SCIENTIFIC CULTURE

www.sci-cult.com

*SCIENTIFIC CULTURE, Vol. 9, No. 2, (2023), pp. 144-176*
*Open Access. Online & Print*

**DOI: 10.5281/zenodo.24320690**

# MAXIMUM COVERAGE AS A MEASURE OF TARGETING FOR STUDENTS AT RISK, TESTED UNDER A FROZEN PLAN

**Luis Eduardo Muñoz Guerrero**[1*]

*[1]PhD, Universidad Tecnológica de Pereira, Facultad de Ingenierías, Pereira, Risaralda, Colombia*

*Received: 20/03/2023*
*Accepted: 30/04/2023*

*Corresponding Author: Luis Eduardo Muñoz Guerrero*

## ABSTRACT

***We formalise the choice of groups of students to support as cardinality-constrained maximum coverage and restate the greedy guarantee of at least 63% of the optimum. On public data from 4,424 students in 17 programmes, we choose groups on nine and test them on eight under a privately tagged plan. The frozen primary analysis is not valid: coverage, the share of test-programme dropouts covered, is 100.0% because the groups reach every student, as training data showed; the placebo test and the negative control on graduates fail. Post hoc, with groups capped at 10% instead of 25% of training students, the greedy rule covers 70.9% of dropouts, reaching 55.4% of students, with a dropout share of 37.1% (base rate 29.0%; random groups 38.2%, or 35.5% at equal reach). The negative control fails; the placebo test passes, although permuted labels reproduce most of the gain over random groups; and the dropout share falls to 32.5% without three attributes of uncertain timing. Across 144 post hoc specifications, the negative control passes in one, at extreme settings, and in a second under one of five seeds. Coverage alone does not establish targeting; reach, precision and a negative control matched on reach should accompany it.***

**KEYWORDS:** Submodular Functions, Approximation Algorithms, NP-Hardness, Integer Programming, Learning Analytics, Higher Education, Open Educational Data, Algorithmic Benchmarking, Systems Engineering, Retention Analytics, Resource Allocation, Placebo Tests.

Copyright: © 2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License. (https://cre-ativecommons.org/licenses/by/4.0/).

## 1. INTRODUCTION

Retention is a central concern of distance and higher education. Learning analytics has been described as a field that draws on technologies and methodologies from several research areas (Siemens, 2013). Wolff et al. (2013), for example, estimate the risk of failure for individual students from virtual-learning-environment activity, so that tutors can identify struggling students and respond in time.

We consider an institution that wants to respond to students at risk and can fund only a limited number of group activities, such as workshops or cohort-level reminders, each of which reaches students who share some characteristics. Choosing which groups to fund, under a fixed number of activities, is an optimisation problem that is distinct from estimating individual risk: it asks for a set of groups that together cover as many students at risk as possible. This is the maximum coverage problem with a cardinality constraint. Feige (1998) notes that maximum k-cover, which selects k sets to cover as many elements as possible, is NP-hard, and shows that, unless P = NP, no polynomial-time algorithm guarantees more than $1 - 1/e$ of the optimum.

The approximation and submodular-optimisation literature reviewed in Section 2.1 leaves two practical questions open. First, does a selection made on one set of programmes cover students at risk in programmes that were not used to make it? Second, does coverage measure targeting, or does it reward broad groups? Under our analysis plan, which was fixed before the outcomes of the confirmatory cohort were analysed (Section 5.6), a positive answer to both questions requires (i) the selection to cover at least 2 percentage points more dropouts than the top-dropout-rate rule (B1), with an interval whose lower limit is above zero, and (ii) the selection to pass the two validity tests that the code applies: a negative control, under which the selection may cover at most 2 points more graduates, who did not drop out, than random groups (upper interval limit), and a placebo test that repeats the selection on permuted training labels. We say that a selection targets dropouts if, at the same reach (the share of students that its groups contain), it covers more dropouts than random subsets of the same family of candidate groups (Section 5.4). The plan's contrasts do not measure this directly: coverage grows with reach, so the primary contrast favours broad rules, whereas a negative control against random groups of the same budget favours narrow ones (Section 6.3). Three labels are used throughout: confirmatory refers to the frozen analysis as the plan specifies it, exploratory to any other analysis, and post hoc to analyses that the plan does not name and that were decided after the frozen result had been seen.

This article makes three contributions. It formulates group-level targeting of students at risk as cardinality-constrained maximum coverage and notes that the decision version is NP-complete for general families of sets, a standard result (identity reduction from maximum k-cover; Section 3); in a limited search (Section 2.3), we found no earlier formulation of this kind, although subgroup discovery may be related (Section 2.2). The greedy rule adds, at each step, the group with the most dropouts not yet covered; its guarantee of at least $1 - 1/e$, about 63% of the optimum, is restated from the literature (Section 4). The article evaluates the selection on a public dataset of 4,424 students in 17 undergraduate programmes (Realinho et al., 2022), under an analysis plan frozen by a version tag, with placebo tests and a negative control (Sections 5 and 6). It reports post hoc reach, precision and checks across programmes alongside coverage. The guarantee's numerical check on the training cohort (Section 6.5), and the measured cost of the greedy rule and of an exact integer programme, with seed-dependent time-outs on synthetic families (Section 7), are supporting material and not contributions.

The main empirical finding is a caution rather than a success. The frozen primary analysis is not valid, because its groups reach every student; no confirmatory or other planned analysis supports targeting, and every positive result on targeting in this article is post hoc. A post hoc variant (Amendment 1), with groups capped at 10% rather than 25% of training students, covers more dropouts than B1 and random groups (70.9% against 33.6% and 45.9%), but it also reaches many students who did not drop out, and its negative control fails at the planned budget. Coverage is therefore not a sufficient measure of targeting: reach and precision should be reported with it, together with a negative control matched on reach, because a rule can reach more students than random groups of the same budget (Section 6.3).

## 2. BACKGROUND AND RELATED WORK

Let E be a finite universe with non-negative weights $w_e$, and let G be a family of subsets of E. For a selection $S \subseteq G$, the coverage function $f(S)$ is the total weight of the union of the sets in S. Lemma 1 shows that f is monotone and submodular: adding a set to a larger selection never helps more than adding it to a smaller one. Greedy and local-search algorithms are treated in the approximation literature (Williamson and Shmoys, 2011),

algorithms for optimising submodular functions are discussed by Krause and Golovin (2014), and an analysis of approximation algorithms for maximising submodular set functions appears in Nemhauser et al. (1978). Submodular maximisation under a knapsack constraint, in which the selected sets have different costs, is studied by Sviridenko (2004). The quality of a greedy k-stage covering algorithm, in which each selected subset must satisfy structural constraints, is analysed by Hochbaum and Pathria (1998). The word ‘group’ has another meaning in the maximum-coverage literature, where group budget constraints are studied by Chekuri and Kumar (2004). Choosing a few individuals or locations to maximise a spread or a detection rate is also studied in influence maximisation, which asks which set of individuals to target (Kempe et al., 2003), and in cost-effective outbreak detection, which asks where to place sensors (Leskovec et al., 2007). Budgeted maximum coverage, in which sets carry costs, is studied by Khuller et al. (1999). Maximal covering location problems, which we take to have the same structure as maximum coverage, appear in facility location (Church and ReVelle, 1974).

Early prediction of student performance in higher education has been examined in a case study (Martins et al., 2021), and the public dataset analysed here is used to build models that predict academic performance and dropout (Realinho et al., 2022). These studies, like the work of Wolff et al. (2013) cited in Section 1, predict outcomes for individual students. Course Signals, an early-intervention system at Purdue, relied on grades, demographic characteristics and past academic history to predict students’ performance, and it gave students real-time feedback (Arnold and Pistilli, 2012). An early-alert initiative predicted academic risk, examined how its models carried over to partner institutions and reported interventions on students at risk (Jayaprakash et al., 2014), and uplift modelling has been used to tailor retention programmes to the students more likely to be retained if targeted (Olaya et al., 2020). Subgroup discovery, which may be related, has been surveyed by Herrera et al. (2011) and applied to e-learning data (Romero et al., 2009); we do not review it. To our knowledge, the other sources reviewed for this article do not formulate the choice of group-level interventions under a fixed budget as a coverage problem or test that choice against a negative control on programmes that were not used to make it. Section 2.3 describes the search behind this statement.

The search fixed its questions and its filters before it began. Crossref was searched with eight queries, each in two publication windows: up to December 2022, the cut-off for the sources cited here, and after it, whose records the cut-off excludes. It returned 189 distinct records published by 2022, and in all 16 query-and-window passes it returned the 25 records requested, so the screening covers only the first 25 records of each pass. arXiv was searched with eight phrase queries, of which six returned results, giving 96 distinct titles published by 2022. Titles, and abstracts where available, were screened for terms related to dropout, retention, targeting, budget, coverage, greedy and submodular methods. The search plan also named OpenAlex, DBLP and Semantic Scholar. In the search environment, OpenAlex and Semantic Scholar returned rate-limit errors and DBLP an automated-access challenge, so their records are not included. No education-specific database, such as the Education Resources Information Center (ERIC), was searched. A second check with the same eight queries returned no arXiv records and errors from OpenAlex and DBLP. The screening supports the statement in Section 2.2 for the reviewed records; it does not establish that no other study exists.

## 3. PROBLEM FORMULATION

Table I lists the notation. The universe E is the set of students in the training cohort. Each element e has a weight we equal to 1 if the student dropped out, and 0 otherwise; E+ denotes the elements with positive weight, that is, the training dropouts. Each candidate group g is defined by one attribute value or by two attribute values (for example, daytime students aged 25 or older, or non-displaced students whose tuition fees are in arrears); as a subset of E it holds the training students with those values, and the same definition gives its members in the confirmatory cohort. G is the family of candidate groups that meet the size limits of Section 5.2. The budget K is the number of group interventions that can be funded; the cost of groups of different size is not modelled (Section 9).

***Table I. Main notation used in the problem formulation and in the evaluation; other symbols are defined where they first appear.***

| Symbol | Meaning | Domain or value |
|---|---|---|
| E | Universe: students in the training cohort | finite set |
| we | Weight of student e: 1 if the student dropped out, 0 otherwise | {0, 1} |

| | | |
|---|---|---|
| E+ | Training dropouts: the elements of E with positive weight | subset of E |
| G | Family of candidate groups | subsets of E |
| g | A candidate group: students who share one value or two values of attributes | subset of E |
| K | Budget: number of groups that can be funded | positive integer |
| S | Selected family of groups | subset of G; $\|S\| \le K$ |
| f(S) | Covered weight: total weight of the union of the groups in S | non-negative real |
| a(e), A | Value of attribute a for student e; the eight attributes used in pairs | attribute values; set of attributes |
| n, τ | Number of training students; size limit as a share of them | positive integer; (0, 1] |
| E′, D, Grad | Confirmed students, dropouts and graduates: those of the confirmatory cohort | subsets of E′ |
| U(S) | Confirmed students whose attribute values place them in at least one group of S | subset of E′ |
| R, S(r) | Number of random draws of B2; the r-th draw | R = 1,000 |
| B, me(b) | Number of bootstrap replicates; multiplicity of student e in replicate b, the number of times it is drawn | B = 2,000; non-negative integer |
| Π, Sπ | Number of placebo permutations; greedy selection on the π-th permutation | Π = 1,000 |
| J, S̃(j) | Number of random subsets of the equal-reach comparator; the j-th subset | J = 40,000 |
| h, Qh | Half-width of the reach window; indices of the subsets kept | 0.005 or 0.01 (half a point or 1 point) |

The objective is the total weight of the students covered by the selected groups:

$$f(S) = \sum_{e \in \cup S} w_e ,$$
$$\max_{S \subseteq G} f(S) \quad \text{subject to} \quad |S| \le K \tag{1}$$

Equation 1 defines the covered weight f, where ∪S is the union of the groups in S, and the problem of choosing at most K groups to maximise it. The decision version of the problem asks whether some S with $|S| \le K$ attains $f(S) \ge \theta$ for a threshold θ.

Theorem 1. The decision version of cardinality-constrained maximum coverage is NP-complete for general families of sets. For every fixed $\varepsilon > 0$, no polynomial-time algorithm achieves a ratio of $1 - 1/e + \varepsilon$ unless P = NP (Feige, 1998). The problem on the attribute-defined family of Section 5.2 is a restriction of the general one; its complexity is not established here.

Proof. Membership in NP is immediate: a selection S is a certificate, and f(S) can be computed in polynomial time. For hardness, take an instance (X, F, k) of maximum k-cover and transform it: E = X, we = 1 for every element, G = F and K = k. The construction takes time linear in the size of the instance. A selection with $f(S) \ge \theta$ exists in the new instance exactly when at most k sets of F cover at least θ elements, so the reduction preserves the decision and the approximation ratio of unit-weight instances. Maximum k-cover is NP-hard, and its approximation threshold is $1 - 1/e$ under the assumption P ≠ NP (Feige, 1998). ■

## 4. ALGORITHMS

Algorithm 1 adds, at each step, the group with the most training dropouts not yet covered, that is, the largest marginal gain in covered weight. Ties are broken by the group name that comes first in code-point order (uppercase letters before lowercase), which makes the output deterministic. A group's name joins the column header of the data file and the value code with an equals sign (for example, Debtor=1); the age band, which is not a column, is named tramo_edad (Spanish for age band), with values 17-19, 20-21, 22-24 and 25 o mas (25 or older); and the two names of a pair are joined by an ampersand between spaces, in the order of the paired attributes listed in Section 5.2. Section 6.1 reports how much this tie-breaking order matters on the frozen design.

**Algorithm 1. Greedy coverage rule**

```
Input: family G of groups; weights w of the students; budget K
Output: selection S with |S| ≤ K
S ← empty family;  C ← empty set
for t = 0 to min(K, |G|) − 1 do
   g ← group in G minus S with largest gain: sum of w_e over e in g minus C
   ties: take the group name that comes first in code-point order
   if the gain of g is 0 then stop
   S ← S union {g};  C ← C union g
end for
return S
```

Each step of Algorithm 1 computes the marginal gain of every remaining group by scanning its members once, keeps the covered students in a hash set and visits the remaining groups in code-point order of their names, which fixes the tie-breaking. The rule therefore runs in expected time

$$O\left(K\left(\sum_{g\in G}|g| + |G|\log|G|\right)\right) \tag{2}$$

The second term comes from sorting the remaining groups at each step, which a single initial sort would avoid; it is smaller than the first here, because every group has at least 30 students. For a fixed family, the first term grows in proportion to the number of students; Section 7 measures this growth on replicated cohorts.

Lemma 1. The coverage function f is monotone and submodular: for all S ⊆ T ⊆ G and every g ∈ G \ T,

$$f(S) \le f(T), \qquad f(S\cup\{g\}) - f(S) \ge f(T\cup\{g\}) - f(T) \tag{3}$$

Proof. Monotonicity holds because adding a set can only add elements to the union, and the weights are non-negative. For submodularity, let S ⊆ T ⊆ G and g ∈ G \ T. The marginal gain of g with respect to S is the weight of the elements of g not covered by S, which is at least the weight of the elements of g not covered by T, because the set covered by T contains the set covered by S and the weights are non-negative. ■

Theorem 2. Let SK be the selection produced by the greedy rule with budget K, and let S∗ be an optimal selection with |S∗| ≤ K. Assume that the weights are non-negative and that the family is finite. Then f(SK) ≥ (1 − 1/e) f(S∗). The binary weights and finite families of this study satisfy these assumptions.

Proof. In this proof, for a set C of elements, f(C) denotes its total weight, so that f(S) = f(∪S) for a selection S. Let Ct be the union of the first t selected sets, with C0 = ∅, and let δt be the gain of the (t+1)-th selection, so that f(Ct+1) − f(Ct) = δt. Write S∗ = {o1, …, oℓ} with ℓ ≤ K. Then

$$f(S^*) - f(C_t) \le \sum_{i=1}^{\ell}\left[f(C_t\cup o_i) - f(C_t)\right] \le \ell\,\delta_t \le K\,\delta_t \tag{4}$$

where the first inequality bounds the union by the sum of the marginal gains, using monotonicity and submodularity (Lemma 1); the second uses the greedy choice, which maximises the gain at each step; and the third uses ℓ ≤ K. Hence

$$f(C_{t+1}) \ge f(C_t) + \frac{f(S^*) - f(C_t)}{K} \tag{5}$$

Subtracting both sides from f(S∗) gives f(S∗) − f(Ct+1) ≤ (1 − 1/K)(f(S∗) − f(Ct)). Iterating this inequality from t = 0 to K − 1 gives

$$f(S^*) - f(C_K) \le \left(1 - \frac{1}{K}\right)^{K} f(S^*) \le \mathrm{e}^{-1} f(S^*) \tag{6}$$

Because (1 − 1/K)K ≤ 1/e. Therefore f(CK) ≥ (1 − 1/e) f(S∗). If the greedy rule stops early, every remaining group has zero gain, so every element of positive weight that a group contains is already covered, and the same holds if it runs out of groups; then f(Ct) is the weight that all of G covers, which is

at least f(S∗), and the bound holds. ■

Theorem 2 is a worst-case guarantee; the proof above follows the standard argument (Nemhauser et al., 1978). The guarantee does not give the actual ratio of the greedy value to the optimum on a given instance. Rearranging Equation 6 also gives, when f(S∗) > 0, a bound for each budget, which equals 1 at K = 1 and decreases towards 1 − 1/e as K grows:

$$\gamma_K = \frac{f(S_K)}{f(S^*)} \geq 1 - \left(1 - \frac{1}{K}\right)^K \geq 1 - \frac{1}{\mathrm{e}} \tag{7}$$

Section 6.5 reports the ratio γK on the training cohort.

An exact optimum is obtained from the following integer programme. Binary variables xg indicate whether group g is selected, and continuous variables ye in [0, 1] stand for the coverage of element e in E+.

$$\begin{aligned} \max_{x,y} \quad & \sum_{e \in E^+} w_e \, y_e \\ \text{s.t.} \quad & y_e \leq \sum_{g \in G \,:\, e \in g} x_g \,, \qquad e \in E^+, \\ & \sum_{g \in G} x_g \leq K, \\ & x_g \in \{0,1\}, \qquad 0 \leq y_e \leq 1. \end{aligned} \tag{8}$$

Because each xg is binary, ye = min(1, the sum of xg over the groups g that contain e) at an optimum, so ye is integral there and declaring it continuous in [0, 1] does not change the optimum. The model has one row per element of E+ that belongs to at least one candidate group, and one row for the budget. We solve it with HiGHS, an open-source solver, through the milp interface of SciPy. Huangfu and Hall (2018) describe the dual revised simplex method, which we take to be the core of HiGHS. The solver has a time limit of 60 s in the analyses of Section 6 and of 120 s in the benchmark and stress tests of Section 7, which also reports checks of the optimal values. The general problem is NP-hard (Theorem 1), so, unless P = NP, no polynomial-time exact algorithm exists for it; for the real frozen family this does not bind, because the exact model solves it in milliseconds, although random synthetic families of the same size can reach the 120 s limit (Section 7).

## 5. DATA AND EVALUATION DESIGN

The dataset Predict Students' Dropout and Academic Success (DOI 10.24432/C5MC89) contains 4,424 student records. It aggregates several disjoint data sources and is part of a learning-analytics tool developed at the Polytechnic Institute of Portalegre, Portugal (Realinho et al., 2022). The outcome has three classes, recorded at the end of the normal duration of the programme: graduate, enrolled or dropout. The data descriptor reports 35 attributes; the file, with 37 columns, and the metadata of the UCI Machine Learning Repository list 36 attributes and the outcome, and we follow the file's count. The records refer to students enrolled between the academic years 2008/2009 and 2018/2019 in 17 undergraduate programmes, and they contain no missing values; the dataset is distributed under a Creative Commons Attribution 4.0 (CC BY 4.0) licence (Realinho et al., 2022). The file was downloaded from the UCI Machine Learning Repository, and its SHA-256 digest, a1b1a6531bbb93a5c7fdf0093b47172776652a4ffb12342fb79655c85b74801b, was checked against the stored value by the download script before the file was used; the analysis reads the file without repeating that check.

All 4,424 records were assigned to one of the two cohorts; none was excluded. The training cohort contains 1,906 students in nine programmes, and the confirmatory cohort contains 2,518 students in eight programmes. The partition is fixed by programme code: the nine programmes with the smallest codes form the training cohort, and the other eight form the confirmatory cohort. The codes are the dataset's programme identifiers (its Course variable; their names are in the note to Table II). The split does not follow fields: the daytime and evening programmes of Management and of Social Service fall in different cohorts. The two cohorts differ in composition. The median age at enrolment is 19 years in the confirmatory cohort, against 21 years in the training cohort, and the confirmatory cohort has a larger share of women and of scholarship holders (Table II). The partition by programme code produces this difference, and it limits what the results can say about other populations.

***Table II. Characteristics of the two cohorts, computed from the file.***

| Characteristic | Training cohort | Confirmatory cohort |
|---|---|---|
| Students | 1,906 | 2,518 |
| Degree programmes | 9 | 8 |
| Median age at enrolment (years) | 21 | 19 |
| Women (Gender = 0) | 54.9% | 72.4% |
| Scholarship holders (Scholarship holder = 1) | 18.4% | 29.7% |
| Dropouts | 692 | 729 |
| Graduates | 806 | 1,403 |
| Enrolled | 408 | 386 |

Training-cohort dropouts are the observed outcome used to calibrate the design; confirmed dropouts are the outcome analysed after the plan was frozen. Programme codes and names (dataset variable description): training, 33 Biofuel Production Technologies, 171 Animation and Multimedia Design, 8014 Social Service (evening attendance), 9003 Agronomy, 9070 Communication Design, 9085 Veterinary Nursing, 9119 Informatics Engineering, 9130 Equinculture and 9147 Management; confirmatory, 9238 Social Service, 9254 Tourism, 9500 Nursing, 9556 Oral Hygiene, 9670 Advertising and Marketing Management, 9773 Journalism and Communication, 9853 Basic Education and 9991 Management (evening attendance).

The groups are defined by 18 attributes that the analysis plan treats as known at enrolment: marital status, application mode and order, daytime or evening attendance, previous qualification, nationality, the mother's and father's qualifications and occupations, displacement, educational special needs, debtor status, tuition-fee payment status, gender, scholarship holder, international status and age band. The age band takes four values from the age at enrolment: 17 to 19, 20 to 21, 22 to 24, and 25 or older, with inclusive limits. The 12 curricular-unit variables of the first and second semesters (credited, enrolled, evaluations, approved, grade and without evaluations) are measured after enrolment, so they are excluded. The macroeconomic indicators, Course (the programme identifier) and the continuous grades (admission grade and previous-qualification grade) are not used to form groups. The dataset does not record when each attribute was measured. Its descriptor groups the attributes into demographic, socioeconomic, macroeconomic and academic data at enrolment, and academic data at the end of the first and second semesters (Realinho et al., 2022), without saying when debtor status, tuition-fee payment status or scholarship holder were recorded. If recorded after enrolment, these three uncertain-timing attributes may already reflect the dropout they are meant to anticipate; the sensitivity analysis S1 (Section 6.3) rebuilds the candidate groups and the comparators B1 and B2 without them.

The codes that label the selected groups in Figure 1(c) mean, in the dataset's variable descriptions: gender 1 for male and 0 for female; displacement, which marks students displaced from their homes (Realinho et al., 2022), and debtor status 1 for yes and 0 for no; tuition-fee payment status 1 for fees paid and 0 for fees in arrears; daytime or evening attendance 1 for daytime and 0 for evening; application order from 0 for the first choice to 9 for the last choice; father's occupation 4 for administrative staff; and mother's occupation 2 for specialists in intellectual and scientific activities. The age bands are defined by this study.

Each candidate group is defined by one value of one of the 18 attributes, or by a pair of values of two attributes from a fixed set of eight: gender, scholarship holder, debtor status, tuition-fee payment status, daytime or evening attendance, age band, displacement and international status. The analysis plan fixes this set. A group is kept only if it has at least 30 training students and does not exceed a size limit, set as a share of the training cohort: 25% in the frozen design and 10% in Amendment 1 (Section 5.6). The plan sets the limit so that no group is almost universal, which would saturate coverage. With a(e) the value of attribute a for student e, where a ranges over the 18 attributes, n the number of training students, τ the size limit and A the set of eight attributes used in pairs, the family is

$$G_\tau = \left\{ g \in G^{(1)} \cup G^{(2)} : 30 \le |g| \le \lfloor \tau n \rfloor \right\},$$
$$G^{(1)} = \{\{e \in E : a(e) = v\}\}_{a,v}, \qquad G^{(2)} = \{\{e \in E : a(e) = v,\ a'(e) = u\}\}_{a,a' \in A,\ a \neq a';\ v,u} \tag{9}$$

In each design, G is Gτ with that design's limit. Both gender groups exceed the 25% limit, so gender enters only through pairs. With the 25% limit the family has 143 candidate groups; with the 10% limit

it has 86. The limits allow training groups of up to 190 students (10%) or 476 (25%) of the 1,906 training students, the largest whole numbers below those shares. The plan and the amendment state these limits rounded up, as 477 and 191; no candidate group has exactly those sizes, so the rounded and the whole-number limits give the same families. The frozen family has 58 one-value groups and 85 pair groups, and the family of Amendment 1 has 37 and 49. Many values of the parents' qualifications and occupations have fewer than 30 training students and form no group, and nationality and educational special needs form none at all: their most common value exceeds both size limits, and their other values have fewer than 30 training students.

We compare the greedy selection with three comparators. B1 takes the K groups with the highest dropout rate in the training cohort, ignoring overlaps, and breaks ties by group name in code-point order; it is the reference rule of the analysis plan. B2 draws K groups at random without replacement from the same family, and is averaged over 1,000 draws with a fixed seed (the starting value of the random-number generator). B3 is the exact optimum on the training cohort, evaluated on the confirmatory cohort; it is a reference for the training objective, not a competing method. As exploratory comparators, we also report the K largest groups in the family, which ignore both labels and overlaps, and the greedy rule applied with equal weights to all training students, which greedily maximises the number of training students reached and ignores outcomes.

The outcome of interest is dropout. The coverage of a selection S in the confirmatory cohort is the share of confirmed dropouts (dropouts in the confirmatory cohort) who belong to at least one group of S, where group membership is computed from the attribute values alone. As a negative control (Lipsitch et al., 2010), the plan's term, we measure the same share for graduates, who are presumed not to be at risk; in diagnostic terms, it checks specificity.

$$\begin{aligned} \mathrm{cov}_D(S) &= \frac{|U(S)\cap D|}{|D|}, \qquad \overline{\mathrm{cov}}_D(\mathrm{B2}) = \frac{1}{R}\sum_{r=1}^{R}\mathrm{cov}_D\left(S^{(r)}\right), \\ \mathrm{H1} &= \mathrm{cov}_D\left(S_{\mathrm{greedy}}\right) - \mathrm{cov}_D(S_{\mathrm{B1}}), \qquad \mathrm{H2} = \mathrm{cov}_D\left(S_{\mathrm{greedy}}\right) - \overline{\mathrm{cov}}_D(\mathrm{B2}), \\ \mathrm{N1} &= \mathrm{cov}_{\mathrm{Grad}}\left(S_{\mathrm{greedy}}\right) - \overline{\mathrm{cov}}_{\mathrm{Grad}}(\mathrm{B2}) \end{aligned} \tag{10}$$

Equation 10 defines the coverage used throughout and the three contrasts of the plan, where D and Grad are the sets of confirmed dropouts and graduates, U(S) is the set of confirmed students whose attribute values place them in at least one group of S, Sgreedy and SB1 are the selections of the greedy rule and of B1, and S(1), …, S(R) are the R = 1,000 random draws of B2; the average for graduates is defined in the same way. H1 is the primary contrast, H2 the secondary contrast and N1 the negative control. The outcome labels of the confirmatory cohort are not used to choose groups in the frozen or the amended analysis, although the decision to amend came after the frozen result (Section 5.6). A rule deployed on new students would apply the chosen groups to the attributes recorded at enrolment, and the confirmatory cohort plays that role here. With dropouts as positives and graduates as negatives, coverage of dropouts is the sensitivity (recall) of a selection and coverage of graduates its false-positive rate (one minus specificity), the terms in which dropout flags have been compared (Bowers et al., 2013); N1 compares that false-positive rate with that of random groups. Three further quantities describe the students that the rule would contact: reach, the share of all confirmed students (the set E′) covered by the selection; precision, the dropout share among the covered students; and lift, precision divided by the base rate of dropout in the cohort:

$$\begin{aligned} \mathrm{reach}(S) &= \frac{|U(S)|}{|E'|}, \qquad \mathrm{prec}(S) = \frac{|U(S)\cap D|}{|U(S)|}, \qquad \mathrm{lift}(S) = \frac{\mathrm{prec}(S)}{|D|/|E'|}, \\ \mathrm{cov}_D(S) &= \mathrm{reach}(S)\,\mathrm{prec}(S)\,\frac{|E'|}{|D|} \end{aligned} \tag{11}$$

The last identity holds for a single selection. For B2, coverage and reach are means over the draws, and precision and lift are computed from them through the identity, as for the random subsets of Section 6.3; this ratio of averages is the reach-weighted mean of the per-draw precisions, which is below their plain mean when draws that reach fewer students have higher precision. Reach, precision, lift and the base rate count all confirmed students, including those still enrolled, whereas the two coverages use only dropouts and graduates; Section 6.3 also reports precision without the enrolled

students. Reach, precision and lift are descriptive and post hoc: the analysis plan does not name them. Differences between two shares are reported in percentage points, written as points or pp. A rule with high lift contacts students who are more likely to drop out, but a rule with high reach contacts more students in total, and a funder whose costs grow with the number of students contacted needs to know both.

The dropout-minus-graduate excess, Δspec, compares H2 and N1 on the difference scale, and the ratio ρ compares the coverage of dropouts with that of graduates:

$$\begin{aligned} \Delta_{\text{spec}}(S) &= \left(\text{cov}_D(S) - \overline{\text{cov}}_D(\text{B2})\right) - \left(\text{cov}_{\text{Grad}}(S) - \overline{\text{cov}}_{\text{Grad}}(\text{B2})\right), \\ \rho(S) &= \frac{\text{cov}_D(S)}{\text{cov}_{\text{Grad}}(S)}, \qquad \rho(\text{B2}) = \frac{\overline{\text{cov}}_D(\text{B2})}{\overline{\text{cov}}_{\text{Grad}}(\text{B2})} \end{aligned} \tag{12}$$

A positive excess means that the selection has a larger advantage over random groups for dropouts than for graduates, and a ratio above 1 means that it covers a larger share of dropouts than of graduates; for random groups the ratio is a ratio of averages over the draws. (With graduates as negatives, the excess is a difference of Youden indices (Youden, 1950), each the coverage of dropouts minus that of graduates; ρ is a positive likelihood ratio; and the plane of the two coverages is the receiver operating characteristic (ROC) space. Fawcett (2006) gives an introduction to ROC analysis.) The excess and ρ are descriptive and post hoc, like reach and precision.

In this article, a selection targets dropouts if, at the same reach, it covers more dropouts than random subsets of the same family; equivalently, its precision is higher than theirs, computed from their average coverage and reach. The comparison also depends on how many groups the random subsets contain (Section 6.3). H1 and H2 compare coverage, which grows with reach, and N1 compares the selection with random groups of the same budget, not of the same reach, so Section 6.3 adds a post hoc comparison at equal reach.

Intervals are percentile bootstrap intervals over students, with B = 2,000 replicates. Each replicate b draws the students with replacement, and the multiplicity me(b) of student e is the number of times the student is drawn in it; resampling is paired, so all comparators are evaluated on the same replicates. For a contrast Δ such as H1, computed from the replicate coverages, the replicate coverage, interval and two-sided counting p-value are

$$\begin{aligned} \text{cov}_D^{(b)}(S) &= \frac{\sum_{e \in D} m_e^{(b)}\ \mathbf{1}[e \in U(S)]}{\sum_{e \in D} m_e^{(b)}}, \qquad b = 1, \ldots, B, \\ \text{CI}_{95} &= \left[\, q_{0.025}\left(\Delta^{(b)}\right),\ q_{0.975}\left(\Delta^{(b)}\right) \right], \\ p &= \min\left\{1,\ \frac{2}{B}\min\left(\sum_{b=1}^{B} \mathbf{1}\left[\Delta^{(b)} \le 0\right],\ \sum_{b=1}^{B} \mathbf{1}\left[\Delta^{(b)} \ge 0\right]\right)\right\} \end{aligned} \tag{13}$$

where 1[•] is 1 when its condition holds and 0 otherwise, and qα is the α-quantile of the B replicates; replicates equal to zero count on both sides, so this p-value approximately inverts the percentile interval. Students are resampled without stratification by programme, so the mix of programmes varies between replicates; programmes themselves are not resampled, and the intervals therefore omit variation between programmes and are probably too narrow (Section 9). The coverage of the random comparator B2 is averaged over 1,000 draws that are fixed once for all replicates; the intervals condition on those draws and on the groups chosen in training. Neither the Monte Carlo error of the draws nor the variability of the training-chosen groups (Section 6.1) is included. No permutation test of H1 itself was run: its p-value comes from the bootstrap replicates, and permutations are used only for the placebo test P1. The negative control passes only if the upper limit of its interval lies at or below 2 percentage points. This one-sided control bound is in the spirit of equivalence testing (Lakens, 2017), but it is not a formal equivalence test with a margin fixed on both sides; as the upper limit of a two-sided 95% interval, it corresponds to a one-sided level of 2.5%. Intervals that are too narrow make both criteria lenient, a lower limit above zero for H1 and an upper limit at or below 2 points for N1, so this level is nominal. The two-point bound is the relevance threshold of the analysis plan: a convention fixed in advance, not a margin derived from a substantive argument.

The analysis plan was committed to a version-controlled repository and tagged, that is, given a fixed label, as plan-congelado (Spanish for frozen

plan) before the confirmatory outcomes were analysed; this is a private form of preregistration (Nosek et al., 2018), whose limits Section 9 discusses. The code refuses to read those outcomes without the tag. That guard and the coverage routines were recorded in the repository with the tag or before it. In order: the tag; the frozen run, about 1 minute later; about 1 minute after that, one commit with the analysis script, the frozen results file and Amendment 1, which states that it was written after the frozen result had been seen and reports a calibration on the training cohort whose computation the repository does not record; and the calibration script, first run about 10 minutes and committed about 27 minutes after the tag, which reproduces that calibration. A later run of the analysis script regenerated the frozen results file, reproducing its numbers except for timings and a status field. With the labels of Section 1, Amendment 1 is both post hoc and exploratory. The plan specifies the primary contrast H1 (greedy minus B1 coverage of dropouts at K = 10) and the secondary contrast H2 (greedy minus B2). It specifies the negative control N1, the coverage of graduates against B2, which passes if the upper limit of its interval is at most 2 points (Section 5.5); for H1, by contrast, the point estimate must reach 2 points. It specifies the placebo test P1, which permutes the training dropout labels 1,000 times, re-runs the greedy rule on each permutation, and passes only if the observed coverage exceeds the 95th percentile of the permuted coverages. The code states the same test as a p-value, the share of the $\Pi = 1{,}000$ permutations whose coverage is at least the observed coverage, where $S_\pi$ is the greedy selection on the $\pi$-th permutation:

$$p_{\mathrm{P1}} = \frac{1}{\Pi}\sum_{\pi=1}^{\Pi} \mathbf{1}\left[\mathrm{cov}_D(S_\pi) \geq \mathrm{cov}_D\left(S_{\mathrm{greedy}}\right)\right] \tag{14}$$

This share must not exceed 1 in 20; counting the observed labels as a permutation would change it by less than 0.001. In both designs reported here the two forms give the same result. The plan further specifies two sensitivity analyses: S1, without debtor status, tuition-fee payment status and scholarship holder; and S2, with K = 5 and K = 20. As exploratory, it lists an analysis with the roles of the two cohorts reversed (E1), which was run after the tag and is reported in Section 6.8, and the size limits of 15% and 35% and minimum sizes of 20 and 50 students, which the multiverse grid of Section 6.7 covers. Table III lists the analyses and their rules.

H1, H2, S1 and the two variants of S2 form the multiplicity family; their p-values are adjusted by Holm's step-down procedure: the five p-values are ranked from smallest to largest, the smallest is multiplied by five, the next by four, and so on, and each adjusted value is the running maximum of these products, capped at 1. For all five, no bootstrap replicate lies on the other side of zero, so the computed p-values are zero, which with 2,000 replicates means only that they are below the resolution of the count: 0.001 before adjustment and 0.005 after it. The decisions use the unadjusted 95% intervals and the validity tests, and no interval in this article is adjusted for multiplicity. In the code, the decision for H1 is positive if the difference is at least 2 percentage points and the lower limit of its interval exceeds zero; null if the lower limit is zero or below; and a positive decision is labelled not valid if a validity test fails. The validity tests are the negative control N1, the placebo test P1, and the sensitivity analyses S1 and S2, as the analysis plan lists them. The code applies only N1 and P1, and the plan gives no pass criterion for S1 and S2, so we follow the code. This does not change the decision on H1: both designs fail N1, so the result is not valid under either reading. The contrast H2 is judged with the same two-point threshold and an interval that excludes zero, which the plan states for it: by that rule the frozen H2 is positive, because its groups reach every student (Section 6.1), and Amendment 1, which is exploratory, gets no confirmatory decision on H2.

The plan leaves four cases open or words them differently, and we follow the code in each. The code makes null a contrast of at least 2 points whose interval includes zero and an interval entirely below zero, labels only positive results not valid, and gives a contrast below 2 points whose interval lies above zero a separate label, positive below the threshold, which no reported result receives. None of these cases changes the decision on H1.

***Table III. Analyses and rules fixed in the analysis plan before the confirmatory outcomes were analysed.***

| Analysis | Contrast | Rule in the plan | Role |
|---|---|---|---|
| H1 | Greedy minus B1: coverage of dropouts at K = 10 | Difference of at least 2 percentage points, and a lower limit of the bootstrap interval above zero | Primary |
| H2 | Greedy minus the mean of B2 | Difference of at least 2 percentage points, and an interval that excludes zero | Secondary |
| N1 | Coverage of graduates: greedy minus the mean of B2 | Upper limit of the interval at most 2 percentage points | Negative control (validity test) |
| P1 | Coverage under permuted training labels (placebo, 1,000 permutations) | Observed coverage above the 95th percentile of the permuted coverages | Placebo (validity test) |
| S1 | H1 without debtor status, tuition-fee payment status and scholarship holder | No pass criterion in the plan | Sensitivity |
| S2 | H1 with K = 5 and K = 20 | No pass criterion in the plan | Sensitivity |
| H3 | Greedy value divided by the training optimum, for K = 5, 10 and 20 | At least 1 − 1/e | Numerical check of the guarantee |
| H4 | B3 minus greedy coverage of dropouts | No pass criterion in the plan | Descriptive |
| E1 | Reversed roles of the two cohorts | No pass criterion in the plan | Exploratory; run after the tag (Section 6.8) |
| Size limits and minimum sizes | Size limits of 15% and 35%; minimum group sizes of 20 and 50 students | No pass criterion in the plan | Exploratory; covered by the multiverse grid (Section 6.7) |

***Only N1 and P1 are applied as validity tests.***

Amendment 1 changes only the size limit, from 25% to 10%. It was decided after the frozen result had been seen, and was committed before its analysis was run, so its analysis is exploratory and not confirmatory. The 10% value was taken from a calibration on the training cohort against a criterion, set in the same amendment, that training coverage at K = 10 stays below 80%, so that the groups do not already cover almost every training dropout (saturation), as the frozen groups did; the value 80% is not justified (Section 6.2).

Table IV summarises the three main analyses: the frozen primary analysis, Amendment 1 and the reversed direction (E1).

With the 25% limit, the greedy rule stopped after eight groups, below the budget of 10, because every training dropout was already covered, so no further group added coverage on the training cohort. The selection covered 100.0% of confirmed dropouts. B1 covered 31.0%, and random groups covered 65.8% on average. The difference from B1 was 69.0 points, with a 95% bootstrap interval from 65.6 to 72.4.

## 6. RESULTS

***Table IV. The three analyses at K = 10: the frozen primary analysis (25% limit) and Amendment 1 (10% limit, post hoc), evaluated on the confirmatory cohort, and the reversed direction (E1), with groups chosen on the confirmatory cohort and evaluated on the training cohort.***

| Quantity | Frozen primary (25%) | Amendment 1 (10%) | Reversed direction (E1) |
|---|---|---|---|
| Candidate groups | 143 | 86 | 139 |

| | | | |
|---|---|---|---|
| Groups selected by the greedy rule | 8 | 10 | 10 |
| Groups in B1 (top rate) | 10 | 10 | 10 |
| Groups in B3 (training optimum) | 6 | 10 | — |
| Reach of the greedy rule | 100.0% | 55.4% | 94.7% |

| | | | |
|---|---|---|---|
| Dropout share among those reached | 29.0% | 37.1% | 37.5% |
| Base rate of dropout | 29.0% | 29.0% | 36.3% |
| Dropouts in the evaluation cohort | 729 | 729 | 692 |
| Coverage of dropouts, greedy | 100.0% | 70.9% | 97.8% |
| Coverage of dropouts, B1 | 31.0% | 33.6% | 33.4% |
| Coverage of dropouts, B2 | 65.8% | 45.9% | 72.6% |
| Coverage of dropouts, B3 | 100.0% | 70.9% | — |
| H1: greedy minus B1 (pp) | 69.0 | 37.3 | 64.5 |
| H1: 95% interval (pp) | [65.6, 72.4] | [33.9, 41.0] | [60.9, 67.8] |
| H2: greedy minus B2 (pp) | 34.2 | 25.0 | 25.2 |
| H2: 95% interval (pp) | [33.0, 35.4] | [22.5, 27.7] | [23.8, 26.6] |
| N1: graduates covered, greedy | 100.0% | 45.3% | 92.9% |
| N1: graduates covered, B2 | 59.0% | 29.3% | 59.4% |
| N1: greedy minus B2 (pp) | 41.0 | 16.0 | 33.5 |
| N1: 95% interval (pp) | [40.1, 41.9] | [13.8, 18.3] | [31.6, 35.4] |
| P1: placebo p-value | 0.725 | 0.011 | not run |
| Decision for H1 (plan's decision rule) | not valid | not valid (post hoc) | not applied |

Coverage is the share of the evaluation cohort's dropouts (for N1, of its graduates) covered by the selected groups; B2 is the mean over 1,000 draws of 10 groups. Differences are in percentage points (pp), computed from unrounded values; intervals are 95% percentile bootstrap intervals over students. A dash marks a quantity not computed for E1.

The result is not informative about targeting. The eight selected groups together contain every student of the confirmatory cohort, because they cover every combination of gender, displacement, age band and attendance that occurs in the data; the training cohort, available before the freeze, already showed this (there the greedy rule reaches every training student after eight groups; Figure 1(b)), but this was not checked before the tag. A simpler check would also have flagged the risk: 10 groups of up to 476 students can hold 4,760 students, more than the 1,906 training students, whereas 10 groups of up to 190 hold at most 1,900. The selection reaches 100.0% of confirmed students, its precision is 29.0%, which is the base rate of dropout, and its lift is 1.00. Covering dropouts therefore means covering graduates too:

the selection covered 100.0% of confirmed graduates, against 59.0% for random groups, so the negative control fails. Once the groups contain every student, coverage of dropouts and of graduates is fixed by construction, whatever the labels are. Under the plan's decision rule, the primary analysis is therefore not valid. The selection's positive contrast with random groups (H2) has the same cause: the selection reaches all students, whereas random groups reach only 61.0% of them on average.

In the placebo test (Section 5.6), the mean coverage of confirmed dropouts across permutations was 99.5%, the 95th percentile was 100.0% and the p-value was 0.725. Under the frozen design, the placebo test therefore does not separate the greedy selection from selections on permuted labels.

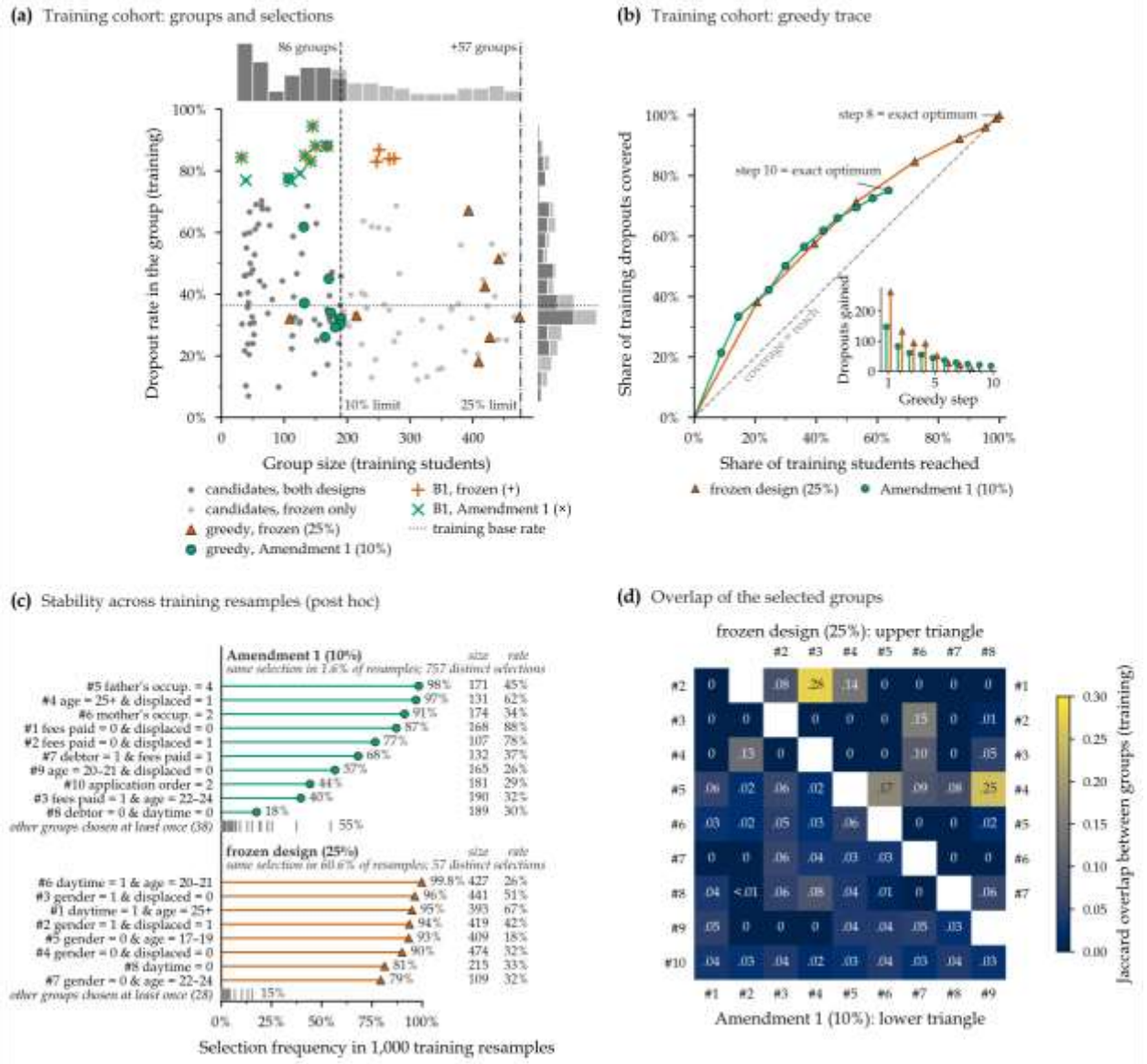



***Figure 1. Candidate groups and the greedy selection on the training cohort, without confirmatory outcomes (post hoc, descriptive). Green circles mark Amendment 1 (10% limit) and vermilion triangles the frozen design (25% limit). (a) Size and training dropout rate of the 143 candidate groups of the frozen family; the 86 at or below the 10% limit of 190 students form the family of Amendment 1. Filled markers are groups chosen by the greedy rule and crosses groups chosen by B1 (× for Amendment 1, + for the frozen design), so a group chosen by both carries both; dashed and dash-dotted lines mark the 10% and 25% limits, the dotted line the training base rate, and the margins give size and rate by family. (b) Share of training dropouts covered against share of training students reached as the greedy rule adds groups; the frozen trace reaches every training student at step 8, labels mark the training optimum at K = 10, and the inset gives the dropouts gained at each step, which never increase (Lemma 1 and the greedy choice). (c) Share of 1,000 bootstrap resamples of the training cohort, with the family fixed (Section 6.1), in which each selected group is chosen***

***again; groups are numbered by greedy step and labelled with dataset codes, with their training size and dropout rate, and grey ticks are other groups chosen at least once. (d) Jaccard overlap between the selected groups, numbered by greedy step: the amended selection in the lower triangle and the frozen one in the upper triangle, on one colour scale.***

The tie-breaking order changes the frozen selection and, slightly, its coverage: with the reverse order of group names, the greedy rule covered 99.9% of confirmed dropouts and 99.4% of graduates, instead of 100.0% of each. No two candidate groups of either design have exactly the same members, so these ties are between different groups with equal gains. The tie-breaking order is therefore part of the specification, and an implementation must follow the one stated in Section 4.1.

Resampling the training cohort is a further check on the selection (post hoc). The family of groups is kept fixed at the one built once from the original training cohort, so the frequencies below are conditional on that family. In 1,000 bootstrap resamples of the training students, the greedy rule with K = 10 reproduces the frozen selection exactly in 60.6% of them, and each of its groups appears in at least 79.4% of them (on average 91.1%). The amended selection is reproduced exactly in only 1.6% of the resamples, which yield 757 distinct selections, and each of its groups appears in at least 17.5% of them. The amended selection is therefore one of many selections that overlap only in part: on average, the selection from a resample contains 67.6% of the amended groups, and their names should not be read as a stable profile of students.

In the amended design (Figure 1), the groups chosen by the greedy rule have a median size of 169.5 students and a median training dropout rate of 35.5%, close to the training base rate of 36.3%; the groups of B1 have a median size of 129 students and a median rate of 83.7%. The groups that the greedy rule adds at steps 1 and 2, both also in B1, are together exactly the 275 training students with tuition-fee payment status 0, that is, with fees in arrears (84.0% dropped out). Too many for one group under the 10% limit of 190, these students enter as two pairs split by displacement. Tuition-fee payment status is one of the three uncertain-timing attributes. Any two groups of the amended selection share at most 12.8% of the students in their union (the Jaccard index), and any two groups of the frozen selection at most 27.5%.

The 10% limit and the saturation criterion were set in Amendment 1 (Section 5.6). The calibration in Figure 2 reports the training coverage for three limits, 10%, 15% and 25%, and for budgets from 1 to 20. With the 10% limit, the training coverage at K = 10 was 75.1%, against 100.0% with the 25% limit. The saturation criterion, that training coverage at the planned budget K = 10 stays below 80%, is an arbitrary choice that the amendment does not justify. With a threshold of 75.1% or lower, none of the three limits would pass, and with any threshold between that value and the coverage under the 15% limit, 97.3%, only the 10% limit passes. From K = 13 on (81.1%), training coverage under the 10% limit exceeds 80%, so the criterion depends on the budget at which it is applied.

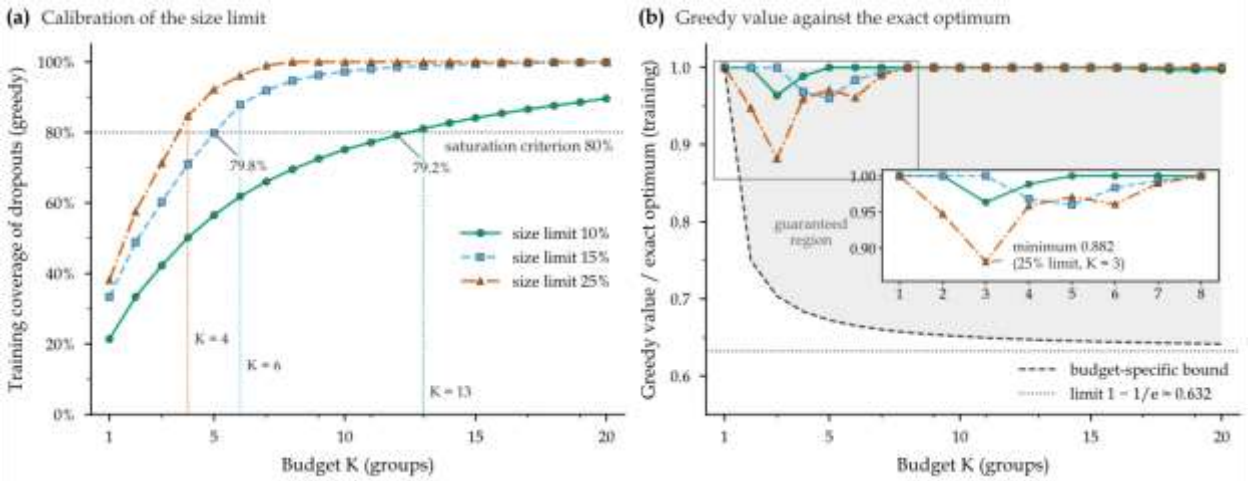



***Figure 2. Calibration and guarantee on the training cohort, by budget K and for three group-size limits; each limit has its own colour, marker and line style (10%: green circles, solid; 15%: sky-blue squares, dashed; 25%: vermilion triangles, dash-dotted). (a) Training coverage of dropouts by the greedy rule; the dotted line marks the saturation criterion of 80% adopted in Amendment 1, the vertical lines the first budget at which each limit reaches it, and the labels the two coverages that fall just below it. (b) Ratio between the greedy value and the exact optimum. The black dashed curve is the budget-specific bound of Equation 7, the grey***

***band the region it guarantees, and the black dotted line its limit 1 − 1/e ≈ 0.632; the inset enlarges the budgets from 1 to 8, marked by the rectangle, and labels the smallest ratio.***

Amendment 1 is exploratory. It has 86 candidate groups, and the greedy rule covered 70.9% of confirmed dropouts, against 33.6% for B1 and 45.9% for random groups on average. The difference from B1 was 37.3 points, with an interval from 33.9 to 41.0. The difference from random groups (H2) was 25.0 points.

Table V separates coverage from reach and precision. The greedy rule reached 55.4% of confirmed students, and the dropout share among them was 37.1%, against a base rate of 29.0%: a lift of 1.28. Excluding the 386 students still enrolled, the precision is 44.8% against a base rate of 34.2% (lift 1.31), and the coverage of dropouts and of graduates does not change. B1 has the highest precision, with a dropout share of 85.1% among the students it reaches (11.4% of all students), but it covers only 33.6% of dropouts. It also covers only 1.9% of graduates, against 29.3% for random groups, so its dropout-minus-graduate excess is 15.2 points and its ratio ρ is 17.5, against 9.1 points and 1.56 for the greedy rule. B1 therefore discriminates better: its ratio is inflated by its small reach, but its excess is also larger, although the excess favours broad rules. It loses H1 because it reaches few students: the primary contrast of the plan favours broad rules, whereas a negative control against random groups of the same budget favours narrow ones. Random groups reach fewer students than the greedy rule (34.8%), with a higher dropout share (38.2%; the mean of the per-draw shares is 39.3%). The greedy rule is thus a broad rule with modest precision, and B1 is a narrow rule with high precision. The contrast with B1 therefore reflects both reach and precision; no reach-matched B1 comparison was run. The greedy rule with equal weights, which ignores labels, reaches even more students (68.0%) and covers dropouts and graduates almost equally (68.6% and 67.4%). Figure 3 shows the five rules of Table V, including the K largest groups, at three budgets under both designs.

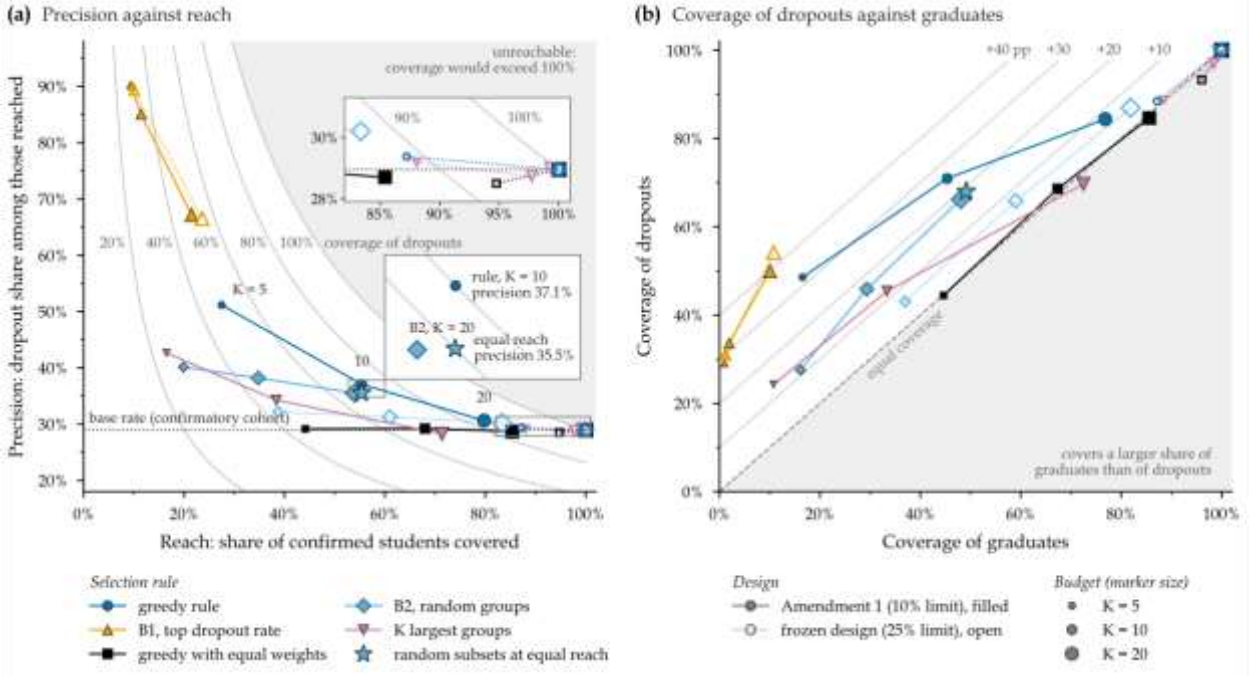



***Figure 3. Reach, precision and coverage of the five rules of Table V on the confirmatory cohort, for budgets K = 5, 10 and 20 (marker size), under Amendment 1 (filled markers, solid lines) and the frozen design (open markers, dotted lines; there the greedy rule and the rule with equal weights coincide at 100% reach for K = 10 and 20); the star marks the random subsets of the Amendment 1 family at equal reach (Section 6.3). (a) Precision, the dropout share among the students reached, against reach, the share of students reached; the black horizontal dotted line marks the base rate, and the grey region above the curve of coverage 100% cannot be reached. The grey curves join the points with equal coverage of dropouts, labelled by that coverage, because, for a single selection, coverage equals precision times reach divided by the base rate (Equation 11); the precision of B2 and of the star is computed from their average coverage and reach through the same identity. Two insets enlarge the marked rectangles, where points would otherwise hide one another. (b) Coverage of dropouts against coverage of graduates, that is, sensitivity against false-positive rate with graduates as negatives (ROC space); points above the dashed diagonal cover a larger share of dropouts than***

***of graduates. The thin grey lines parallel to the diagonal mark constant differences between the two coverages (constant Youden index), labelled at the top.***

The negative control fails at the planned budget. The selection covered 45.3% of graduates, against 29.3% for random groups: a difference of 16.0 points, with an interval from 13.8 to 18.3. Its upper limit exceeds the two-point bound, so under the plan's decision rule H1 is not valid. On the difference scale, however, the selection's advantage over random groups is larger for dropouts than for graduates at K = 5 (20.6 points; interval 17.1 to 24.1) and K = 10 (9.1 points; interval 5.7 to 12.3); at K = 20 the sign reverses (−10.6 points; interval −13.4 to −7.8). Under the frozen design at K = 10, the dropout-minus-graduate excess is −6.8 points, with an interval from −8.3 to −5.5. On the ratio scale, the share of dropouts that the selection covers is 1.56 times the share of graduates it covers, and random groups do the same (1.56 times, as a ratio of averages over the draws). On both scales the comparison is confounded by reach, which at K = 10 is 55.4% for the selection and 34.8% for random groups. A selection's coverage of dropouts minus its coverage of graduates equals its coverage of graduates times (ρ − 1), so at the same ratio above 1 a broader selection has a larger excess, whereas the ratio of random groups falls as their reach grows (1.71, 1.56 and 1.38 at K = 5, 10 and 20). The comparison at equal reach below removes this confounding.

The negative control compares the rule with random groups that use the same budget, so the effects of reach and labels are confounded. As a post hoc check, we compared the rule with random subsets of the Amendment 1 family, drawn without replacement with a size uniform between 1 and 40 groups. Of the 40,000 subsets, the 630 whose reach lies within half a point of the rule's 55.4% are kept; they contain 22.3 groups on average, because random subsets of 10 groups reach fewer students than the rule. With $\tilde{S}(1)$, …, $\tilde{S}(J)$ the J = 40,000 random subsets and h the half-width of the window, the indices kept, Qh, and the comparator's coverage of dropouts are

$$Q_h = \{j : |\mathrm{reach}(\tilde{S}^{(j)}) - \mathrm{reach}(S_{\mathrm{greedy}})| \le h\}, \qquad \overline{\mathrm{cov}}_{D,h} = \frac{1}{|Q_h|} \sum_{j \in Q_h} \mathrm{cov}_D\left(\tilde{S}^{(j)}\right) \tag{15}$$

and its coverage of graduates is the same average for graduates. With the window of half a point, the kept subsets cover 68.0% of dropouts and 49.1% of graduates. The rule, with 10 groups, covers 70.9% and 45.3%. The differences are 2.9 points for dropouts (95% bootstrap interval from 0.6 to 5.4) and −3.8 points for graduates (from −6.0 to −1.5). At equal reach, the amended selection therefore covers more dropouts and fewer graduates, and it leads on both scales: its excess over the kept subsets is positive (Table VI), and its ratio is 1.56 against 1.38 for the kept subsets, a ratio of averages (1.22 to 1.38 across the comparators of Table VI). The intervals of this comparison omit variation between selections, and the selection is unstable: it is reproduced exactly in only 1.6% of the training resamples (Section 6.1), so the intervals may be too narrow. The comparison was not repeated without the three uncertain-timing attributes.

***Table V. Exploratory selections under Amendment 1 (10% limit, K = 10) in the confirmatory cohort.***

| Rule | Students reached | Dropout share among those reached | Lift over the base rate | Coverage of dropouts | Coverage of graduates |
|---|---|---|---|---|---|
| Greedy rule | 55.4% | 37.1% | 1.28 | 70.9% | 45.3% |
| B1, top dropout rate | 11.4% | 85.1% | 2.94 | 33.6% | 1.9% |
| B2, random groups | 34.8% | 38.2% | 1.32 | 45.9% | 29.3% |
| Greedy with equal weights | 68.0% | 29.2% | 1.01 | 68.6% | 67.4% |
| K largest groups | 38.4% | 34.3% | 1.18 | 45.4% | 33.4% |

Students reached is the share of all confirmed students covered by the selected groups; the dropout share among them is the precision, and lift divides it by the base rate of 29.0%. For B2, reach and coverage are means over 1,000 draws, and the dropout share and lift are computed from them through Equation 11, as for the subsets at equal reach; the mean of the per-draw shares is 39.3%.

Table VI repeats the comparison with a window of 1 point and with subsets of up to 20 and up to 60 groups. The window and the larger range change the differences little. With subsets of up to 20 groups, the maximum size is reached in 14.7% of the kept subsets, and the differences grow to 7.3 points for dropouts and −6.7 for graduates. The dropout-minus-graduate excess, defined as in Equation 12 with the kept subsets in place of B2, is positive in every row, with an interval above zero. At equal reach, the random subsets of the comparator, with

precision computed from their average coverage and reach, have a dropout share of 35.5%, against 37.1% for the rule (descriptive). The direction holds in every comparator tried, but the size of the difference depends on how the comparator is built, and the comparator's group count is not matched, so the result does not settle the question of targeting.

With permuted training labels, the greedy rule of Amendment 1 covered 65.1% of confirmed dropouts on average, and the 95th percentile of the permuted coverages was 69.3%, below the observed 70.9%, so the placebo test passes (p-value 0.011). The permuted-label selections also cover more dropouts than random groups: 19.3 of the 25.0 points by which the rule exceeds random groups (H2) are reproduced with permuted labels. Most of the advantage over random groups therefore does not depend on the observed labels, and at equal reach only 2.9 of these 25.0 points remain, and at most 7.4 with the other comparators of Table VI.

Figure 4 summarises the comparisons at K = 10.

**Figure 4. Differences in coverage, in percentage points computed from unrounded values, with 95% bootstrap intervals (K = 10), in three blocks: the frozen primary analysis (the confirmatory test), the exploratory analysis under Amendment 1, and the post hoc comparison of Amendment 1 with random subsets of its family at equal reach (Section 6.3). Under the titles of the first two blocks are the reach of the rule and the p-value of the placebo test P1. Rows give H1 (rule minus B1) and H2 (rule minus B2) for dropouts (diamonds), the negative control N1 (rule minus B2) for graduates (squares), and the dropout-minus-graduate excess (circles), which in the first two blocks is that of Equation 12 and in the third compares the rule with the subsets at equal reach; the right column lists each estimate with its interval, the outcome of the negative control and, in italics, the decision on H1 under the plan's decision rule, which for Amendment 1 is what that rule would give, not a confirmatory decision. The dashed line, drawn through the first two blocks, marks the two-point threshold: the estimates of H1 and H2 must reach it, and the upper limit of N1 must lie at or below it for the control to pass; it does not apply to the excess or to the comparison at equal reach.**

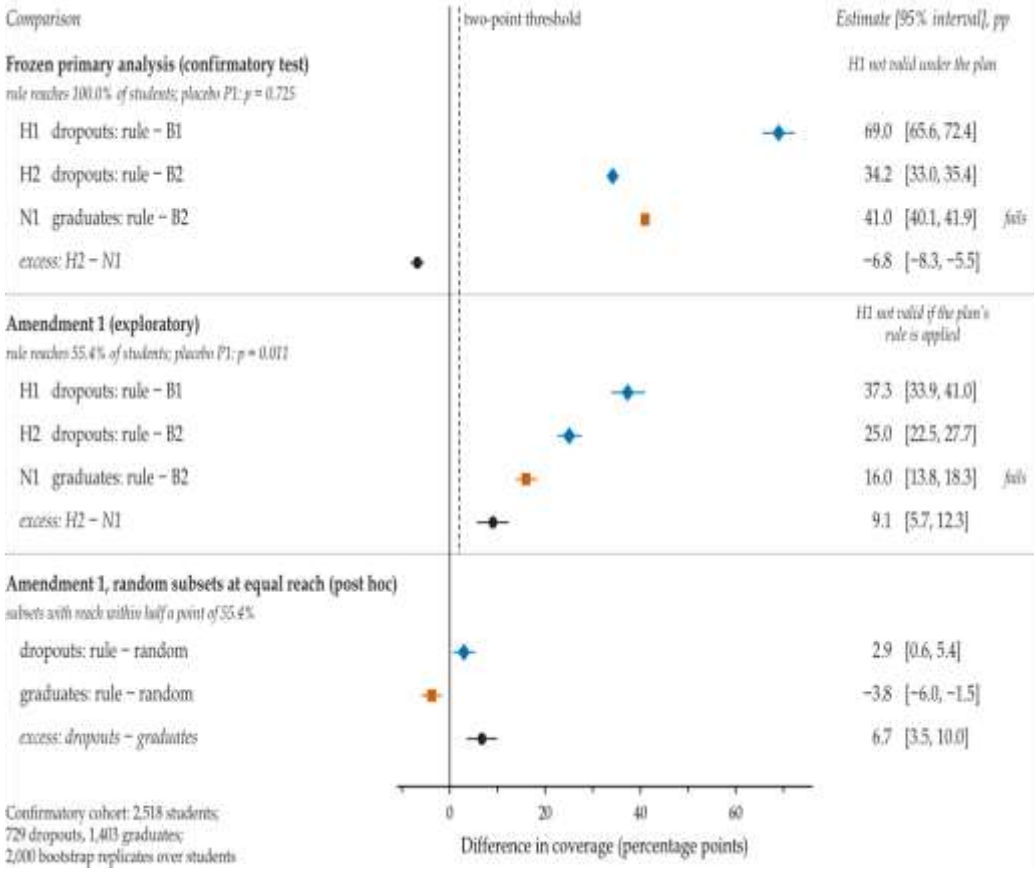



Table VII reports the sensitivity analyses for both designs. The difference from B1 stays positive at the three budgets analysed, K = 5, 10 and 20, under both designs. Under Amendment 1 it was 19.5 points with K = 5; with K = 20 it was 34.4, and excluding the three uncertain-timing attributes it was 33.6. The negative control of Amendment 1 fails at every budget tested, narrowly at K = 5 and clearly at K = 10 and K = 20. At

K = 5 its point estimate is 0.4 points and its upper limit 2.3, just above the two-point bound; whether it fails depends on the seed, which sets both the random groups and the bootstrap replicates (Section 6.7). When the three uncertain-timing attributes are excluded (S1, K = 10), the selection of Amendment 1, rebuilt without them, has a precision of 32.5%; the dropout-minus-graduate excess falls to 0.9 points (interval from −2.7 to 4.7), although this scale favours a selection that reaches more students than random groups, as this one does (57.0% against 35.1%). The signs of targeting in Amendment 1 therefore depend on these attributes.

***Table VI. Rule minus random subsets of the Amendment 1 family at equal reach (post hoc), for three ranges of subset size and two reach windows. Differences in percentage points, with 95% bootstrap intervals.***

| Groups per subset | Window | Subsets kept | Mean groups | Dropouts | Graduates | Dropout-minus-graduate excess |
|---|---|---|---|---|---|---|
| 1 to 40 | ±0.5 pp | 630 | 22.3 | 2.9 [0.6, 5.4] | −3.8 [−6.0, −1.5] | 6.7 [3.5, 10.0] |
| 1 to 40 | ±1 pp | 1,299 | 22.4 | 3.1 [0.8, 5.5] | −3.9 [−6.1, −1.6] | 7.0 [3.7, 10.2] |
| 1 to 20 | ±0.5 pp | 510 | 16.7 | 7.3 [5.0, 9.9] | −6.7 [−9.0, −4.3] | 14.1 [10.7, 17.4] |
| 1 to 20 | ±1 pp | 1,041 | 16.8 | 7.4 [5.1, 9.9] | −6.7 [−9.0, −4.3] | 14.0 [10.7, 17.3] |
| 1 to 60 | ±0.5 pp | 428 | 22.3 | 3.0 [0.8, 5.5] | −3.9 [−6.1, −1.6] | 6.9 [3.6, 10.2] |
| 1 to 60 | ±1 pp | 879 | 22.6 | 3.0 [0.8, 5.5] | −3.8 [−6.1, −1.6] | 6.9 [3.6, 10.2] |

Subsets are drawn without replacement from the Amendment 1 family, with a number of groups uniform over the range shown, and are kept when their reach lies within the window (h in Equation 15) around the rule's reach of 55.4%; their coverages are averages over the kept subsets. The first row is the main comparison of Section 6.3. The excess is the dropout difference minus the graduate difference; differences are computed from unrounded values. The number of groups is not matched: the rule uses 10. The kept subsets are fixed from the full-sample reach and reused in every bootstrap replicate, so the intervals omit the variation of the comparator's draws and of the selection; the rows are not adjusted for multiplicity.

To check transfer beyond one split, and whether the pooled result depends on one programme, we re-selected the amended rule's groups on seven of the eight confirmatory programmes, using their dropout labels, and evaluated the selected groups on the eighth, for each programme in turn (the eight-programme check). The greedy rule covered more dropouts than random groups in seven of the eight programmes. The unweighted mean of the eight programme-level differences was 21.0 points, and the individual differences ranged from −12.0 to 37.5. This check is post hoc and not an independent test, because it reuses the confirmatory cohort's labels; the family of groups comes from the training cohort, and the budget from the plan. The programme with the lowest difference, code 9991 (Management, evening attendance), has the most dropouts of the eight: 136 of its 268 students. There the rule covers fewer dropouts than random groups, and it covers 26.9% of the programme's graduates, against 67.5% for random groups. Within programmes, the rule's advantage over random groups among graduates exceeds the two-point bound in seven of the eight programmes, and the dropout-minus-graduate excess is zero or negative in three of them, including the largest, programme 9500 (−0.4 points). Figure 5 shows the programmes and both checks.

***Table VII. Sensitivity analyses, frozen design (25% limit) and exploratory Amendment 1 (10% limit). Differences in percentage points.***

| Analysis | Frozen: H1 | Frozen: N1 | Amendment 1: H1 | Amendment 1: N1 |
|---|---|---|---|---|
| S2: K = 5 | 57.8 | 50.2 | 19.5 | 0.4 |
| 95% interval | [53.8, 61.4] | [48.1, 52.2] | [16.7, 22.4] | [−1.4, 2.3] |
| K = 10 (plan) | 69.0 | 41.0 | 37.3 | 16.0 |
| 95% interval | [65.6, 72.4] | [40.1, 41.9] | [33.9, 41.0] | [13.8, 18.3] |
| S2: K = 20 | 45.8 | 18.1 | 34.4 | 28.8 |
| 95% interval | [42.2, 49.4] | [17.4, 18.8] | [30.7, 38.0] | [26.9, 30.6] |
| S1: three attributes excluded (K = 10) | 54.6 | 38.0 | 33.6 | 20.5 |
| 95% interval | [51.2, 58.3] | [37.5, 38.5] | [28.9, 38.2] | [18.3, 22.7] |

H1 is greedy minus B1 coverage of dropouts; N1 is greedy minus random-group coverage of

graduates. Intervals are 95% bootstrap intervals over students. The plan defines S1 and S2 for H1 only, so N1 at K = 5, at K = 20 and without the three uncertain-timing attributes is post hoc.

In the 17-programme check (post hoc), we left out one of the 17 programmes at a time and re-selected the groups with the other 16. The labels of the other 16 programmes are used for selection in each fold, and the family of groups is rebuilt from those programmes (105 to 112 groups); the size limit is 10% of all students in the 16 programmes, between 365 and 441 students, instead of 190 (10% of the training cohort) in the eight-programme check. Like the eight-programme check, it reuses labels that earlier analyses had already used, so it is not an independent test. The rule covers more dropouts than random groups of the same budget in all 17 programmes, by between 3.1 and 39.9 points (mean 26.0); the smallest advantage is in programme 8014, and the largest in programme 9147. For graduates, the rule covers more than random groups in all but three programmes: programme 33 (Biofuel Production Technologies), whose 12 students include a single graduate, so that its graduate difference (−7.5 points) rests on one student, and programmes 8014 and 9991. This pattern is consistent with a reach effect, which this check does not measure per programme. The advantage is larger for dropouts than for graduates in 14 of the 17 programmes. Programme 9991, which was negative in the eight-programme check above, is positive here (9.1 points for dropouts). The two checks differ in the programmes used for selection, in the family of groups and in the size limit, so the change of sign cannot be attributed to the selection programmes alone.

Figure 2(b) shows the ratio between the greedy value and the exact optimum on the training cohort, for each budget from 1 to 20 and for the three size limits; every exact solve of this grid returned status optimal. The lowest ratio was 0.964 for the 10% limit, 0.960 for the 15% limit and 0.882 for the 25% limit. Every ratio is above the guarantee of $1 - 1/e \approx 0.632$ in Theorem 2. Every ratio is also at or above the budget-specific bound of Equation 7, with equality only at K = 1, where the greedy rule is optimal. At K = 10 the greedy value equalled the optimum for all three limits. The check is numerical and concerns these instances; the guarantee itself is the theorem.

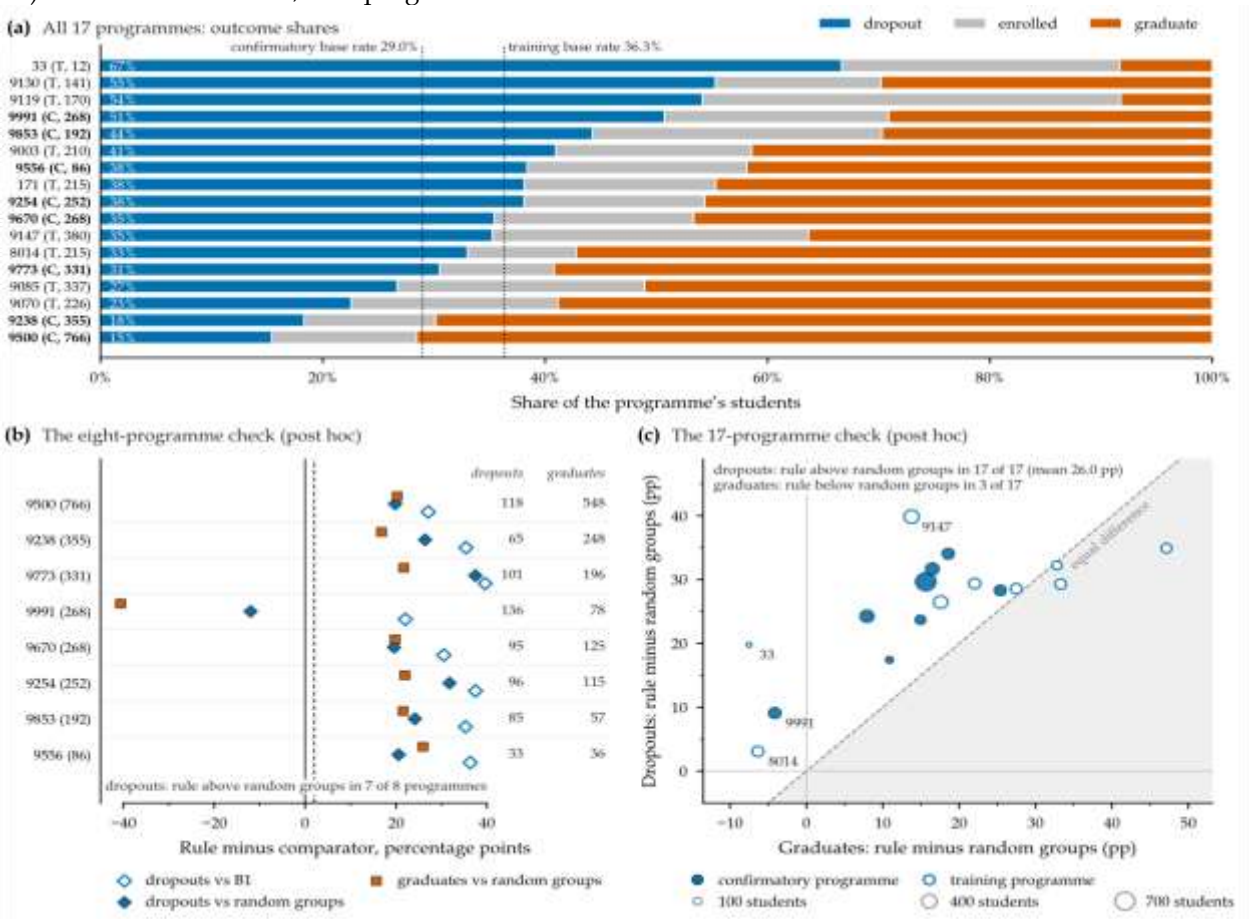



***Figure 5. Outcomes by programme and the two checks across programmes. (a) Outcome composition of the 17 programmes, ordered by dropout share, with each programme's cohort (T, training; C, confirmatory, in***

***bold) and number of students; vertical lines mark the base rates of the two cohorts. (b) The eight-programme check (post hoc; Section 6.4), with each programme's number of students in parentheses and its numbers of dropouts and graduates at the right: the rule and B1 are re-selected on the other seven programmes and evaluated on the one shown; markers give the rule's coverage minus that of B1 for dropouts (open diamonds) and minus that of random groups for dropouts (filled diamonds) and graduates (squares), in percentage points, as in Figure 4; the dashed line marks the two-point bound of the negative control. (c) The 17-programme check, leaving one programme out (post hoc): the rule's coverage minus that of random groups, for dropouts against graduates, one marker per held-out programme, sized by its number of students; points above the dashed diagonal have a larger advantage for dropouts than for graduates. Both checks reuse confirmatory labels, so they are not independent tests, and they report no intervals.***

Table VIII reports the planned checks H3 and H4 for the three budgets of both designs. The exact optimum on the training cohort, evaluated on the confirmatory cohort, covered more confirmed dropouts than the greedy rule in two of the six cases, at K = 5 under the frozen design and at K = 20 under Amendment 1; in the other four, the two selections cover the same share, although they may differ in groups. These are out-of-sample differences: the exact solution is optimal only on the training cohort. When several selections are optimal, the solver returns one of them: the optimal training value does not depend on that choice, but the listed groups, and hence the confirmatory coverage of B3 (H4), may. Under Amendment 1 at K = 5 and K = 10, for example, the greedy selection is itself optimal on the training cohort, but it is not the selection that the solver returns.

The sensitivity analyses above change one choice at a time. To check how several choices act together, we ran a multiverse analysis (Steegen et al., 2016), shown as a specification curve (Simonsohn et al., 2020) in Figure 6, over a grid of 288 runs: 144 specifications, each specification run with two tie-breaking orders; in 90 of those specifications the tie-breaking order changes the set of selected groups. The grid crosses the group-size limit (10%, 15%, 25% and 35%), the budget (K = 5, 10 and 20), the minimum group size (20, 30 and 50 students), the family of groups (one-value and pair groups, or one-value groups only), the set of attributes (all, or without the three uncertain-timing attributes), and the tie-breaking order of the greedy rule. The plan names the 15% and 35% limits and the minimum sizes of 20 and 50 students as exploratory and the reduced attribute set as S1 (Section 5.6); the 10% limit comes from Amendment 1, and the family of one-value groups only and the reverse tie-breaking order were added here. The grid was committed to the repository before it was run, after the frozen and amended results were known, so it is post hoc. The five-seed and one-choice-at-a-time analyses below were added after the first results, without changing the grid or its decision rule. Age bands, preprocessing and comparators are not varied, so the grid does not cover every defensible choice. Every specification uses the grid's seed for the draws of B2 and the bootstrap, and the sensitivity analyses use other seeds, so a specification can differ from its entry in Table VII, for example by about 1 point in the negative control for the frozen design with K = 5.

The contrast with B1 is positive in every post hoc run, with a student-level interval above zero in all 288 of them, and it reflects reach as well as precision (Section 6.3). The difference ranges from 11.2 to 80.9 points, with a median of 38.2. The negative control behaves differently. Its upper limit is at most 2 points in two of the 288 runs, and its median upper limit is 23.9 points. When the plan's decision rule is applied to each specification, with the negative control as its only test (the placebo test was not repeated), the control passes in only one of the 144 specifications. The two tie-breaking orders select the same groups in the passing one: a size limit of 10%, K = 5 and a minimum size of 50 students. In that specification H1 is 17.6 points, with an interval from 14.9 to 20.4. Its control estimate is −2.6 points, with an interval from −4.4 to −0.8, which lies entirely below zero: the selection covers fewer graduates than random groups. Its five groups are those of Amendment 1 at K = 5, and three of them use tuition-fee payment status; only the comparators B1 and B2 change, because the minimum size of 50 students leaves 64 candidate groups instead of 86. The control therefore passes because the comparator changes, not because the selection does.

***Table VIII. Planned checks H3 and H4 by budget in both designs: H3 compares the greedy rule with the exact optimum on the training cohort, and H4 compares their confirmatory coverage.***

| Design | K | Training dropouts covered, greedy | Training dropouts covered, optimum | Ratio (H3) | Confirmatory coverage, greedy | Confirmatory coverage, B3 |
|---|---|---|---|---|---|---|
| Frozen (25%) | 5 | 638 | 657 | 0.971 | 88.5% | 95.6% |
| Frozen (25%) | 10 | 692 | 692 | 1.000 | 100.0% | 100.0% |
| Frozen (25%) | 20 | 692 | 692 | 1.000 | 100.0% | 100.0% |
| Amendment 1 (10%) | 5 | 391 | 391 | 1.000 | 48.6% | 48.6% |
| Amendment 1 (10%) | 10 | 520 | 520 | 1.000 | 70.9% | 70.9% |
| Amendment 1 (10%) | 20 | 620 | 622 | 0.997 | 84.4% | 88.2% |

The training cohort has 692 dropouts and the confirmatory cohort 729. Every exact solve returned status optimal. The plan gives no pass criterion for H4. At K = 10 the frozen greedy rule stops after eight groups, and the frozen optimum uses six.

When the grid is repeated with five seeds (the grid's own and four more), the specification with the 10% limit, K = 5 and a minimum size of 50 students passes under all five seeds, and the set of passing specifications is the same under four of them. In the remaining seed, the neighbour with a minimum size of 30 students, which is Amendment 1 at K = 5, also passes. Across the five seeds, the upper limit of the control for the passing specification lies between −1.3 and −0.7 points, and for that neighbour between 1.98 and 2.44 points. Whether that neighbour passes therefore depends on the seed.

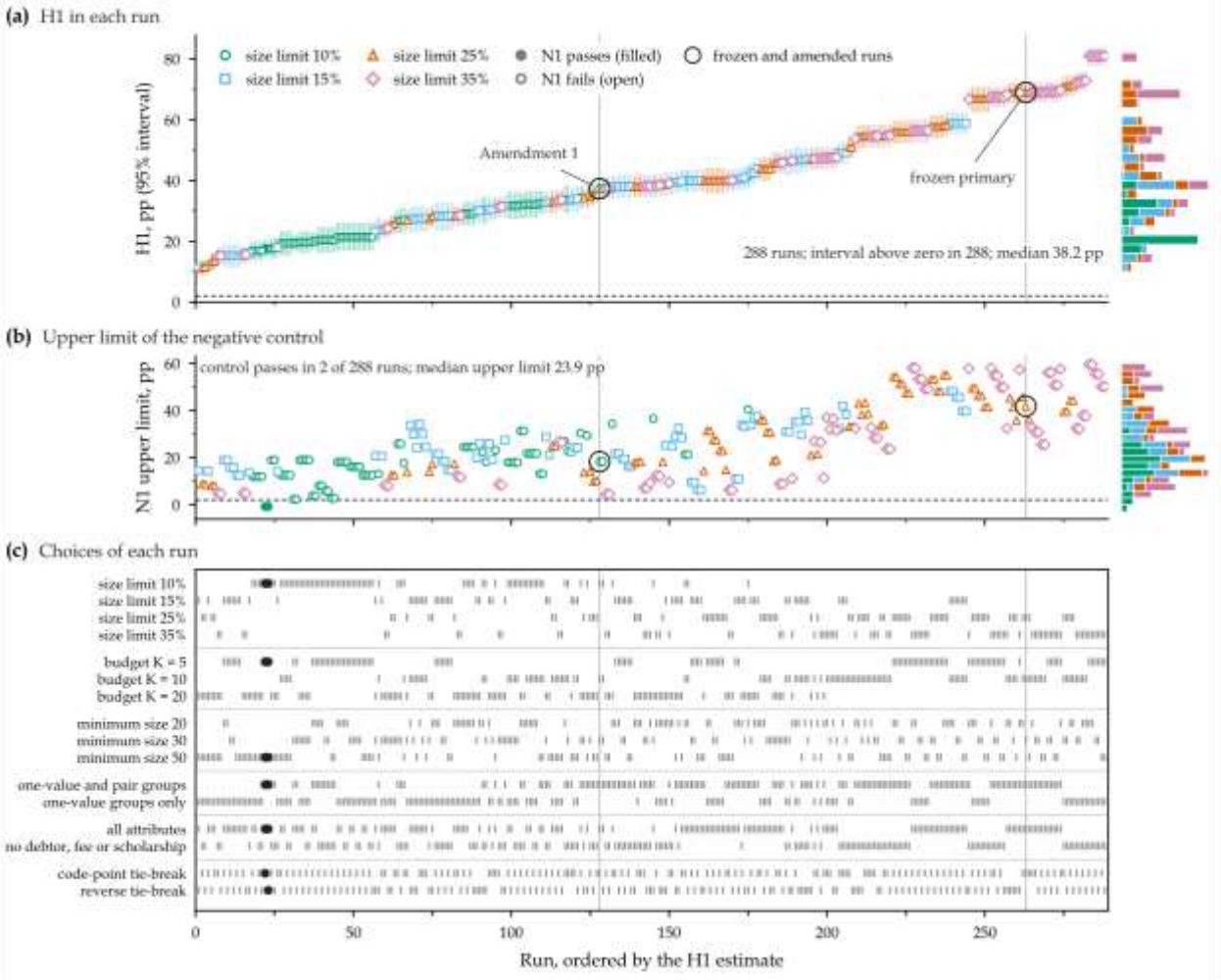



*Figure 6. Specification curve of the post hoc multiverse, with one column per run (two runs per*

***specification), ordered by the H1 estimate. Colours and markers give the size limit, as in Figure 2, with pink diamonds for the 35% limit; filled markers denote runs in which the negative control passes and open markers those in which it fails. (a) Difference in coverage of dropouts between the greedy rule and B1, in percentage points, with its 95% bootstrap interval; black rings mark the run of the frozen primary specification and the run of Amendment 1 with the plan's code-point tie-break, and the dashed line the two-point threshold of the H1 decision rule. (b) Upper limit of the negative-control interval, against the two-point bound; the two runs of the passing specification have the same selection and identical values, and sit side by side. The histograms at the right of (a) and (b) give the distribution of the runs, stacked by size limit, and the text in each panel summarises the runs. (c) Choices of each run: one row per level of the six choices of the grid, with a tick in the column of each run that uses that level, and a filled circle for runs in which the control passes. The grey vertical lines mark the frozen and the amended specifications, and the attribute row 'no debtor, fee or scholarship' is the set without the three uncertain-timing attributes.***

When one choice at a time is changed from the passing specification, the upper limit of the control is 2.3 points when the minimum size is 30 students and 2.7 points without the three uncertain-timing attributes. With a size limit of 15% it is 6.2 points, with K = 10 it is 13.1 points, and with one-value groups only it is 12.5 points. The control passes in no specification with a size limit of 15% or more, nor with K = 10 or K = 20. Table IX summarises the runs by the level of each choice. The size limit moves H1 most: its median rises from 21.4 points with the 10% limit to 54.7 with the 35% limit.

***Table IX. Post hoc multiverse summarised by the level of each choice, over the runs that use that level. Differences in percentage points.***

| Choice and level | Runs | H1, median | H1, range | Runs in which N1 passes | N1 upper limit, median |
|---|---|---|---|---|---|
| Size limit 10% | 72 | 21.4 | 16.9 to 41.4 | 2 | 18.0 |
| Size limit 15% | 72 | 38.0 | 11.2 to 58.8 | 0 | 24.2 |
| Size limit 25% | 72 | 49.1 | 11.2 to 71.3 | 0 | 33.4 |
| Size limit 35% | 72 | 54.7 | 15.2 to 80.9 | 0 | 31.2 |
| Budget K = 5 | 96 | 40.1 | 15.2 to 80.9 | 2 | 22.3 |
| Budget K = 10 | 96 | 47.3 | 19.3 to 72.8 | 0 | 31.1 |
| Budget K = 20 | 96 | 31.8 | 11.2 to 47.3 | 0 | 17.5 |
| Minimum size 20 | 96 | 40.0 | 15.2 to 80.9 | 0 | 29.1 |
| Minimum size 30 | 96 | 38.1 | 15.2 to 80.9 | 0 | 24.3 |
| Minimum size 50 | 96 | 33.9 | 11.2 to 80.9 | 2 | 18.2 |
| One-value and pair groups | 144 | 41.2 | 17.6 to 69.7 | 2 | 27.9 |
| One-value groups only | 144 | 30.2 | 11.2 to 80.9 | 0 | 21.6 |
| All attributes | 144 | 39.9 | 11.2 to 69.7 | 2 | 24.3 |
| Without the three uncertain-timing attributes | 144 | 38.0 | 11.2 to 80.9 | 0 | 23.6 |

| Choice and level | Runs | H1, median | H1, range | Runs in which N1 passes | N1 upper limit, median |
|---|---|---|---|---|---|
| Code-point tie-break | 144 | 38.1 | 11.2 to 80.9 | 1 | 23.9 |
| Reverse tie-break | 144 | 38.3 | 11.2 to 80.9 | 1 | 23.9 |

Each specification is run with both tie-breaking orders, so every level of the other choices contains both runs of its specifications. The two passing runs are the two tie-breaking orders of one specification.

The grid does not change the decision on the frozen analysis, which remains not valid, because the plan's decision rule applies to the frozen specification alone. In the grid, the contrast with B1 holds across the varied choices, whereas the negative control does not. Its failure means that, in most specifications, the selection's advantage over random groups on graduates may exceed 2 points. A coverage advantage on dropouts that comes with the same advantage on graduates does not show targeting. The specification that passes sits at the extreme levels of three choices (the lowest size limit, the lowest budget and the largest minimum size), so we do not recommend it: it is one result of a search over 144 specifications (288 runs) and would need its own confirmation.

The plan's analysis with the roles of the two cohorts reversed (E1) was run after the tag: the groups are chosen on the confirmatory cohort, using its labels, and evaluated on the training cohort. The candidate family has 139 groups under the same 25% limit, applied to the confirmatory cohort of 2,518 students, and B1 and B2 are defined as in the frozen analysis. On the training cohort, the greedy rule covers 97.8% of dropouts, against 33.4% for B1 and 72.6% for random groups; the difference from random groups is 25.2 points (interval from 23.8 to 26.6), and Table IV gives its reach, precision, H1 and negative control, whose upper limit exceeds the two-point bound by a wide margin. The pattern resembles the frozen design: reach and coverage of dropouts are high and precision is close to the base rate. Because it chooses groups with confirmatory labels that earlier analyses had already used, this check is not an independent test.

## 7. COMPUTATIONAL COST AND SCALING

Timings were measured on a computer with an eight-core, 16-thread processor and 16 GiB (17.2 GB) of memory. HiGHS ran with its default presolve, and its optimal values for the frozen and amended families at K = 5, 10 and 20 are confirmed with presolve disabled and by a separate branch-and-bound search with a Lagrangian bound, which uses neither HiGHS nor linear programming, in all six cases (post hoc). These checks matter because a property test found a three-element instance on which HiGHS, called through the SciPy version of Section 8, returned a value of 2 with status optimal, where the optimum is 3; that case is kept as an expected failure in the test suite. Disabling presolve also gives the same exact values in all 60 training instances of the calibration grid. The benchmark and stress instances were not checked this way, and their timings use the default setting. The thread pools of the numerical libraries, including the Basic Linear Algebra Subprograms (BLAS), were set to one through environment variables, but HiGHS's own threads were not limited, so its timings may include parallel work, and the plan's single-thread rule for HiGHS was not enforced. Each timing is the median of up to five repetitions (three for the exact solver on the replicated cohorts, and one when the solver reaches the time limit). The analysis plan also asked for interquartile ranges, which are not reported; the results file records the minimum and maximum of the repetitions for the timings by number of groups, and only medians for the replicated cohorts. Memory is the peak Python allocation measured with tracemalloc, which does not count memory allocated by native solver code, so the memory figures are lower bounds. Table X and Figure 7 report the wall-clock time of the whole exact routine, including the assembly of the model, on random subsets of the pool of 482 candidate groups (the frozen family plus groups defined by values of three of the eight attributes used in pairs, within the same size limits), on the frozen family and on the whole pool.

***Table X. Median wall-clock time with BLAS limited to one thread, for the greedy rule and the exact integer programme.***

| Candidate groups \|G\| | Greedy (s) | Exact (s) |
|---|---|---|
| 25 | 0.001 | 0.004 |
| 50 | 0.002 | 0.006 |
| 100 | 0.004 | 0.057 |
| 143 | 0.006 | 0.014 |
| 482 | 0.015 | 0.035 |

Rows 1 to 3 are random subsets, drawn with a fixed seed, of the pool of 482 candidate groups (the frozen family and groups of three attributes); row 4 is the frozen family and row 5 the whole pool. Exact times include model assembly; every exact solve returned status optimal.

For the frozen family, the memory peak was 107,136 bytes for the greedy rule and 1,169,718 bytes for the exact model. Replicating the training cohort 16 times, to 30,496 students, gave a synthetic test for cost only, and raised the greedy time to 0.110 s and the exact time to 0.257 s.

***Figure 7. Wall-clock time, with BLAS limited to one thread and model assembly included. (a) Time against the number of candidate groups: medians of repeated runs on random subsets of the pool (solid lines; error bars give the range where it exceeds the marker), on the frozen family and on the whole pool, which adds groups of three attributes (labelled points); and the single original run of each synthetic stress family (dashed lines; Table XI reports three seeds, with solver-only times). Open markers are runs stopped at the 120 s limit (dotted line); labels give the exact-to-greedy time ratio for the frozen family and the times of two synthetic runs that finished just below the limit. (b) Time against the number of students in replicated training cohorts (synthetic, for cost only), which measures the size of the model, not the difficulty of the choice; the grey dashed line has slope 1, the growth in proportion to n that Equation 2 implies for the greedy rule on a fixed family.***

The stress test uses random families of training students, weighted by their dropout labels, with group sizes drawn uniformly between 30 and 477 students (the generator's upper bound, just above the size limit of 476 used in the analysis), and solves them with K = 10. This is not the size distribution of the real groups: the frozen family has 143 candidate groups, with a median of 169 students and a maximum of 474, so the scale reported below describes these synthetic families and does not transfer to the real family. The original run used one seed. We repeated it with three seeds; the times of the first seed differ from the original ones by less than 10%. These repeated runs time the solver call alone, without model assembly, and their numerical library had loaded before those environment variables were set, so the thread limit does not apply to them. Table XI reports, for each size, how many seeds returned status optimal within the limit, and the median and maximum times: the scale at which the exact method stops being viable depends on the seed, and the 120 s limit locates it only roughly.

***Table XI. Exact solver on synthetic random families, three seeds, time limit 120 s. Medians and maxima over the seeds that ran the size; the first seed reproduces the outcomes and dual bounds of the original run.***

| Candidate groups \|G\| | Seeds with status optimal | Median exact time (s) | Maximum exact time (s) |
|---|---|---|---|
| 25 | 3 of 3 | 0.1 | 0.1 |
| 50 | 3 of 3 | 6.4 | 13.7 |
| 75 | 3 of 3 | 15.9 | 113.7 |
| 100 | 2 of 3 | 113.2 | 120.0 |
| 143 | 0 of 2 | 120.0 | 120.0 |

A seed that reaches the time limit stops the sequence for that seed, as in the original run. Values at the time limit (120 s) are not finished runs. Synthetic families draw group sizes uniformly from 30 to 477 students; they are for cost only and are not used for coverage. Times are the solver call alone,

without model assembly, and BLAS threads were not limited in these runs.

Unlike the random families, the frozen family has nested sets: a pair group lies inside each of its two one-value groups when those also pass the size limit. The gap between the real and synthetic timings may come from this structure or from the strength of the linear relaxation; we did not test either explanation.

## 8. SOFTWARE ENGINEERING AND REPRODUCIBILITY

The analysis is scripted. The primary, amended and multiverse results record the commit of the code, the tag of the analysis plan, the seed and the digest of the data; the primary and amended results also record the versions of the libraries used to compute them. The benchmark and stress results do not record the commit of the code, which the analysis plan asks for. The placebo test and the sensitivity analyses use seeds derived from the base seed by fixed offsets; the results files record the base seed, but not the additional offsets used by the two variant runs of the equal-reach analysis (Table VI). In the primary and amended analyses, the draws for B2 and the bootstrap come from one generator, used in that order, whereas the plan asks for an independent generator for each task; this is a departure from the plan. Raw data are not tracked in Git; a script downloads the file and checks its digest (Section 5.1). The manuscript is built from the results files: numbers in the prose are read from stored results at build time, except a few written by hand. Automated checks block the build when they fail: they verify the schema and geometry of the document against the journal's template, the citations against their verified records, and the origin of every number written by hand. The test suite covers the algorithm, the formulation and the partitions. Figure 8 summarises the pipeline.

An independent implementation, written from the analysis plan and its amendment without the pipeline code, matched the coverage results and both decisions for the two designs at K = 10. Its independence is partial, because the amendment file reports the primary results. The tolerances were set after the first comparison, not declared in advance: 0.5 points for the random-group means, 1 point for interval limits and 0.02 for placebo p-values. The deterministic quantities agree exactly and the others within these tolerances, and the independent implementation found that the frozen coverage depends on the tie-breaking order (Section 6.1). A second independent implementation, written from the analysis plan and from Sections 3 to 5 and two passages of Section 1 of this article, without the code, the amendment file or the results files, recomputed most point estimates of Sections 6.1 and 6.3, but not their intervals or the resampling check, against tolerances declared before the comparison: the deterministic quantities, including the selected groups, agree exactly, and the averages over random draws agree within 1 point. Its independence is also partial, because those passages report three coverages of Amendment 1. It found that the selections of S1, of the rule with equal weights and of Amendment 1 at K = 20 depend on the tie-breaking order. Clean clones of the repository reproduce the analysis numbers from the downloaded data and rebuild the manuscript with identical numbers and citations.

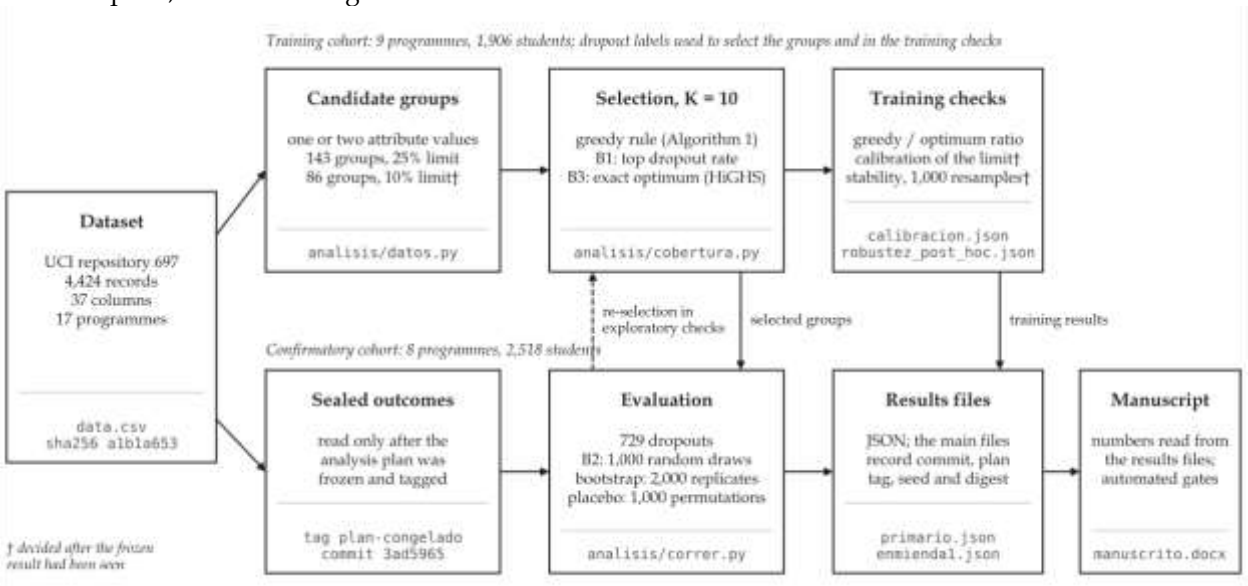



***Figure 8. Architecture of the study. Upper lane: the training cohort, whose dropout labels are used to select***

***the groups and in the training checks; building the candidate groups uses only the attribute values and the group sizes. Lower lane: the confirmatory cohort, whose outcomes the code does not read until the analysis plan is frozen and tagged (totals for the whole dataset had been seen; Section 9); the selected groups are evaluated there, and the results go to files from which the manuscript reads its numbers. The monospaced line at the foot of each box names its artefact, with the repository's own names: the script or results file, the digest of the data or the commit of the plan tag. A dagger marks what was decided after the frozen result had been seen, and the dashed arrow the exploratory checks that re-select groups or specifications with confirmatory labels (Sections 6.4, 6.7 and 6.8).***

***The software used was Python 3.11, NumPy 1.24.1 (Harris et al., 2020), pandas 1.5.2 (McKinney, 2010) and SciPy 1.9.3 (Virtanen et al., 2020), with Matplotlib 3.6.2 (Hunter, 2007) for the figures. Table XII lists the departures from the analysis plan reported in this article.***

***Table XII. Departures from the analysis plan, as reported in the text.***

| What the plan specifies | What was done | Section |
|---|---|---|
| Size limit of 25% of the training cohort | Amendment 1 changed it to 10% after the frozen result had been seen; reported as exploratory | 5.6, 6.3 |
| Decision rule for H1 | The code is followed in four cases that the plan leaves open or words differently | 5.6 |
| Validity tests N1, P1, S1 and S2 | Only N1 and P1 are applied; the plan gives no pass criterion for S1 and S2 | 5.6 |
| An independent random generator for each task | One generator, used first for the draws of B2 and then for the bootstrap | 8 |
| HiGHS limited to one thread | Library thread pools set to one; HiGHS's own threads not limited | 7 |
| Interquartile ranges of the timings | Medians reported; the results file records the minimum and maximum by number of groups | 7 |
| The code commit recorded in every results file | The benchmark and stress results do not record it | 8 |

## 9. THREATS TO VALIDITY

Construct validity. Dropout is the student's final status in the programme, as defined by the dataset description (Realinho et al., 2022). Coverage measures the share of dropouts that a group intervention would reach, not the change in dropout that the intervention would produce; reach, the share of all students covered, is reported separately. The budget counts groups, so a group of 30 students and a group at the size limit cost the same. No limit below 10% was tested, so the groups studied can be larger than a single workshop or tutoring group. Intervention timing is not modelled either: the outcome is the final status at the end of the programme, while groups would be funded earlier.

Internal validity. The groups are defined by attributes that the analysis plan treats as known at enrolment, but their timing is not recorded (Section 5.1). Debtor status, tuition-fee payment status and scholarship holder may have been recorded after enrolment. The sensitivity analysis S1 excludes them, whereas at K = 10 every group of the reference rule B1 uses at least one of them, so its high precision may rest partly on them. Under Amendment 1, the placebo test passes (p-value 0.011), which is a post-selection check rather than a confirmatory validity test; under the frozen design it does not (p-value 0.725). Neither result shows that the selection is specific to dropout. The comparison at equal reach addresses that question without settling it (Section 6.3).

External validity. The data come from one dataset of students in 17 undergraduate programmes. The primary validation uses held-out programmes rather than years, because the dataset has no dates for individual students; Section 6.4 adds post hoc checks across programmes. The confirmatory programmes differ from the training programmes in age and in composition, so the test of transfer crosses a change of population, and one split of one dataset cannot show that the results apply elsewhere. The study does not compare the rule with selections built from individual risk estimates. Such selections choose students, not groups, whereas a group activity reaches the students who share its attributes, so individual selections answer a different question. Methods that learn a budget-constrained treatment assignment policy from observational data (Athey and Wager, 2021), or maximise empirical welfare to choose whom to treat (Kitagawa and Tetenov, 2018), need records of which students received which treatment; this dataset has none. No replication on a second dataset was run. Of the candidate datasets in a catalogue built for this study, the one used here was judged closest to our educational question, because it records dropout and allows a budgeted selection to be studied. The Open University Learning Analytics

Dataset (Kuzilek et al., 2017) could not be downloaded from its official link when the catalogue was built.

Conclusion validity. A tag that precedes the frozen run froze the primary analysis before the confirmatory outcomes were analysed; the tag fixed the plan and the code's guard, not the analysis script, which was committed after it (Section 5.6). The tag is in a private repository, so its date cannot be checked by a reader until the repository is shared, and even then its commit dates are recorded by the author's own system, not by an independent registry. The calibration (Figure 2) first appears in Amendment 1, which is post hoc. Before the tag, the outcome totals of the training cohort and of the whole dataset, which includes the confirmatory cohort, had been seen; no relation between covariates and outcome was computed in the confirmatory cohort. The confirmatory cohort has been used in several analyses: the frozen primary analysis, Amendment 1, the checks across programmes (which re-select groups with its labels), the reversed-direction check, the exploratory variants and the multiverse grid with five seeds. No untouched cohort remains in this dataset. The bootstrap intervals omit variation between programmes. Cluster bootstrap methods exist (Field and Welsh, 2007), but, in our view, a cluster bootstrap would be unreliable with only eight programmes. The placebo p-values have a Monte Carlo resolution of 1 in 1,000.

Social and ethical consequences. Groups defined by attributes such as displacement, gender, scholarship holder or debtor status place students in categories that can carry stigma (nationality and educational special needs were also considered, but form no candidate group here), and a funder that allocates support by such groups must weigh that cost. The amended selection includes tuition-fee payment status (four of its 10 groups), a sensitive measure recorded at an uncertain time; the frozen selection includes none. A funder would have to check such rules against its own non-discrimination obligations, which this study does not assess. The literature on algorithmic bias in education (Baker and Hawn, 2022) bears on such rules; this study does not review it, nor the uplift-modelling literature. Table XIII gives the reach of the amended selection within 10 subgroups of the confirmatory cohort. The rule reaches every debtor and every student aged 22 to 24, but only 37.0% of the students aged 17 to 19, who make up 50.7% of the cohort, and it reaches women less often than men.

***Table XIII. Reach of the amended selection (K = 10) within subgroups of the confirmatory cohort (post hoc, descriptive). The overall reach is 55.4%.***

| Subgroup | Share of the cohort | Reach within the subgroup | Ratio to the overall reach |
|---|---|---|---|
| Women (Gender = 0) | 72.4% | 51.1% | 0.92 |
| Men (Gender = 1) | 27.6% | 66.7% | 1.20 |
| Displaced students (Displaced = 1) | 57.0% | 48.7% | 0.88 |
| Scholarship holders (Scholarship holder = 1) | 29.7% | 41.7% | 0.75 |
| Debtors (Debtor = 1) | 9.5% | 100.0% | 1.81 |
| Special needs (Educational special needs = 1) | 1.2% | 50.0% | 0.90 |
| Age 17 to 19 | 50.7% | 37.0% | 0.67 |
| Age 20 to 21 | 19.1% | 62.4% | 1.13 |
| Age 22 to 24 | 7.3% | 100.0% | 1.81 |
| Age 25 or older | 22.9% | 76.1% | 1.37 |

Reach within a subgroup is the share of its confirmed students who belong to at least one selected group; the ratio divides it by the overall reach. These rates describe whom the rule reaches, not whom it benefits.

Actionability of the groups. The frozen groups combine gender, displacement, daytime or evening attendance and age band; the amended groups keep displacement, daytime or evening attendance and age band, drop gender, and add tuition-fee payment status, debtor status, the parents' occupations and the application order. Daytime or evening attendance and age band can be targeted through scheduling and outreach, but an activity directed at these groups would need its own evaluation before any reduction in dropout could be claimed. Tuition-fee payment status and debtor status are recorded at a time that the dataset does not show, so a funder would first have to confirm when they were recorded. A funder cannot act directly on the parents' occupations or on the application order; these attributes describe whom a group contains rather than what the funder can change.

## 10. DISCUSSION AND CONCLUSIONS

The formulation treats choosing groups under a budget as cardinality-constrained maximum coverage; the decision version is NP-complete for general families of sets, and the greedy rule carries the $1 - 1/e$ guarantee. On the training cohort, the greedy values respect that guarantee for every budget from 1 to 20 and for the 10%, 15% and 25% size limits, which checks the implementation rather than the theorem. On the machine described in Section 7, the exact integer programme solves the frozen family in milliseconds, and the stress tests show that the time at a given size varies widely across seeds. The conclusions concern a budget that counts groups; a cost per group, such as its number of students, as in budgeted maximum coverage (Khuller et al., 1999), could favour other groups.

The empirical result is less favourable. On the first question, the amended selection, made on the training programmes, covers 70.9% of the dropouts of the confirmatory programmes, against 45.9% for random groups (Section 6.3); re-selected within the confirmatory cohort, the rule beats random groups in seven of the eight held-out programmes, with differences from −12.0 to 37.5 points (Section 6.4; post hoc, reusing confirmatory labels). On the second, coverage does not measure targeting.

The frozen primary analysis is not valid. Its eight groups reach every student, so its coverage of dropouts and of graduates is 100% by construction, as the training cohort had already shown. Amendment 1 covers more dropouts than B1 and random groups while reaching a broad share of students, with a precision of 37.1%, above the base rate (29.0%), below that of random groups (38.2%) and slightly above that of random subsets of equal reach (35.5%, a descriptive value). Its negative control fails at the planned budget, and although its placebo test passes, permuted training labels reproduce 19.3 of its 25.0 points of advantage over random groups (Section 6.3). Across 144 specifications, the negative control passes in one of them, at the extreme levels of three choices, and only because its comparator changes (Section 6.7). Under Amendment 1 at $K = 10$, the advantage over random groups of the same budget is larger for dropouts than for graduates on the difference scale (9.1 points), whereas on the ratio scale the rule and random groups are about equal (both 1.56); both comparisons are confounded by reach, and at equal reach the selection leads on both scales (Section 6.3). The sign of the excess over random groups of the same budget changes with the budget and with the design (Section 6.3), and without the three uncertain-timing attributes it falls to 0.9 points, with an interval that includes zero: the signs of targeting that remain depend on those attributes. Against the greedy rule with equal weights, which ignores labels and reaches more students (Section 6.3), the greedy rule, which counts only dropouts, covers slightly more dropouts (70.9% against 68.6%) but far fewer graduates (45.3% against 67.4%). Because the two rules differ in reach, this exploratory difference mixes the role of labels with reach and does not isolate it. The high-precision rule, B1, reaches few students, but it discriminates better than the greedy rule: its excess is larger, although the excess favours broad rules, and it loses the primary contrast of the plan. Reach and precision therefore trade off, and coverage alone hides the trade-off. Against random subsets with more groups on average, the amended selection still covers more dropouts at equal reach, by between 2.9 and 7.4 points depending on the size range of the comparator and the reach window, and covers fewer graduates; this is an exploratory result, with the limits stated in Section 6.3. The reversed-direction check (Section 6.8), which reuses the confirmatory labels, shows the frozen design's pattern: a large dropout difference and a failed negative control.

In this dataset, most of the greedy rule's coverage advantage over random groups came from reaching more students, not from reaching students at higher risk. For a funder, we suggest that a targeting rule be reported with its reach, its precision, its lift, a negative control stated before the analysis and matched on reach, and its behaviour across budgets and programmes. Reported alone, coverage of dropouts would look like success. Evaluations of targeted anti-poverty programmes, for example, compare the share of benefits that reaches the poor with the share under universal allocation, a measure akin to lift, and find wide variation in targeting performance (Coady et al., 2004). The present results do not establish that coverage-based targeting improves student retention. A future study should fix the budget and the size limit before any confirmatory cohort is analysed, check on the training cohort before the freeze that the design does not saturate (the frozen design saturated after eight groups; Section 6.1), and report cost per student as well as the number of groups. It should measure whether, and by how much, an intervention changes dropout, instead of measuring coverage, and examine the fairness of group-based allocation, including whether groups defined by sensitive attributes should be used at all.

**DECLARATIONS**

**Ethics:** No ethics approval was sought. The study analyses a public dataset of student records, as distributed by its source. The author did not verify how the records were extracted or anonymised. The study collects no new data from people and involves no intervention.

**DATA AND CODE AVAILABILITY:** The dataset is available from the UCI Machine Learning Repository (DOI 10.24432/C5MC89), and is licensed under CC BY 4.0 (Realinho et al., 2022). The analysis code and the results files are available from the corresponding author on reasonable request.

**AUTHOR CONTRIBUTIONS:** The author conceived the study, formulated the problem, designed and wrote the analysis software, analysed the data and wrote the manuscript.

## REFERENCES

Arnold, K. E. and Pistilli, M. D. (2012) Course signals at Purdue: using learning analytics to increase student success. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, pp. 267–270. https://doi.org/10.1145/2330601.2330666

Athey, S. and Wager, S. (2021) Policy Learning With Observational Data. Econometrica, 89, pp. 133–161. https://doi.org/10.3982/ECTA15732

Baker, R. S. and Hawn, A. (2022) Algorithmic Bias in Education. International Journal of Artificial Intelligence in Education, 32, pp. 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Bowers, A. J., Sprott, R. and Taff, S. A. (2013) Do We Know Who Will Drop Out? A Review of the Predictors of Dropping out of High School: Precision, Sensitivity, and Specificity. The High School Journal, 96, pp. 77–100. https://doi.org/10.1353/hsj.2013.0000

Chekuri, C. and Kumar, A. (2004) Maximum Coverage Problem with Group Budget Constraints and Applications. In: Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques. Lecture Notes in Computer Science. Springer Berlin Heidelberg. pp. 72–83. https://doi.org/10.1007/978-3-540-27821-4_7

Church, R. and ReVelle, C. (1974) The maximal covering location problem. Papers of the Regional Science Association, 32, pp. 101–118. https://doi.org/10.1007/bf01942293

Coady, D., Grosh, M. and Hoddinott, J. (2004) Targeting Outcomes Redux. The World Bank Research Observer, 19, pp. 61–85. https://doi.org/10.1093/wbro/lkh016

Fawcett, T. (2006) An introduction to ROC analysis. Pattern Recognition Letters, 27, pp. 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Feige, U. (1998) A threshold of ln n for approximating set cover. Journal of the ACM, 45, pp. 634–652. https://doi.org/10.1145/285055.285059

Field, C. A. and Welsh, A. H. (2007) Bootstrapping Clustered Data. Journal of the Royal Statistical Society Series B: Statistical Methodology, 69, pp. 369–390. https://doi.org/10.1111/j.1467-9868.2007.00593.x

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. and Oliphant, T. E. (2020) Array programming with NumPy. Nature, 585, pp. 357–362. https://doi.org/10.1038/s41586-020-2649-2

Herrera, F., Carmona, C. J., González, P. and del Jesus, M. J. (2011) An overview on subgroup discovery: foundations and applications. Knowledge and Information Systems, 29, pp. 495–525. https://doi.org/10.1007/s10115-010-0356-2

Hochbaum, D. S. and Pathria, A. (1998) Analysis of the greedy approach in problems of maximum k-coverage. Naval Research Logistics, 45, pp. 615–627. https://doi.org/10.1002/(sici)1520-6750(199809)45:6<615::aid-nav5>3.0.co;2-5

Huangfu, Q. and Hall, J. A. J. (2018) Parallelizing the dual revised simplex method. Mathematical Programming Computation, 10, pp. 119–142. https://doi.org/10.1007/s12532-017-0130-5

Hunter, J. D. (2007) Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9, pp. 90–95. https://doi.org/10.1109/mcse.2007.55

Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R. and Baron, J. D. (2014) Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. Journal of Learning Analytics, 1, pp. 6–47. https://doi.org/10.18608/jla.2014.11.3

Kempe, D., Kleinberg, J. and Tardos, É. (2003) Maximizing the spread of influence through a social network. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data

mining, pp. 137–146. https://doi.org/10.1145/956750.956769
Khuller, S., Moss, A. and Naor, J. (1999) The budgeted maximum coverage problem. Information Processing Letters, 70, pp. 39–45. https://doi.org/10.1016/s0020-0190(99)00031-9
Kitagawa, T. and Tetenov, A. (2018) Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. Econometrica, 86, pp. 591–616. https://doi.org/10.3982/ECTA13288
Krause, A. and Golovin, D. (2014) Submodular Function Maximization. In: Tractability: Practical Approaches to Hard Problems. Cambridge University Press. pp. 71–104. https://doi.org/10.1017/cbo9781139177801.004
Kuzilek, J., Hlosta, M. and Zdrahal, Z. (2017) Open University Learning Analytics dataset. Scientific Data, 4, 170171. https://doi.org/10.1038/sdata.2017.171
Lakens, D. (2017) Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. Social Psychological and Personality Science, 8, pp. 355–362. https://doi.org/10.1177/1948550617697177
Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J. and Glance, N. (2007) Cost-effective outbreak detection in networks. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 420–429. https://doi.org/10.1145/1281192.1281239
Lipsitch, M., Tchetgen Tchetgen, E. and Cohen, T. (2010) Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology, 21, pp. 383–388. https://doi.org/10.1097/EDE.0b013e3181d61eeb
Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M. T. and Realinho, V. (2021) Early Prediction of Student's Performance in Higher Education: A Case Study. In: Trends and Applications in Information Systems and Technologies. Advances in Intelligent Systems and Computing. Springer International Publishing. pp. 166–175. https://doi.org/10.1007/978-3-030-72657-7_16
McKinney, W. (2010) Data Structures for Statistical Computing in Python. Proceedings of the Python in Science Conference, pp. 56–61. https://doi.org/10.25080/majora-92bf1922-00a
Nemhauser, G. L., Wolsey, L. A. and Fisher, M. L. (1978) An analysis of approximations for maximizing submodular set functions—I. Mathematical Programming, 14, pp. 265–294. https://doi.org/10.1007/bf01588971
Nosek, B. A., Ebersole, C. R., DeHaven, A. C. and Mellor, D. T. (2018) The preregistration revolution. Proceedings of the National Academy of Sciences, 115, pp. 2600–2606. https://doi.org/10.1073/pnas.1708274114
Olaya, D., Vásquez, J., Maldonado, S., Miranda, J. and Verbeke, W. (2020) Uplift Modeling for preventing student dropout in higher education. Decision Support Systems, 134, 113320. https://doi.org/10.1016/j.dss.2020.113320
Realinho, V., Machado, J., Baptista, L. and Martins, M. V. (2022) Predicting Student Dropout and Academic Success. Data, 7, 146. https://doi.org/10.3390/data7110146
Romero, C., González, P., Ventura, S., del Jesus, M. and Herrera, F. (2009) Evolutionary algorithms for subgroup discovery in e-learning: A practical application using Moodle data. Expert Systems with Applications, 36, pp. 1632–1644. https://doi.org/10.1016/j.eswa.2007.11.026
Siemens, G. (2013) Learning Analytics: The Emergence of a Discipline. American Behavioral Scientist, 57, pp. 1380–1400. https://doi.org/10.1177/0002764213498851
Simonsohn, U., Simmons, J. P. and Nelson, L. D. (2020) Specification curve analysis. Nature Human Behaviour, 4, pp. 1208–1214. https://doi.org/10.1038/s41562-020-0912-z
Steegen, S., Tuerlinckx, F., Gelman, A. and Vanpaemel, W. (2016) Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science, 11, pp. 702–712. https://doi.org/10.1177/1745691616658637
Sviridenko, M. (2004) A note on maximizing a submodular set function subject to a knapsack constraint. Operations Research Letters, 32, pp. 41–43. https://doi.org/10.1016/s0167-6377(03)00062-2
Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E.,

Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G. L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O. and Vázquez-Baeza, Y. (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods, 17, pp. 261–272. https://doi.org/10.1038/s41592-019-0686-2

Williamson, D. P. and Shmoys, D. B. (2011) The Design of Approximation Algorithms. Cambridge University Press. https://doi.org/10.1017/cbo9780511921735

Wolff, A., Zdrahal, Z., Nikolov, A. and Pantucek, M. (2013) Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment. Proceedings of the Third International Conference on Learning Analytics and Knowledge, pp. 145–149. https://doi.org/10.1145/2460296.2460324

Youden, W. J. (1950) Index for rating diagnostic tests. Cancer, 3, pp. 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3