 0.4 to 0.8 below. The section design removes between a third and two-thirds of the bias rather than all of it, because each student belongs to four sections and randomisation by one of them leaves three-quarters of the peer exposure uncontrolled, and it costs precision: cluster-robust standard errors are 0.6 to 0.8 points against 0.5 for the user design. For streaks, whose only social channel is the general peer-activity norm, the week-16 effect is near zero under every design; the week-4 effect of 1.9 to 2.7 points is underestimated by 0.5 to 1.1 points under user randomisation. The multi-campus decomposition (Figure 9d) finds no between-campus heterogeneity beyond sampling error for 29 of the 30 pattern-metric pairs (I-squared 0%) and I-squared of 57% for social proof on on-time submission, with a between-campus standard deviation of 1.3 points; with hypothetical 4,000-student samples per campus the pooled random-effects estimates have standard errors of 0.5 to 0.8 points, close to the user-level separate experiment on 20,000 per arm.

Figure 9: Interference and the randomisation unit for social proof and streaks at three interference strengths, and the multi-campus decomposition with random-effects pooling

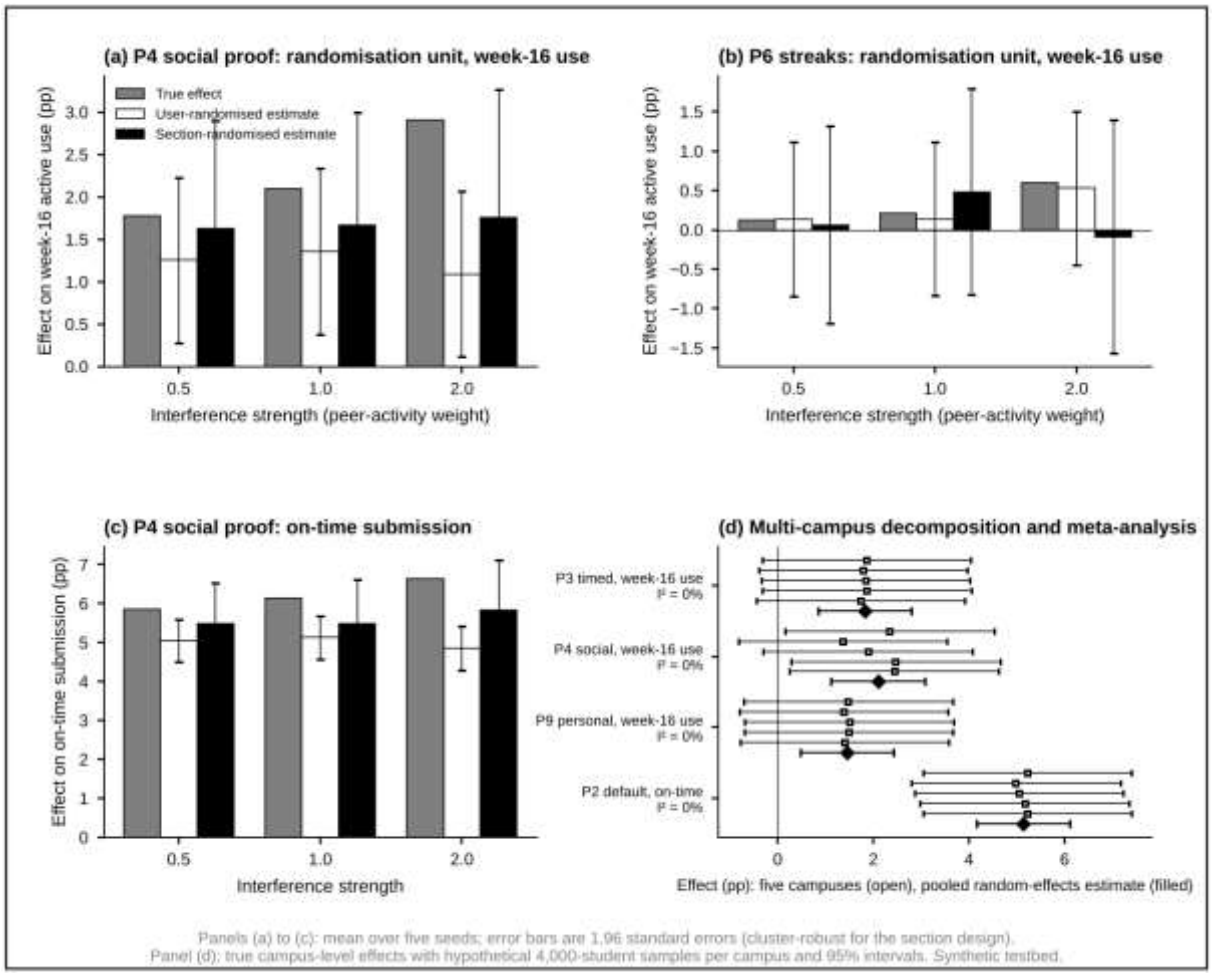


Table 12: User-level and section-cluster estimates against the true effects of social proof and streaks at three interference strengths

Pattern	Interference strength	True WAU16 (pp)	User-randomised (SE)	Bias	Section-randomised (SE)	Bias	True on-time (pp)	User bias	Section bias
P4 Social proof	0.5	+1.8	+1.3 (0.5)	-0.5	+1.6 (0.6)	-0.1	+5.9	-0.8	-0.4
P4 Social proof	1.0	+2.1	+1.4 (0.5)	-0.7	+1.7 (0.7)	-0.4	+6.1	-1.0	-0.7
P4 Social proof	2.0	+2.9	+1.1 (0.5)	-1.8	+1.8 (0.8)	-1.1	+6.6	-1.8	-0.8
P6 Streaks	0.5	+0.1	+0.1 (0.5)	0.0	+0.1 (0.6)	-0.1	+1.2	-0.2	-0.4
P6 Streaks	1.0	+0.2	+0.1 (0.5)	-0.1	+0.5 (0.7)	+0.3	+1.4	-0.6	0.0
P6 Streaks	2.0	+0.6	+0.5 (0.5)	-0.1	-0.1 (0.8)	-0.7	+1.9	-0.8	-1.1

Method properties: peeking, power, errors and corrections

Table 13 and Figure 10 report the method checks. A/A tests with a single pre-registered look at the end of term declared 1 of 40 replications significant (2.5%); the same tests examined every week and stopped at the first significant week declared 11 of 40 significant (27.5%), five times the nominal rate, which is the peeking inflation that always-valid sequential methods are designed to remove. CUPED with the prior-term opening covariate reduces standard errors by 12% for activation, 8 to 10% for weekly active use and retention and 37 to 38% for on-time submission and opens per term, in proportion to the correlation of each outcome with the covariate (0.45 to 0.48 for the binary engagement metrics, 0.79 for

on-time submission). Power for the small persistent effects is low at the sample sizes a single campus has: for timed reminders on week-16 use (1.8 points), a two-arm test with 2,000 students per arm has 21% power (25% with CUPED), 4,000 per arm 36% (45%), 8,000 per arm 64% (73%) and 16,000 per arm 90% (95%); for personalised timing (1.4 points) 16,000 per arm gives 73% (83%); for social proof on activation (1.0 point) 16,000 per arm gives 47%. The large activation effects need little: the checklist's 11.2 points has 72% power at 250 per arm and full power at 1,000. On-time submission, with its lower variance and higher covariate correlation, is the most detectable metric: a 5.1-point effect has 90% power at 250 per arm with CUPED. When a small effect is tested with few students and found significant, the estimate exaggerates it: the exaggeration ratio for timed reminders on week-16 use is 5.8 at 250 per arm, 3.0 at 1,000, 2.2 at 2,000 and 1.6 at 4,000, and at 250 per arm 14% of significant estimates have the wrong sign. The multiple-testing exercise over 60 pattern-metric tests at 2,000 per arm finds that 24.5 tests are significant uncorrected on average, of which 0.8 have true effects below 0.5 points, and 20.1 after Benjamini-Hochberg correction, of which 0.3; of the 40 true effects of at least 1 point, 23.1 are detected uncorrected and 19.5 after correction. False discoveries are rare in both cases because most of the 60 true effects are not null; the cost of testing at this sample size is the missed small effects, not the false ones.

Figure 10: Method properties: power by sample size with and without CUPED, Type M exaggeration, CUPED variance reduction, and peeking and multiple testing

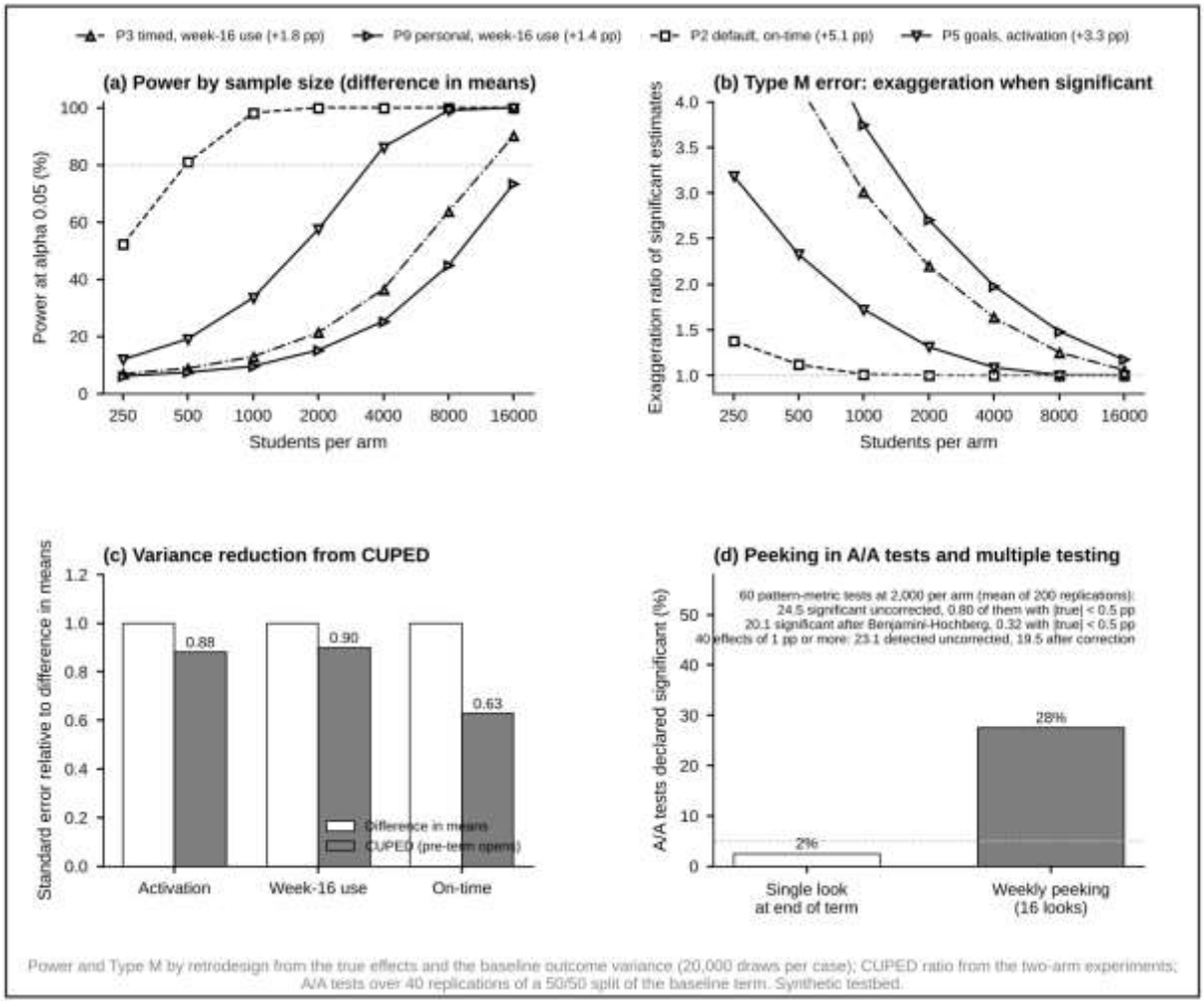


Table 13: Properties of the evaluation methods in the testbed: A/A tests with peeking, CUPED, power, Type S and Type M errors, multiple testing, multi-campus heterogeneity and Monte Carlo rank stability

Check	Result in the testbed
A/A tests, single look at end of term	2.5% of 40 replications significant at alpha = 0.05 (nominal 5%)
A/A tests, weekly peeking (16 looks)	27.5% of 40 replications significant at some look
CUPED standard error relative to difference in means	Activation 0.88; WAU week 4 0.90; WAU week 16 0.90; on-time 0.63; retention 0.92
Power (%), P3 timed reminders, WAU week 16 (true +1.8 pp), n per arm 250 / 1,000 / 2,000 / 4,000 / 8,000 / 16,000	7 / 13 / 21 / 36 / 64 / 90; with CUPED at 2,000: 25
Power (%), P9 personalised timing, WAU week 16 (true +1.4 pp), n per arm 250 / 1,000 / 2,000 / 4,000 / 8,000 / 16,000	6 / 10 / 15 / 25 / 45 / 73; with CUPED at 2,000: 18
Power (%), P2 default opt-in, on-time (true +5.1 pp), n per arm 250 / 1,000 / 2,000 / 4,000 / 8,000 / 16,000	52 / 98 / 100 / 100 / 100 / 100; with CUPED at 2,000: 100
Power (%), P5 goal setting, activation (true +3.3 pp), n per arm 250 / 1,000 / 2,000 / 4,000 / 8,000 / 16,000	12 / 34 / 57 / 86 / 99 / 100; with CUPED at 2,000: 69
Power (%), P1 checklist, activation (true +11.2 pp), n per arm 250 / 1,000 / 2,000 / 4,000 / 8,000 / 16,000	72 / 100 / 100 / 100 / 100 / 100; with CUPED at 2,000: 100
Type M (exaggeration ratio) and Type S (wrong sign, %), P3, WAU week 16, n per arm 250 / 1,000 / 4,000	5.84 / 3.01 / 1.63; Type S 14 / 2 / 0
Type M (exaggeration ratio) and Type S (wrong sign, %), P9, WAU week 16, n per arm 250 / 1,000 / 4,000	7.29 / 3.75 / 1.98; Type S 17 / 4 / 0
Type M (exaggeration ratio) and Type S (wrong sign, %), P5, activation, n per arm 250 / 1,000 / 4,000	3.19 / 1.72 / 1.09; Type S 3 / 0 / 0
Multiple testing, 60 pattern-metric tests at 2,000 per arm (200 replications)	40 tests have a true effect of at least 1 pp and 14 below 0.5 pp; per replication 24.5 tests are significant uncorrected (0.80 with a true effect below 0.5 pp) and 20.1 after Benjamini-Hochberg (0.32 below 0.5 pp); of the effects of at least 1 pp, 23.1 are detected uncorrected and 19.5 after correction
Multi-campus meta-analysis (30 pattern-metric pairs, five campuses)	I-squared 0% to 57%; between-campus standard deviation of the true effect 0.00 to 1.27 pp; campus range of true effects up to 4.1 pp
Monte Carlo ranking stability (200 draws)	P1 or P8 ranks first on activation in 82% of draws (both in the top two in 56%); P4 ranks first on WAU week 16 in 54% of draws; P2 lowers WAU week 16 in 70% of draws; P6 lowers it in 28%; P10 raises complaints above 5 per 100 in 100%

Harms, combinations and robustness

Figure 11 and Table 14 complete the picture. The harm metrics separate the patterns as sharply as the engagement metrics: default opt-in raises opt-outs by 11.3 points, loss framing by 2.4 points, and the urgency countdown produces 11.1 complaints per 100 students and reduces end-of-term trust from 1.0 to 0.03, while timed reminders, personalised timing and social proof lower opt-outs (by 2.0, 2.9 and 0.4 points) because they make the reminders students receive more useful or fewer. Combinations are not additive. The two onboarding patterns together raise activation by 19.9 points against a sum of 23.1 for the singles, a sub-additive interaction of 3.2 points because both act on the same friction; adding timed reminders (P1 + P3 + P8) gives 22.6 points. Default opt-in with timed reminders is super-additive: 14.5 points on week-4 use against a sum of 9.0, 3.8 on week-16 use against 0.9, 14.9 on on-time submission against 9.0, with opt-outs 3.2 points lower than the sum, because the timed, less annoying reminder reaches the 90% of students the default enrols; default opt-in with personalised timing behaves the same way (week-16 use 2.8 against 0.5, opt-outs 4.7 points below the sum). Timed reminders with social proof and streaks with urgency are additive. All seven benign patterns together raise activation by 40.6 points, week-4 use by 25.4, week-16 use by 9.3 and on-time submission by 32.2, with no net change in opt-outs, and the interactions are positive on every engagement metric (6.0 points on activation, 10.7 on on-time submission).

The Monte Carlo analysis (Figure 11c and 11d, Table 13) draws 200 combinations of the fourteen parameters in Table 4 on 10,000-student populations. The ranking on activation is stable: friction reduction or the checklist ranks first in 82% of draws (the urgency countdown in the rest), with medians of 10.4 and 12.4 points and 10th to 90th percentile ranges of 5.5 to 13.8 and 7.9 to 16.5 points. The ranking on week-16 use is stable at the top and ambiguous below: social proof ranks first in 54% of draws and in the top three in 96%, timed reminders rank first in 30% and in the top three in 69%, personalised timing is in the top three in 59% and the urgency countdown in 53%; their medians are 2.0, 1.5, 1.3 and 1.1 points. Default opt-in lowers week-16 use in 70% of draws (median 0.6 points below control, range 3.4 below to 3.3 above) and its opt-out effect ranges from 0.6 to 45.7 points (median 10.5), driven by annoyance per reminder and tolerance (rank correlations of 0.65 and 0.66 in absolute value); streaks lower week-16 use in 28% of draws. The urgency countdown produces 9 to 12 complaints per 100 students in every draw. The parameters that matter are the expected ones: the reminder multiplier for timed reminders (rank correlation 0.83 with its week-16 effect), the norm amplification and the interference strength for social proof (0.74 and 0.59), and the activation link for the engagement effects of the onboarding patterns (0.86 for the checklist). The qualitative conclusions (onboarding patterns for activation, deadline and habit patterns for sustained engagement, novelty patterns for neither, default opt-in and urgency as the patterns with harms) hold across the ranges; the sizes do not.

Figure 11: Harm metrics by pattern, interactions between combined patterns, and the Monte Carlo distributions of the week-16 and activation effects under joint parameter uncertainty

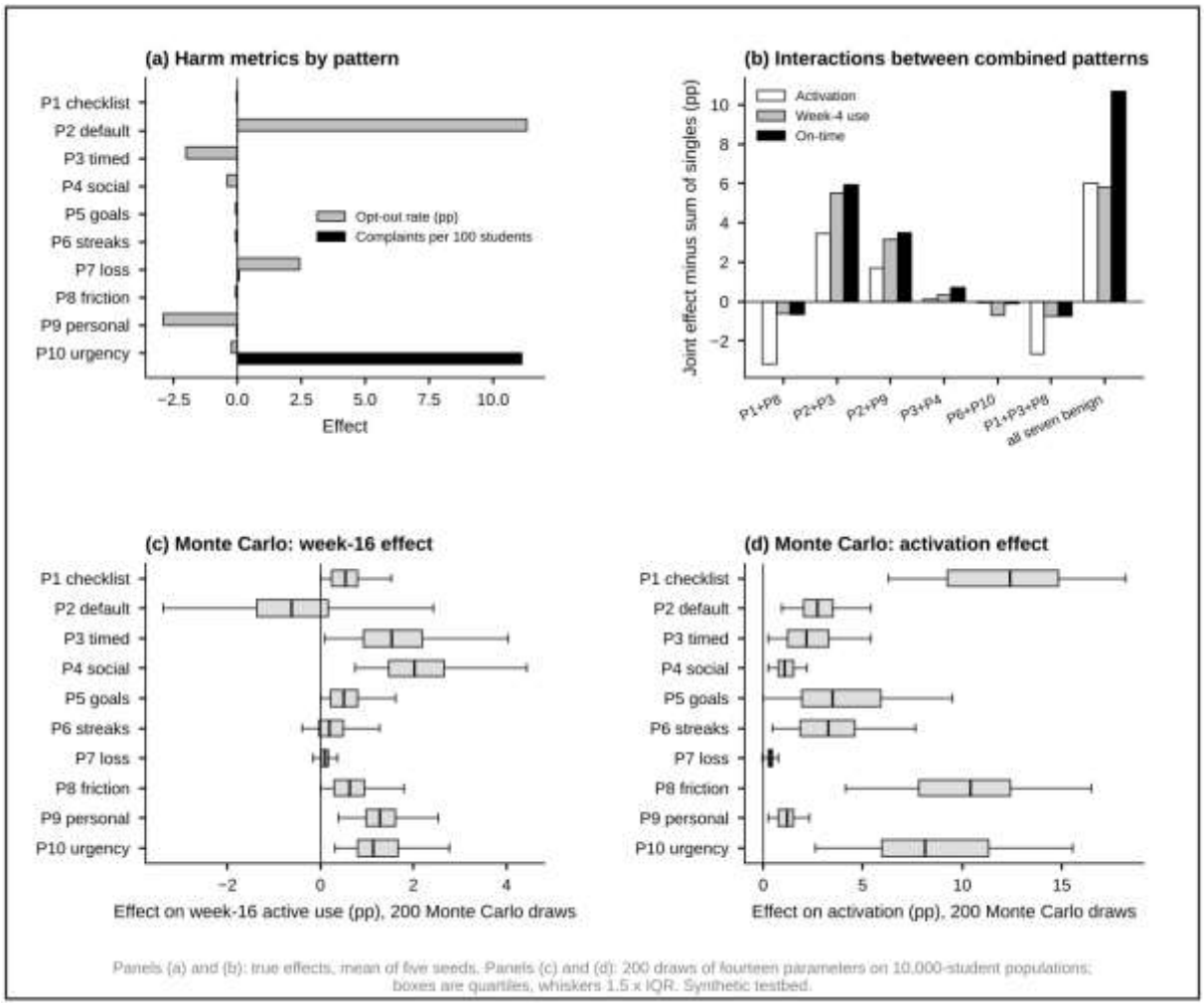


Table 14: Combinations of patterns: joint effect and interaction (joint effect minus the sum of the single effects) for activation, week-16 active use, on-time submission and opt-out

Combination	Activa-tion: joint (pp)	Inter-action	WAU week 16: joint (pp)	Inter-action	On-time: joint (pp)	Inter-action	Opt-out: joint (pp)	Inter-action
P1 + P8 (checklist and friction reduction)	+19.9	-3.2	+0.9	-0.3	+1.8	-0.6	-0.1	0.0
P2 + P3 (default opt-in and timed reminders)	+9.5	+3.4	+3.8	+2.9	+14.9	+5.9	+6.1	-3.2
P2 + P9 (default opt-in and personalised timing)	+6.6	+1.7	+2.8	+2.3	+10.8	+3.5	+3.8	-4.7
P3 + P4 (timed reminders and social proof)	+3.3	+0.1	+4.0	+0.1	+10.7	+0.7	-2.4	0.0
P6 + P10 (streaks and urgency)	+11.7	0.0	+1.5	0.0	+6.8	-0.1	-0.3	0.0
P1 + P3 + P8	+22.6	-2.7	+2.6	-0.4	+5.6	-0.7	-2.1	0.0
All seven benign patterns (P1 to P5, P8, P9)	+40.6	+6.0	+9.3	+3.3	+32.2	+10.7	-0.3	-6.1

Discussion

Answers to the research questions

RQ1 asked how large the effects are, how they evolve and how they differ across students. In the testbed the answer depends on which metric is asked. Activation is moved by removing friction and by guiding the steps (about 12 and 11 percentage points, a third of the baseline rate), by the urgency countdown (9 points) and modestly by defaults, goals and streaks (3 to 4 points); sustained engagement, measured in the week without deadlines, is moved by social proof, timed reminders and personalised timing (1.4 to 2.1 points, 3 to 5% of the baseline rate) and by nothing else beyond 1.3 points; on-time submission, the academic proxy, is moved by social proof, urgency, defaults and timed reminders (4 to 6 points). The time paths divide the patterns into those whose effects build over the term because they act on a repeated behaviour (reminders, social proof, personalised timing, and the onboarding patterns through calendar sync) and those whose effects decay because they act through novelty or a time-limited prompt (urgency, goals, streaks). The decay is the testbed's version of the novelty effects that experimentation practice warns about and of the fading that field studies of nudges report when the intervention stops (Kohavi et al., 2012; Hohnhold et al., 2015; Allcott & Rogers, 2014; Milkman et al., 2021a). Heterogeneity across cohorts is small by construction, but across intent it is large and non-monotonic for the onboarding patterns, which help the middle of the motivation distribution most, and it reverses the sign of the streak effect for the most motivated students, as the crowding-out literature would predict (Deci et al., 1999; Hanus & Fox, 2015).

RQ2 asked what harms the patterns produce and how they trade against engagement. The harms are concentrated: default opt-in produces an 11-point rise in opt-outs for a 5-point gain in on-time submission and a small loss in week-16 use; the urgency countdown produces 11 complaints per 100 students and the loss of nearly all trust for a 13-point week-2 gain that is 1.3 points by week 16; loss framing produces opt-outs without engagement. Three patterns reduce harm while raising engagement (timed reminders, personalised timing, gated social proof), because they make the reminders a student receives more useful or fewer. The trade-off is therefore not a frontier along which every pattern lies but a sorting: patterns that respect the student's attention budget are better on both axes, and patterns that spend it are worse on both once the term is long enough for the harm to appear. The sizes are assumptions, but the structure, in which annoyance accumulates and opt-out is absorbing, follows the notification literature and the testbed makes its consequence visible on a term's horizon (Pielot et al., 2014; Mehrotra et al., 2016; Sahami Shirazi et al., 2014; Mathur et al., 2019; Federal Trade Commission, 2022).

RQ3 asked how well standard methods recover the truth. Four results matter for a campus product. First, sample size: the persistent effects that matter most for a product's long-run engagement are 1 to 2 points, and a two-arm test needs 8,000 to 16,000 students per arm for 65 to 90% power on a binary weekly-use metric even with CUPED, which is most or all of a large campus; a megastudy with eleven arms on 40,000 students had 1.2-point standard errors and could not distinguish these effects from zero in most seeds, while it measured the 9 to 12-point activation effects precisely. On-time submission, with lower variance and a strong prior-term covariate, is detectable at a tenth of the sample. Second, interference: randomisation by user understates the effect of social proof by a third at the default interference strength and by 62% at double strength, and understates the effect of every pattern that raises activity on the on-time metric by 10 to 15%, because treated students raise their control peers' activity through the section norm; randomisation by section removes a third to two-thirds of the bias at a cost of 20 to 60% in standard error, which is the pattern that the interference literature reports for graph-cluster designs with imperfect clusters (Eckles et al., 2017; Ugander et al., 2013; Hudgens & Halloran, 2008; Aronow & Samii, 2017). Third, inference: weekly peeking raised the A/A false-positive rate from 2.5% to 27.5%, and a significant estimate of a 1.8-point effect at 250 students per arm overstated it by a factor of six, with a 14% chance of the wrong sign, which is the Type M and Type S pattern that Gelman and Carlin describe and that the replication literature finds in published effects (Gelman & Carlin, 2014; Johari et al., 2017; Simmons et al., 2011; Ioannidis, 2005; Open Science Collaboration, 2015). Fourth, surrogates: the week-2 effect ranking carried no information about the week-16 ranking, and the week-4 ranking little; a team that ships on week-2 results would ship urgency, defaults, goals and streaks and leave social proof, timed reminders and personalised timing on the

shelf. Benjamini-Hochberg correction cost little at 2,000 per arm because most true effects were not null, and the dominant error was missed small effects rather than false discoveries, which is the regime of a product with many modest real effects rather than the regime of a literature with many null ones (Benjamini & Hochberg, 1995; Athey et al., 2019; Kohavi et al., 2020).

RQ4 asked how stable the answers are under the assumed mechanisms. The Monte Carlo analysis shows that the ranking of the onboarding patterns on activation, the position of social proof and timed reminders at the top of the week-16 ranking, the decay of the novelty patterns, the opt-out cost of default opt-in and the complaint cost of urgency hold across the ranges in Table 4, while every effect size moves by a factor of two or more and the sign of the default opt-in and streak effects on week-16 use is uncertain (negative in 70% and 28% of draws). The parameters that drive the uncertainty are ones a campus could measure in a baseline term: annoyance per reminder and tolerance, the strength of peer influence, and the link from calendar sync to deadline salience, which is an argument for the measurement foundation in the roadmap below (Law, 2015; Sargent, 2013).

Comparison with the literature

The effect sizes in the testbed were set by the author with reference to the field literature, so agreement with it is partly construction; what the testbed adds is the comparison on one population. Three points of contact are worth stating. First, the small persistent effects of reminders, defaults and norms (1 to 6 points) are in the range that nudge units report at scale (DellaVigna and Linos report an average take-up effect of 1.4 percentage points, an 8.1% increase, across 126 trials of two U.S. nudge units, against 8.7 points, a 33.5% increase, in the published academic trials they compare), and smaller than the early laboratory and field results that the catalogues of choice architecture summarise (DellaVigna & Linos, 2022; Thaler & Sunstein, 2008; Hummel & Maedche, 2019; Mertens et al., 2022; Maier et al., 2022). Second, the small and decaying effects of goals and planning prompts match the large field experiments in higher education that found little or no effect of goal-setting, coaching and text nudges on persistence and grades, and the null results of planning prompts scaled across online courses (Dobronyi et al., 2019; Oreopoulos & Petronijevic, 2019; Bird et al., 2021; Kizilcec et al., 2020; Clark et al., 2020). Third, the activation effects of friction reduction are large in the testbed because onboarding friction is a direct cost on a one-time action, which is where simplification has had its largest documented effects (Bhargava & Manoli, 2015; Madrian & Shea, 2001; Johnson & Goldstein, 2003). The megastudy design is borrowed from the vaccination and gym-attendance megastudies, which found that most interventions had effects of a few percentage points, that the best were modest, and that effects faded when the intervention ended; the testbed reproduces the fading and shows what the megastudy's shared control can and cannot resolve at campus scale (Milkman et al., 2021b; Milkman et al., 2021a). The engagement literature's warning that software engagement is a proxy for behavioural engagement, not for learning, is reflected in the inclusion of on-time submission and idle opens as separate metrics (Fredricks et al., 2004; Henrie et al., 2015; O'Brien et al., 2018).

Design implications

Six implications follow for a campus product and its evaluation framework (Figures 1 and 3). Define activation and engagement as separate primary metrics with separate owners, because the patterns that move one do not move the other and the ranking of patterns inverts between them. Pre-register the week-16 (or end-of-term) metric as primary and treat week-2 and week-4 metrics as diagnostic, or use on-time submission as the early surrogate, which in the testbed tracked the long-run ranking far better than early use. Measure harm as guardrails with the same status as engagement (opt-outs, complaints, idle opens, trust), with automatic ramp-down; the patterns that survive these guardrails are also the ones with persistent effects. Randomise by course section, or by a graph cluster that covers more of a student's peers, for any pattern with a social mechanism, and expect a downward bias in user-level estimates for every pattern that raises activity. Plan sample sizes with retrodesign on the effect the product needs, not the effect the literature hopes for: 1 to 2 points on a binary weekly metric needs most of a campus, and pooling across campuses with cluster-robust or random-effects methods is the realistic route to power for the persistent patterns. Treat uplift targeting as an evaluation question rather than a default, because in the testbed two learners found heterogeneity that was not there and a targeted half did no better than a random half; a targeting policy should be evaluated against treating everyone and treating a random subset in a held-out sample before it is trusted, and should pass the fairness

review that the regulatory mapping below requires (Künzel et al., 2019; Wager & Athey, 2018; Athey & Imbens, 2017; Gutierrez & Gérardy, 2017).

Ethics, policy and regulatory mapping

Table 15 maps the patterns and the practices of the evaluation framework to the instruments that govern them. Three points follow from the testbed results rather than from the texts alone. First, the pattern with the largest short-run engagement effect, the urgency countdown, is also the one that the FTC staff report and Article 25 of the Digital Services Act describe most directly: false urgency and interfaces that distort decisions (Federal Trade Commission, 2022; European Parliament and Council, 2022; Mathur et al., 2019). A team that ranks patterns on week-2 engagement without a harm guardrail would ship it; a team with the complaint and trust guardrails of Table 5 would not. Second, defaults are lawful and effective, but the testbed shows that an opt-in default raises opt-outs by more than ten percentage points and lowers week-16 use for the students it annoys most; the ethics of nudging literature asks that defaults be transparent and easy to reverse, and the product consequence is a one-tap opt-out and an annoyance budget rather than a hidden setting (Sunstein, 2015; Sunstein, 2022; Thaler & Sunstein, 2008). Third, uplift targeting and personalised timing use models to decide which students are nudged; under FERPA these are education records used by school officials, under the GDPR they are profiling, and under the EU AI Act systems that evaluate or prioritise students in education are high-risk, so a data-protection impact assessment, human oversight and transparency about targeting are the price of the modest gains that Section 5 reports (U.S. Department of Education, n.d.; European Parliament and Council, 2016; European Parliament and Council, 2024; Prinsloo & Slade, 2017; Slade & Prinsloo, 2013). The Common Rule's exemption for normal educational practice covers much of this work when it is run as institutional improvement, but the determination belongs to the IRB, and the resentment of undisclosed A/B tests that Meyer et al. (2019) document argues for plain-language notice regardless of the legal exemption (U.S. Department of Health and Human Services, 2018; Meyer et al., 2019; Kramer et al., 2014). Accessibility review under WCAG 2.2 applies to countdown timers and animated streaks specifically, because time limits and motion are among its criteria (World Wide Web Consortium, 2023). The instruments cited are read as of their cited editions; statutes, guidance and enforcement priorities change and differ by jurisdiction and institution, and the mapping is a reading for design, not legal advice.

Table 15: Mapping of the patterns and the evaluation framework to policy, regulatory and ethical instruments

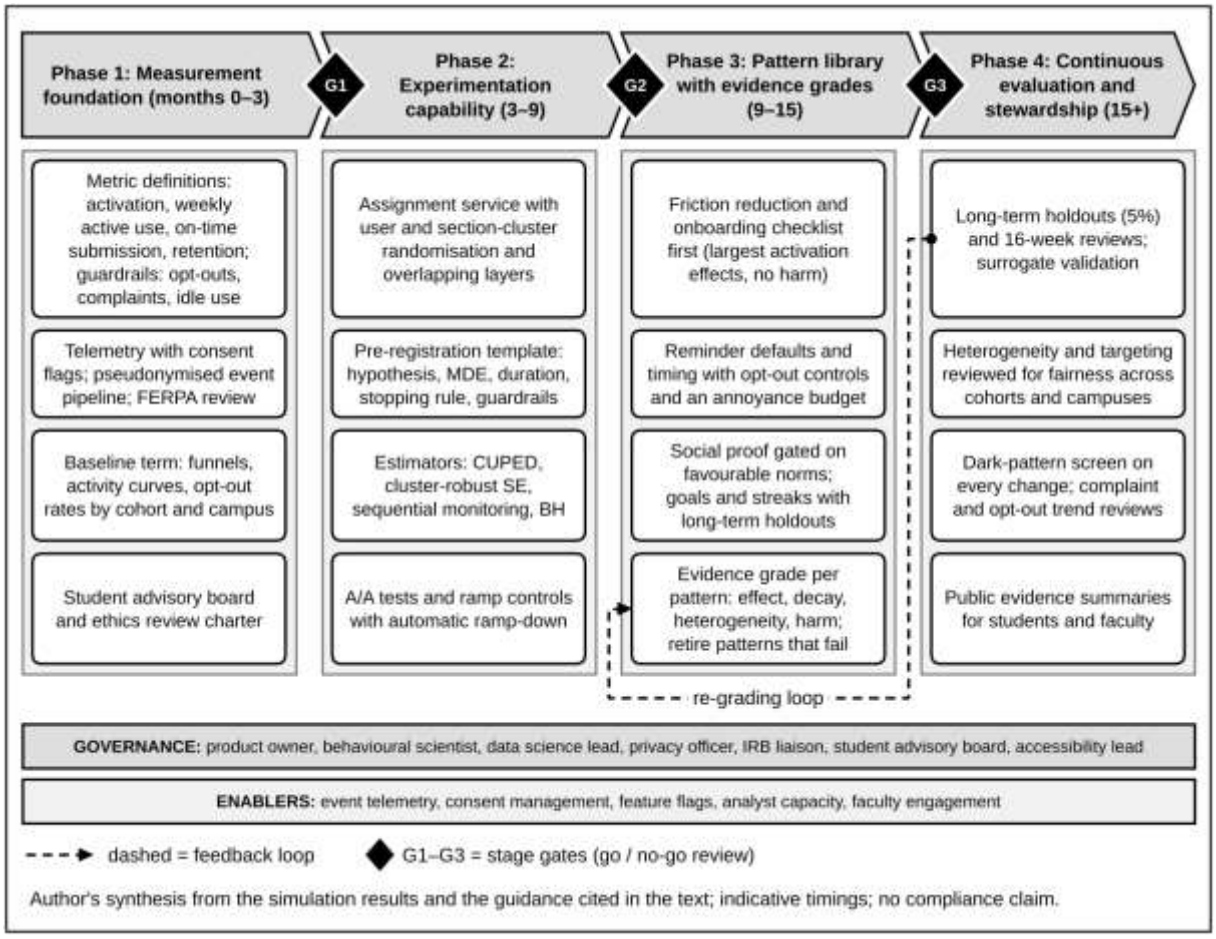
Instrument	What it requires or prohibits	Patterns and practices affected	Implication for the workflow
FERPA (U.S. Department of Education, n.d.)	Education records may be used by school officials with a legitimate educational interest; disclosure to vendors needs the school-official exception or consent	Event pipeline, pseudonymisation, vendor experimentation services, uplift models on academic data	Data-use agreements; pseudonymised events; experiments run as institutional improvement with the registrar informed
Common Rule (U.S. Department of Health and Human Services, 2018)	Research on human subjects needs IRB review unless exempt; normal educational practice in established settings is exempt (category 1)	Experiments published as research; experiments with more than minimal risk (urgency, loss framing)	An IRB determination for each campaign; exemption documented; consent where the exemption does not apply
FTC Act section 5 (U.S. Congress, 2023) and the FTC dark-pattern report (Federal Trade Commission, 2022)	Unfair or deceptive practices are unlawful; the report lists false urgency, obstruction of cancellation and hidden defaults	P10 urgency countdown; opt-out that is hard to find; P2 defaults without clear notice	A dark-pattern screen before any test; urgency only when true; one-tap opt-out; defaults disclosed
EU Digital Services Act, Article 25 (European Parliament and Council, 2022)	Online platforms must not design interfaces that deceive or manipulate users or distort their decisions	The same patterns for European campuses or vendors; nudges that impair autonomous choice	The screen applies to all markets; documented rationale per pattern
GDPR (European Parliament and Council, 2016) and the EU AI Act (European Parliament and Council, 2024)	Lawful basis and purpose limitation for behavioural data; profiling rights; AI systems used in education are high-risk	Uplift targeting and personalised timing; any model that prioritises or evaluates students	Data-protection impact assessment; human oversight; transparency about targeting
WCAG 2.2 (World Wide Web Consortium, 2023)	Interfaces must be perceivable, operable and understandable; time limits adjustable; no flashing	Countdown timers; streak and badge animations; reminder cards	Accessibility review of every pattern surface; timers adjustable
Ethics of nudging (Sunstein, 2015, 2022) and of experiments (Meyer et al., 2019; Kramer et al., 2014)	Nudges should be transparent and preserve choice; people object to experiments on them less than to untested policies, but A/B deception is resented	All patterns; the fact of experimentation	Public pattern library; plain-language notice of experimentation; student advisory board
Learning analytics ethics (Slade & Prinsloo, 2013; Prinsloo & Slade, 2017) and the NIST AI Risk Management Framework (Tabassi, 2023)	Students as agents, not data subjects; transparency and recourse; AI risks mapped, measured, managed and governed	Dashboards, risk scores and uplift models that decide who is nudged	Students can see and contest their status; model cards; a risk register

Implementation roadmap

Figure 12 arranges the findings as a four-phase roadmap with stage gates. Phase 1 (months 0 to 3) builds the measurement foundation: the metric and guardrail definitions of Table 5, a consented and pseudonymised event pipeline with a FERPA review, a baseline term that records funnels, activity curves and opt-out rates by cohort and campus, and a student advisory board with an ethics review

charter. Phase 2 (months 3 to 9) adds the experimentation capability of Figure 1: an assignment service with user and section-cluster randomisation and overlapping layers, a pre-registration template, the estimators of Table 7, and A/A tests with automatic ramp-down. Phase 3 (months 9 to 15) builds the pattern library in the order the testbed suggests: friction reduction and the onboarding checklist first, because they have the largest activation effects and no measured harm; reminder defaults and timing with opt-out controls and an annoyance budget; social proof gated on favourable norms; goals and streaks only with long-term holdouts, because their effects decay within weeks. Each pattern receives an evidence grade (effect, decay, heterogeneity, harm) and patterns that fail are retired. Phase 4 (from month 15) is continuous evaluation: long-term holdouts and 16-week reviews that validate the early surrogates, fairness reviews of targeting across cohorts and campuses, a dark-pattern screen on every change, and public evidence summaries for students and faculty. The governance group that owns the gates includes a product owner, a behavioural scientist, a data science lead, a privacy officer, an IRB liaison, the student advisory board and an accessibility lead. The timings are indicative; a campus with an existing experimentation platform would compress Phases 1 and 2 (Kohavi et al., 2020; EDUCAUSE, 2023; Dahlstrom et al., 2014).

Figure 12: Implementation roadmap for behavioural pattern evaluation in a campus product: four phases with stage gates, the re-grading loop, governance and enablers



Limitations

Table 16 lists the limitations. The first dominates: the behaviour is synthetic and the pattern mechanisms are stylised, so the effect sizes are consequences of the assumptions in Table 4 and the Monte Carlo ranges, not estimates of what any campus would observe. The ranking of patterns on activation (friction reduction and the checklist first) and the harm profile of the urgency countdown are robust within those ranges, but the ranges were chosen by the author with reference to the literature, and a different reading of that literature would move them. The testbed has one channel from activation

to later engagement (deadline salience after calendar sync), one channel of interference (course sections), one harm mechanism for reminders (annoyance against tolerance) and one for urgency (trust), and these are the channels the results describe. The estimators compared are the standard ones; always-valid sequential methods, regression adjustment with many covariates and doubly robust estimators would narrow some of the gaps reported in Section 5. The multi-campus analysis finds little between-campus heterogeneity because the campuses differ only in an intent offset; real campuses differ in calendars, platforms, populations and norms, and the I-squared values reported here are a lower bound on what a multi-campus trial would find. The regulatory mapping summarises texts as of their cited editions (Tipton, 2014; Law, 2015; Sargent, 2013).

Table 16: Limitations of the study and what would address them

Limitation	Why it matters	What would address it
Synthetic behaviour with stylised mechanisms	Effect sizes, decay rates and harms are assumptions, not estimates; the testbed cannot show which pattern works on a real campus	Field experiments on real campus products with the designs tested here; calibration of the testbed to logged behaviour
Pattern mechanisms simplified	Each pattern is one or two parameters; real implementations differ in copy, timing and placement, which the literature finds matter	Multiple implementations per pattern; a design-space study rather than one point per pattern
Interference only through sections	Spillovers through friends, residence halls and social media are absent; cluster designs on sections would not remove them	Multi-level cluster designs; exposure-mapping estimators; network data where available and permitted
One activation-to-engagement link	Downstream effects of onboarding patterns depend on an assumed salience link that the Monte Carlo analysis varies but cannot validate	Longitudinal data on activation and later engagement; mediation analysis in field data
Opt-out, complaints and trust are proxies for harm	Real harms (stress, distraction, inequity) are not modelled; effects on learning are proxied by on-time submission	Student-reported outcomes; academic outcomes; subgroup analyses with fairness metrics
One term, five campuses, no attrition	Term-to-term carry-over, cohort replacement and withdrawal are absent; campus heterogeneity is modelled only as intent offsets	Multi-term simulation; richer campus models; external validity indices (Tipton, 2014)
Estimators are standard	Sequential methods, always-valid inference, regression adjustment with many covariates and doubly robust estimators are not compared	A methods benchmark on the same testbed with those estimators
Regulatory mapping is a reading, not advice	Statutes and guidance change and vary by jurisdiction and by institution; the mapping summarises texts as of their cited editions	Legal review at each institution; monitoring of FTC, EU and state rules

Conclusion

A campus software product that adopts behavioural design patterns without an evaluation framework will learn the wrong lessons from its own data: it will see large early effects from urgency and streaks and ship them, miss the small persistent effects of reminders, social proof and personalised timing because its tests are underpowered, mistake a positive early estimate for a large effect, and count opt-outs as noise rather than as harm. The testbed built here, on which every true effect is known, shows each of these failures and the design that prevents it. Activation is moved most by removing friction and by a checklist, with effects of about 11 to 12 percentage points that carry through to later engagement only weakly; sustained engagement is moved by patterns that act on deadlines and habits,

with effects of one to two percentage points on week-16 use that take weeks to build and need thousands of students per arm to detect; the patterns with the largest short-run effects are the ones whose effects decay or whose harms are largest.

REFERENCES

[1]. Adnan, S. (2024). Human Trust in AI-Assisted Project Decision Systems: A Mixed-Methods Analysis of Explainability, Adoption, and Managerial Decision Quality. Journal of Sustainable Development and Policy, 3(04), 287-332. https://doi.org/10.63125/9c139h44

[2]. Allcott, H. (2011). Social norms and energy conservation. Journal of Public Economics, 95(9-10), 1082–1095. https://doi.org/10.1016/j.jpubeco.2011.03.003

[3]. Allcott, H., & Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. American Economic Review, 104(10), 3003–3037. https://doi.org/10.1257/aer.104.10.3003

[4]. Aminul Islam, M. (2024). Interpreting Cyber-Threat Indicators in Enterprise Memory Forensics: A Mixed-Methods Study of Machine Learning Performance and Explainability. Review of Applied Science and Technology, 14(12), 01–13. https://doi.org/10.63125/2p8hte92

[5]. Aminul Islam, M., & Md, A. (2023). Adaptive Class-Imbalance Learning for High-Sensitivity Fraud Detection in Large-Scale Digital Transaction Networks. International Journal of Business and Economics Insights, 13(09), 01-33. https://doi.org/10.63125/bmxwv741

[6]. Arnold, K. E., & Pistilli, M. D. (2012). Course Signals at Purdue: Using learning analytics to increase student success. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12) (pp. 267–270). ACM. https://doi.org/10.1145/2330601.2330666

[7]. Aronow, P. M., & Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. The Annals of Applied Statistics, 11(4), 1912–1947. https://doi.org/10.1214/16-AOAS1005

[8]. Asheke, M. (2023). Machine Learning–Based Predictive Quality Modeling for Defect Detection and Process Optimization in Advanced Manufacturing Systems. Review of Applied Science and Technology, 13(12), 01–49. https://doi.org/10.63125/thy73b16

[9]. Asheke, M., & Md Maksudul, H. (2022). AI-Driven Demand Forecasting for Sporting-Goods and Teamwear Supply Chains Under Seasonal and Tournament Demand. International Journal of Business and Economics Insights, 2(1), 73–110. https://doi.org/10.63125/ed6by557

[10]. Asheke, M., & Mohammad Badrul, A. (2022). Development Of Functionalized Cellulose Nanofiber Adsorbents for Selective Critical-Metal Recovery from Industrial Wastewater. American Journal of Scholarly Research and Innovation, 1(02), 270–320. https://doi.org/10.63125/v6fxzq09

[11]. Ashik, M. (2024). Meta-Analysis of Risk-Based Inspection Strategies for Improving Reliability and Safety Across Critical Energy Infrastructure Systems. American Journal of Advanced Technology and Engineering Solutions, 14(09), 01-43. https://doi.org/10.63125/tgttkt09

[12]. Athey, S., & Imbens, G. W. (2017). The econometrics of randomized experiments. Handbook of Economic Field Experiments, 1, 73–140. https://doi.org/10.1016/bs.hefe.2016.10.003

[13]. Athey, S., Chetty, R., Imbens, G. W., & Kang, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely (NBER Working Paper 26463). National Bureau of Economic Research. https://doi.org/10.3386/w26463

[14]. Bakshy, E., Eckles, D., & Bernstein, M. S. (2014). Designing and deploying online field experiments. In Proceedings of the 23rd International Conference on World Wide Web (WWW '14) (pp. 283–292). ACM. https://doi.org/10.1145/2566486.2567967

[15]. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[16]. Bhargava, S., & Manoli, D. (2015). Psychological frictions and the incomplete take-up of social benefits: Evidence from an IRS field experiment. American Economic Review, 105(11), 3489–3529. https://doi.org/10.1257/aer.20121493

[17]. 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. https://doi.org/10.1016/j.jebo.2020.12.022

[18]. Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. Educational Evaluation and Policy Analysis, 29(1), 30–59. https://doi.org/10.3102/0162373707299550

[19]. 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. https://doi.org/10.1016/j.jebo.2014.12.008

[20]. Castleman, B. L., & Page, L. C. (2016). Freshman year financial aid nudges: An experiment to increase FAFSA renewal and college persistence. Journal of Human Resources, 51(2), 389–415. https://doi.org/10.3368/jhr.51.2.0614-6458R

[21]. Chen, N., Liu, M., & Xu, Y. (2019). How A/B tests could go wrong: Automatic diagnosis of invalid online experiments. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19) (pp. 501–509). ACM. https://doi.org/10.1145/3289600.3291000

[22]. Clark, D., Gill, D., Prowse, V., & Rush, M. (2020). Using goals to motivate college students: Theory and evidence from field experiments. The Review of Economics and Statistics, 102(4), 648–663. https://doi.org/10.1162/rest_a_00864

[23]. Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates. https://www.routledge.com/Statistical-Power-Analysis-for-the-Behavioral-

Sciences/Cohen/p/book/9780805802832
[24]. Dahlstrom, E., Brooks, D. C., & Bichsel, J. (2014). The current ecosystem of learning management systems in higher education: Student, faculty, and IT perspectives. EDUCAUSE. https://library.educause.edu/resources/2014/9/the-current-ecosystem-of-learning-management-systems-in-higher-education-student-faculty-and-it-perspectives
[25]. Dai, H., Milkman, K. L., & Riis, J. (2014). The fresh start effect: Temporal landmarks motivate aspirational behavior. Management Science, 60(10), 2563–2582. https://doi.org/10.1287/mnsc.2014.1901
[26]. Damgaard, M. T., & Nielsen, H. S. (2018). Nudging in education. Economics of Education Review, 64, 313–342. https://doi.org/10.1016/j.econedurev.2018.03.008
[27]. Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin, 125(6), 627–668. https://doi.org/10.1037/0033-2909.125.6.627
[28]. DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. Econometrica, 90(1), 81–116. https://doi.org/10.3982/ECTA18709
[29]. Deng, A., Xu, Y., Kohavi, R., & Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM '13) (pp. 123–132). ACM. https://doi.org/10.1145/2433396.2433413
[30]. DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. Controlled Clinical Trials, 7(3), 177–188. https://doi.org/10.1016/0197-2456(86)90046-2
[31]. Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification". In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (MindTrek '11) (pp. 9–15). ACM. https://doi.org/10.1145/2181037.2181040
[32]. Dmitriev, P., Gupta, S., Kim, D. W., & Vaz, G. (2017). A dirty dozen: Twelve common metric interpretation pitfalls in online controlled experiments. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17) (pp. 1427–1436). ACM. https://doi.org/10.1145/3097983.3098024
[33]. Dobronyi, C. R., Oreopoulos, P., & Petronijevic, U. (2019). Goal setting, academic reminders, and college success: A large-scale field experiment. Journal of Research on Educational Effectiveness, 12(1), 38–66. https://doi.org/10.1080/19345747.2018.1517849
[34]. Eckles, D., Karrer, B., & Ugander, J. (2017). Design and analysis of experiments in networks: Reducing bias from interference. Journal of Causal Inference, 5(1), Article 20150021. https://doi.org/10.1515/jci-2015-0021
[35]. EDUCAUSE. (2023). 2023 EDUCAUSE Horizon Report: Teaching and learning edition. EDUCAUSE. https://library.educause.edu/resources/2023/5/2023-educause-horizon-report-teaching-and-learning-edition
[36]. Epstein, J. M. (1999). Agent-based computational models and generative social science. Complexity, 4(5), 41–60. https://doi.org/10.1002/(SICI)1099-0526(199905/06)4:5<41::AID-CPLX9>3.0.CO;2-F
[37]. European Parliament and Council. (2016). Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation). Official Journal of the European Union, L 119, 1-88. https://eur-lex.europa.eu/eli/reg/2016/679/oj
[38]. European Parliament and Council. (2022). Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2022/2065/oj
[39]. European Parliament and Council. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
[40]. Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). "Counting your customers" the easy way: An alternative to the Pareto/NBD model. Marketing Science, 24(2), 275–284. https://doi.org/10.1287/mksc.1040.0098
[41]. Federal Trade Commission. (2022). Bringing dark patterns to light (FTC staff report). Federal Trade Commission. https://www.ftc.gov/reports/bringing-dark-patterns-light
[42]. Fogg, B. J. (2009). A behavior model for persuasive design. In Proceedings of the 4th International Conference on Persuasive Technology (Persuasive '09) (pp. 1–7). ACM. https://doi.org/10.1145/1541948.1541999
[43]. Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. Review of Educational Research, 74(1), 59–109. https://doi.org/10.3102/00346543074001059
[44]. Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. Perspectives on Psychological Science, 9(6), 641–651. https://doi.org/10.1177/1745691614551642
[45]. Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. Journal of Consumer Research, 35(3), 472–482. https://doi.org/10.1086/586910
[46]. Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. American Psychologist, 54(7), 493–503. https://doi.org/10.1037/0003-066X.54.7.493
[47]. Gray, C. M., Kou, Y., Battles, B., Hoggatt, J., & Toombs, A. L. (2018). The dark (patterns) side of UX design. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18) (pp. 1–14). ACM. https://doi.org/10.1145/3173574.3174108
[48]. Gupta, S., Kohavi, R., Tang, D., Xu, Y., Andersen, R., Bakshy, E., Cardin, N., Chandran, S., Chen, N., Coey, D., Curtis, M., Deng, A., Duan, W., Forbes, P., Frasca, B., Guy, T., Imbens, G. W., Saint Jacques, G., Kantawala, P., ... Yashkov, I. (2019). Top challenges from the first practical online controlled experiments summit. ACM SIGKDD Explorations Newsletter, 21(1), 20–35. https://doi.org/10.1145/3331651.3331655
[49]. Gutierrez, P., & Gérardy, J.-Y. (2017). Causal inference and uplift modelling: A review of the literature. Proceedings

of Machine Learning Research, 67, 1–13. https://proceedings.mlr.press/v67/gutierrez17a.html
[50]. Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? A literature review of empirical studies on gamification. In 2014 47th Hawaii International Conference on System Sciences (HICSS) (pp. 3025–3034). IEEE. https://doi.org/10.1109/HICSS.2014.377
[51]. Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. Computers & Education, 80, 152–161. https://doi.org/10.1016/j.compedu.2014.08.019
[52]. Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. Computers & Education, 90, 36–53. https://doi.org/10.1016/j.compedu.2015.09.005
[53]. Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in Medicine, 21(11), 1539–1558. https://doi.org/10.1002/sim.1186
[54]. Hohnhold, H., O'Brien, D., & Tang, D. (2015). Focusing on the long-term: It's good for users and business. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15) (pp. 1849–1858). ACM. https://doi.org/10.1145/2783258.2788583
[55]. Holland, P. W. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354
[56]. Hudgens, M. G., & Halloran, M. E. (2008). Toward causal inference with interference. Journal of the American Statistical Association, 103(482), 832–842. https://doi.org/10.1198/016214508000000292
[57]. Hummel, D., & Maedche, A. (2019). How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies. Journal of Behavioral and Experimental Economics, 80, 47–58. https://doi.org/10.1016/j.socec.2019.03.005
[58]. Ifenthaler, D., & Schumacher, C. (2016). Student perceptions of privacy principles for learning analytics. Educational Technology Research and Development, 64(5), 923–938. https://doi.org/10.1007/s11423-016-9477-y
[59]. Imbens, G. W., & Rubin, D. B. (2015). Causal inference for statistics, social, and biomedical sciences: An introduction. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751
[60]. Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124
[61]. Johari, R., Koomen, P., Pekelis, L., & Walsh, D. (2017). Peeking at A/B tests: Why it matters, and what to do about it. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17) (pp. 1517–1525). ACM. https://doi.org/10.1145/3097983.3097992
[62]. Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? Science, 302(5649), 1338–1339. https://doi.org/10.1126/science.1091721
[63]. Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux. https://openlibrary.org/isbn/9780374275631
[64]. Karlan, D., McConnell, M., Mullainathan, S., & Zinman, J. (2016). Getting to the top of mind: How reminders increase saving. Management Science, 62(12), 3393–3411. https://doi.org/10.1287/mnsc.2015.2296
[65]. Kivetz, R., Urminsky, O., & Zheng, Y. (2006). The goal-gradient hypothesis resurrected: Purchase acceleration, illusionary goal progress, and customer retention. Journal of Marketing Research, 43(1), 39–58. https://doi.org/10.1509/jmkr.43.1.39
[66]. Kizilcec, R. F., Reich, J., Yeomans, M., Dann, C., Brunskill, E., Lopez, G., Turkay, S., Williams, J. J., & Tingley, D. (2020). Scaling up behavioral science interventions in online education. Proceedings of the National Academy of Sciences, 117(26), 14900–14905. https://doi.org/10.1073/pnas.1921417117
[67]. Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., & Xu, Y. (2012). Trustworthy online controlled experiments: Five puzzling outcomes explained. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12) (pp. 786–794). ACM. https://doi.org/10.1145/2339530.2339653
[68]. Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: Survey and practical guide. Data Mining and Knowledge Discovery, 18(1), 140–181. https://doi.org/10.1007/s10618-008-0114-1
[69]. Kohavi, R., Tang, D., & Xu, Y. (2020). Trustworthy online controlled experiments: A practical guide to A/B testing. Cambridge University Press. https://doi.org/10.1017/9781108653985
[70]. Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences, 111(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111
[71]. Kuh, G. D. (2008). High-impact educational practices: What they are, who has access to them, and why they matter. Association of American Colleges and Universities. https://www.aacu.org/publication/high-impact-educational-practices-what-they-are-who-has-access-to-them-and-why-they-matter
[72]. Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences, 116(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116
[73]. Kynatun, M., & Saba, A. (2024). Machine Learning Enhanced Intrusion Detection and Anomaly Classification for MIS-Driven Enterprise Network Security. International Journal of Scientific Interdisciplinary Research, 14(09), 01–41. https://doi.org/10.63125/xk6kpc72
[74]. Lally, P., van Jaarsveld, C. H. M., Potts, H. W. W., & Wardle, J. (2010). How are habits formed: Modelling habit

formation in the real world. European Journal of Social Psychology, 40(6), 998–1009. https://doi.org/10.1002/ejsp.674

[75]. Lattimore, T., & Szepesvári, C. (2020). Bandit algorithms. Cambridge University Press. https://doi.org/10.1017/9781108571401

[76]. Law, A. M. (2015). Simulation modeling and analysis (5th ed.). McGraw-Hill Education. https://www.mheducation.com/

[77]. Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. Journal of Marketing, 80(6), 69–96. https://doi.org/10.1509/jm.15.0420

[78]. Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In Proceedings of the 19th International Conference on World Wide Web (WWW '10) (pp. 661–670). ACM. https://doi.org/10.1145/1772690.1772758

[79]. Macal, C. M., & North, M. J. (2010). Tutorial on agent-based modelling and simulation. Journal of Simulation, 4(3), 151–162. https://doi.org/10.1057/jos.2010.3

[80]. Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. Computers & Education, 54(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

[81]. Madrian, B. C., & Shea, D. F. (2001). The power of suggestion: Inertia in 401(k) participation and savings behavior. The Quarterly Journal of Economics, 116(4), 1149–1187. https://doi.org/10.1162/003355301753265543

[82]. Maier, M., Bartoš, F., Stanley, T. D., Shanks, D. R., Harris, A. J. L., & Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias. Proceedings of the National Academy of Sciences, 119(31), e2200300119. https://doi.org/10.1073/pnas.2200300119

[83]. Mansura Akter, E., & Sumaiya, M. (2022). Antimicrobial Resistance Profiles of Acinetobacter Baumannii and Vibrio Cholerae in Hospital Wastewater and Adjacent Household Water Sources. American Journal of Data Science and Analytics, 3(09), 01-55. https://doi.org/10.63125/gbervt14

[84]. Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark patterns at scale: Findings from a crawl of 11K shopping websites. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), Article 81. https://doi.org/10.1145/3359183

[85]. Mathur, A., Kshirsagar, M., & Mayer, J. (2021). What makes a dark pattern... dark? Design attributes, normative considerations, and measurement methods. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21) (pp. 1–18). ACM. https://doi.org/10.1145/3411764.3445610

[86]. Md Imran, H., & Jannatul, F. (2022). Molecular and Precision Rehabilitation Molecular Mechanisms Linking Mitochondrial Dysfunction, Inflammation, and Muscle Degeneration with Functional Decline and Rehabilitation Response in Neuromuscular Disease. American Journal of Interdisciplinary Studies, 12(08), 01-44. https://doi.org/10.63125/0ym5bh79

[87]. Md Maksudul, H. (2024). Supplier-Risk Scoring and Procurement Resilience for Distributed Athletic-Apparel and Equipment Supply Networks. American Journal of Advanced Technology and Engineering Solutions, 14(09), 44-90. https://doi.org/10.63125/a9fmgj50

[88]. Md Maksudul, H., & Md Tazul Islam, C. (2022). Machine Learning-Based Cash-Flow Forecasting for Improving Financial Planning in U.S. Small and Mid-Sized Businesses. International Journal of Scientific Interdisciplinary Research, 12(12), 01–38. https://doi.org/10.63125/w3pgjx97

[89]. Md Shahid, N. (2023). Digital Twin-Based Network Reliability Modeling for LAN And WAN Infrastructure Resilience. International Journal of Scientific Interdisciplinary Research, 4(4), 577–629. https://doi.org/10.63125/ahjvxj14

[90]. Md Shahid, N. (2024). Predictive Analytics Models for Network Traffic Optimization in Distributed Enterprise Environments. Review of Applied Science and Technology, 3(01), 367-419. https://doi.org/10.63125/qm6cyf84

[91]. Md Tazul Islam, C. (2024). Explainable AI For Financial and Operational Decision Support: Improving Transparency and Managerial Trust. American Journal of Advanced Technology and Engineering Solutions, 14(12), 01-38. https://doi.org/10.63125/kwfvzw68

[92]. Md Tazul Islam, C., & Ahmed, H. C. H. R. (2023). Predictive Operations Analytics for Resource Allocation and Service-Capacity Planning in Data-Driven Organizations. Review of Applied Science and Technology, 13(02), 01–41. https://doi.org/10.63125/z3sm3089

[93]. Md. Morshedul, I., & Aminul Islam, M. (2022). A Quantitative Evaluation of Explainable Ensemble Learning for Probabilistic Risk Estimation and Decision Reliability in Enterprise Analytics. American Journal of Data Science and Analytics, 3(02), 01-38. https://doi.org/10.63125/k22shk74

[94]. Mehrotra, A., Pejovic, V., Vermeulen, J., Hendley, R., & Musolesi, M. (2016). My phone and me: Understanding people's receptivity to mobile notifications. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16) (pp. 1021–1032). ACM. https://doi.org/10.1145/2858036.2858566

[95]. Mertens, S., Herberz, M., Hahnel, U. J. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. Proceedings of the National Academy of Sciences, 119(1), e2107346118. https://doi.org/10.1073/pnas.2107346118

[96]. Meyer, M. N., Heck, P. R., Holtzman, G. S., Anderson, S. M., Cai, W., Watts, D. J., & Chabris, C. F. (2019). Objecting to experiments that compare two unobjectionable policies or treatments. Proceedings of the National Academy of Sciences, 116(22), 10723–10728. https://doi.org/10.1073/pnas.1820701116

[97]. Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., Eccles, M. P., Cane, J., & Wood, C. E. (2013). The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: Building an

international consensus for the reporting of behavior change interventions. Annals of Behavioral Medicine, 46(1), 81–

95. https://doi.org/10.1007/s12160-013-9486-6

[98]. Michie, S., van Stralen, M. M., & West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions. Implementation Science, 6(1), Article 42. https://doi.org/10.1186/1748-5908-6-42

[99]. Milkman, K. L., Beshears, J., Choi, J. J., Laibson, D., & Madrian, B. C. (2011). Using implementation intentions prompts to enhance influenza vaccination rates. Proceedings of the National Academy of Sciences, 108(26), 10415–10420. https://doi.org/10.1073/pnas.1103170108

[100]. Milkman, K. L., Gromet, D., Ho, H., Kay, J. S., Lee, T. W., Pandiloski, P., Park, Y., Rai, A., Bazerman, M., Beshears, J., Bonacorsi, L., Camerer, C., Chang, E., Chapman, G., Cialdini, R., Dai, H., Eskreis-Winkler, L., Fishbach, A., Gross, J. J., ... Duckworth, A. L. (2021a). Megastudies improve the impact of applied behavioural science. Nature, 600(7889), 478–483. https://doi.org/10.1038/s41586-021-04128-4

[101]. Milkman, K. L., Patel, M. S., Gandhi, L., Graci, H. N., Gromet, D. M., Ho, H., Kay, J. S., Lee, T. W., Akinola, M., Beshears, J., Bogard, J. E., Buttenheim, A., Chabris, C. F., Chapman, G. B., Choi, J. J., Dai, H., Fox, C. R., Goren, A., Hilchey, M. D., ... Duckworth, A. L. (2021b). A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. Proceedings of the National Academy of Sciences, 118(20), e2101165118. https://doi.org/10.1073/pnas.2101165118

[102]. Monge Roffarello, A., & De Russis, L. (2019). The race towards digital wellbeing: Issues and opportunities. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19) (pp. 1–14). ACM. https://doi.org/10.1145/3290605.3300616

[103]. O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. Journal of the American Society for Information Science and Technology, 59(6), 938–955. https://doi.org/10.1002/asi.20801

[104]. O'Brien, H. L., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. International Journal of Human-Computer Studies, 112, 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004

[105]. Oinas-Kukkonen, H., & Harjumaa, M. (2009). Persuasive systems design: Key issues, process model, and system features. Communications of the Association for Information Systems, 24(1), 485–500. https://doi.org/10.17705/1CAIS.02428

[106]. Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716. https://doi.org/10.1126/science.aac4716

[107]. Oreopoulos, P., & Petronijevic, U. (2019). The remarkable unresponsiveness of college students to nudging and what we can learn from it (NBER Working Paper 26059). National Bureau of Economic Research. https://doi.org/10.3386/w26059

[108]. Pielot, M., Church, K., & de Oliveira, R. (2014). An in-situ study of mobile phone notifications. In Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services (MobileHCI '14) (pp. 233–242). ACM. https://doi.org/10.1145/2628363.2628364

[109]. Prinsloo, P., & Slade, S. (2017). An elephant in the learning analytics room: The obligation to act. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK '17) (pp. 46–55). ACM. https://doi.org/10.1145/3027385.3027406

[110]. Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Sage. https://www.sagepub.com/shop/buy-a-book/hierarchical-linear-models-2-9230

[111]. Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5), 688–701. https://doi.org/10.1037/h0037350

[112]. Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

[113]. Sadia, A., & Md. Mosheur, R. (2023). Consumer Trust, Perceived Authenticity, and Personalized Digital Marketing: A Quantitative Decision-Science Study. International Journal of Scientific Interdisciplinary Research, 4(2), 252–295. https://doi.org/10.63125/cajkqt60

[114]. Sahami Shirazi, A., Henze, N., Dingler, T., Pielot, M., Weber, D., & Schmidt, A. (2014). Large-scale assessment of mobile notifications. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14) (pp. 3055–3064). ACM. https://doi.org/10.1145/2556288.2557189

[115]. Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. Educational Psychology Review, 32(1), 77–112. https://doi.org/10.1007/s10648-019-09498-w

[116]. Sargent, R. G. (2013). Verification and validation of simulation models. Journal of Simulation, 7(1), 12–24. https://doi.org/10.1057/jos.2012.20

[117]. Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. Psychological Science, 18(5), 429–434. https://doi.org/10.1111/j.1467-9280.2007.01917.x

[118]. Siemens, G. (2013). Learning analytics: The emergence of a discipline. American Behavioral Scientist, 57(10), 1380–1400. https://doi.org/10.1177/0002764213498851

[119]. Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22(11), 1359–1366. https://doi.org/10.1177/0956797611417632

[120]. Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. American Behavioral Scientist, 57(10),

1510–1529. https://doi.org/10.1177/0002764213479366
[121]. Stabak, D., & Md Shahid, N. (2022). Empirical Assessment of Uncertainty-Aware Real-Time Decision-Making for Reliable and Resilient Cyber-Physical Systems. International Journal of Scientific Interdisciplinary Research, 1(01), 432–477. https://doi.org/10.63125/q3r4kk16
[122]. Sumaiya, M. (2024). Hospital Wastewater as A Reservoir of Antimicrobial Resistance Genes: Implications for Public Health and Genomic Surveillance in The United States. Review of Applied Science and Technology, 3(04), 448–508. https://doi.org/10.63125/ahvas976
[123]. Sunstein, C. R. (2015). The ethics of nudging. Yale Journal on Regulation, 32(2), 413–450. https://www.yalejreg.com/print/the-ethics-of-nudging/
[124]. Sunstein, C. R. (2022). Sludge audits. Behavioural Public Policy, 6(4), 654–673. https://doi.org/10.1017/bpp.2019.32
[125]. Syed Feroz, A., & Md Mahamud Pervaz, P. (2022). Root-Cause Analysis and FMEA-Driven Reliability Improvement in Gas Production Facilities: A Multi-Year Empirical Study. International Journal of Scientific Interdisciplinary Research, 12(10), 01–53. https://doi.org/10.63125/70msp878
[126]. Syed Feroz, A., & Shabbir, A. (2022). Finite Element-Based Thermomechanical Optimization of Pressure Components for Extending Industrial Boiler Service Life. Journal of Sustainable Development and Policy, 12(12), 01-51. https://doi.org/10.63125/x24pht28
[127]. Taher, M. A., Sikder, M. A., Ahmed, H., & Bashar, M. (2024). Integrating IOT and Big Data Analytics for Enhanced Supply Chain Performance in Industrial Engineering Sectors: A Cross-Market Study. International Journal of Science and Engineering, Volume 1, 1-14.
[128]. Tang, D., Agarwal, A., O'Brien, D., & Meyer, M. (2010). Overlapping experiment infrastructure: More, better, faster experimentation. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10) (pp. 17–26). ACM. https://doi.org/10.1145/1835804.1835810
[129]. Thaler, R. H., & Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Yale University Press. https://openlibrary.org/books/OL10319463M
[130]. Tinto, V. (1993). Leaving college: Rethinking the causes and cures of student attrition (2nd ed.). University of Chicago Press. https://eric.ed.gov/?id=ED371658
[131]. Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. Journal of Educational and Behavioral Statistics, 39(6), 478–501. https://doi.org/10.3102/1076998614558486
[132]. Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. The Quarterly Journal of Economics, 106(4), 1039–1061. https://doi.org/10.2307/2937956
[133]. U.S. Congress. (2023). Unfair methods of competition unlawful; prevention by Commission (15 U.S.C. 45, United States Code, 2023 edition). U.S. Government Publishing Office. https://www.govinfo.gov/content/pkg/USCODE-2023-title15/html/USCODE-2023-title15-chap2-subchapI-sec45.htm
[134]. U.S. Department of Education. (n.d.). Family Educational Rights and Privacy Act (FERPA). U.S. Department of Education. https://www.ed.gov/laws-and-policy/ferpa
[135]. U.S. Department of Health and Human Services. (2018). Protection of human subjects (45 CFR Part 46). Electronic Code of Federal Regulations. https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46
[136]. Ugander, J., Karrer, B., Backstrom, L., & Kleinberg, J. (2013). Graph cluster randomization: Network exposure to multiple universes. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13) (pp. 329–337). ACM. https://doi.org/10.1145/2487575.2487695
[137]. Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839
[138]. Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. Science, 331(6023), 1447–1451. https://doi.org/10.1126/science.1198364
[139]. Wood, W., & Neal, D. T. (2007). A new look at habits and the habit-goal interface. Psychological Review, 114(4), 843–863. https://doi.org/10.1037/0033-295X.114.4.843
[140]. World Wide Web Consortium. (2023). Web Content Accessibility Guidelines (WCAG) 2.2. W3C. https://www.w3.org/TR/WCAG22/
[141]. Xu, Y., Chen, N., Fernandez, A., Sinno, O., & Bhasin, A. (2015). From infrastructure to culture: A/B testing challenges in large scale social networks. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15) (pp. 2227–2236). ACM. https://doi.org/10.1145/2783258.2788602