Welfare-Maximizing Pooled Testing*

Simon Finster† Michelle González Amador‡ Edwin Lock§
Francisco Marmolejo-Cossío¶ Evi Micha‖ Ariel D. Procaccia**

August 15, 2023

Abstract

Large-scale testing is crucial in pandemic containment, but resources are often prohibitively constrained. We study the optimal application of pooled testing for populations that are heterogeneous with respect to an individual's infection probability and utility that materializes if included in a negative test. We show that the welfare gain from overlapping testing over non-overlapping testing is bounded. Moreover, non-overlapping allocations, which are both conceptually and logistically simpler to implement, are empirically near-optimal, and we design a heuristic mechanism for finding these near-optimal test allocations. In numerical experiments, we highlight the efficacy and viability of our heuristic in practice. We also implement and provide experimental evidence on the benefits of utility-weighted pooled testing in a real-world setting. Our pilot study at a higher education research institute in Mexico finds no evidence that performance and mental health outcomes of participants in our testing regime are worse than under the first-best counterfactual of full access for individuals without testing.

Keywords: pooled testing, welfare maximization, approximation guarantees, non-overlapping, COVID-19, experiment, algorithms.

JEL Codes: C02, C61, C63, C93, D61, I23.

*We thank our collaborators at IPICYT, Luis Salazar Olivo, Rubén López-Revilla, Salvador Ruiz-Correa and Viridiana Robledo Valero, as well as Ángel Gabriel Alpuche Solís and Edgar Daniel Páez Perez from the LANBAMA laboratory, and Kat Molinet. We are grateful to Paul Klemperer, Alex Teytelboym, Robin Cowan, and participants at EC'23 for valuable feedback and comments.

†CREST-ENSAE (Paris) and CNRS/Université Grenoble Alpes, simon.finster@ensae.fr
‡UNU-MERIT & Maastricht University, The Netherlands, mgonzalez@merit.unu.edu
§University of Oxford, United Kingdom, edwin.lock@economics.ox.ac.uk
¶Harvard University, United States, fjmarmol@seas.harvard.edu
‖University of Toronto, Canada, emicha@cs.toronto.edu
**Harvard University, United States, arielpro@seas.harvard.edu

1 Introduction

In a pandemic, large-scale testing forms a crucial part of containment strategies. This has been highlighted in the COVID-19 pandemic, during which social distancing measures, including lockdowns, came with high economic and social costs (Deb et al., 2022; Camera and Gioffré, 2021). Instead, differentiated isolation and quarantining aids in balancing containment and activities (both social and productive). But testing resources can be prohibitively constrained in terms of the supply of reagents, trained personnel or lab equipment. Extensive individual testing is often infeasible, as demonstrated by ample evidence from lower-and-middle income countries during the COVID-19 pandemic (Kavanagh et al., 2020; Dhabaan et al., 2020; Abera et al., 2020). Pooled testing has been a longstanding mechanism for expanding the reach of limited testing resources (Dorfman, 1943). In a pooled test, the samples of multiple individuals are pooled together and tested in a single test. If this test is positive, at least one individual in the pool is infected; otherwise, all pooled individuals are healthy.[1] While a systematic allocation of testing resources is indispensable to maximize the social gain of a limited testing budget, the allocation problem with a large population of individuals is non-trivial, and has been addressed, for instance, by Augenblick et al. (2020); Bobkova et al. (2023) and Lipnowski and Ravid (2021).

Our point of departure from prior work is the observation that different individuals have different utilities for resuming in-person activities. The goal, therefore, should be to maximize an expected aggregation of these utilities. Put differently, a social planner might attribute higher utilities to individuals whose in-person contributions are crucial to their organization. We assume that a given population is *heterogeneous* in that each individual has their own probability of being infected and their own utility for testing negative and therefore being able to resume in-person activities. Our goal is to allocate tests to subsets of the population as pooled tests in order to maximize aggregate expected social welfare. Only negatively tested individuals contribute to social welfare, whereas individuals in positively tested pools remain quarantined and earn a utility normalized to zero.[2]

Our contribution is three-fold: Firstly, we quantify the loss from restricting a testing regime to non-overlapping tests. In non-overlapping tests, each individual can be pooled into at most one test, a constraint that is appealing especially in smaller testing laboratories, for logistical reasons. Secondly, we develop a non-overlapping test allocation mechanism with surprising theoretical guarantees for optimal expected social welfare given *any* population, which achieves high-quality approximations of optimal welfare for realistic populations. The procedure is both conceptually simple, computationally efficient and practical on typical computing facilities. Finally, we implemented our testing framework in a small scale pilot study at the Potosinian Institute of Scientific and Technological Research (IPICYT). We provide numerical evidence of the feasibility and performance of our procedure based on real-world data, and we provide evidence from a randomized controlled trial that our testing mechanism does not negatively impact the participating individuals' productivity and performance at work, nor their subjective well-being and mental health, when the counterfactual is a world without quarantine.

Our problem setting is fundamentally motivated and informed by a collaboration with the Potosinian Institute of Scientific and Technological Research (IPICYT), a higher education research institution in Mexico. In 2021, the entire institute was working remotely with access to campus only in exceptional circumstances, and the IPICYT administration was keen to explore safe and resource-optimal alternatives to ease the social and economical costs of a fully virtual work environment. In September 2022, coinciding with a general reopening of campus facilities, a heterogeneous population of 130 individuals participated in our pilot study. The treatment group

[1] Pooled tests can also address privacy concerns, as pooling provides a certain degree of anonymity to test takers who may not agree to be tested individually.

[2] Unlike in some pooled testing regimes, we do not perform subsequent individual, or multi-stage adaptive testing, following a positive pooled test result, as this is would be prohibitively expensive in a strongly resource-constrained setting.

in this study was invited for qPCR testing based on pooled testing allocations computed by our test allocation mechanism, and allowed access to campus only following a negative qPCR test result.

The National Laboratory of Agricultural, Medical and Environmental Biotechnology (LANBAMA), housed within IPICYT, generously provided qPCR testing. A crucial building block in the application of our algorithms was the determination of population data, consisting of the individuals' utilities from in-person access and their health probabilities. We devised a survey protocol for determining the utilities and worked closely with epidemiologists at IPICYT and in the local state of San Luis Potosí to determine accurate estimates of the probabilities of infection for each member of the population (details are given in Section 5.2).

It is important to note that, although pooled testing increases resource efficiency with regards to testing reagents, this can come at a significant logistical cost for laboratory personnel when pools are overlapping, especially if the series of tests to be performed is complicated and requires delicate tab-keeping of results. The pooled testing regime in which no individual forms part of more than one pooled test is not only conceptually simpler to study but, more importantly, also logistically simpler to implement. We call this testing regime 'non-overlapping'. By contrast, the more general overlapping testing regime allows for individual samples to be allocated to arbitrarily many tests. The following example demonstrates that allowing overlapping tests can result in a higher welfare than non-overlapping testing, even in simple cases. Interestingly, we see that the healthy individual is tested twice in the optimal overlapping test allocation. This is intuitive, as a healthy individual can be included in many tests without affecting the probability of each test being negative.

Example 1.1. Suppose we have a testing budget of $B = 2$, and consider a population of three individuals $\{1, 2, 3\}$ with health probabilities $q_1 = q_2 = 1/2$, $q_3 = 1$ and utilities $u_1 = u_2 = u_3 = 1$. As individuals 1 and 2 are identical, there are four possible non-overlapping test allocations up to symmetry. Allocation $(\{1\}, \{2\})$ achieves welfare 1, and the remaining three allocations $(\{1\}, \{3\})$, $(\{1, 2\}, \{3\})$ and $(\{1, 3\}, \{2\})$ achieve welfare $\frac{3}{2}$. The best overlapping test allocation $T^* = (\{1, 3\}, \{2, 3\})$ tests individual 3 twice and achieves a welfare of $u(T^*) = \frac{7}{4}$.

Our first research question is whether non-overlapping tests may be vastly outperformed by overlapping regimes. After all, if that is the case then supporting overlaps may be worth the logistical overhead. During a pandemic, testing regimes have to be robust to a vast multitude of infection states and heterogeneity between individuals. Thus, our focus is on the analysis of worst-case scenarios, i.e. we establish welfare guarantees that hold for any configuration of health probabilities and individual utilities. In this vein, we show that the worst-case ratio between the welfare of the best overlapping testing regime and the best non-overlapping regime is at most 4, and in special cases even smaller. While a factor of 4 is admittedly significant, the worst example we know of gives a ratio of $7/6$. Qualitatively, we interpret these results as a justification to focus on non-overlapping testing regimes. This is confirmed by our empirical results, which indicate only small gains from overlaps in practice.

Turning to the challenge of computing testing regimes, even without overlaps, it is known from previous literature in discrete mathematics that the problem of determining an optimal regime is computationally intractable if the pool size is unbounded[3] and even with a bounded pool size, the problem is not computationally feasible in practice (see details in Section 4). However, we were able to design a greedy polynomial-time algorithm that finds an approximate solution to the welfare-maximization problem, which, in the worst case, is guaranteed to achieve at least one fifth of welfare in the optimal non-overlapping testing regime. We then compare the performance of our greedy algorithm with optimal non-overlapping testing empirically. In order to compute (approximately) optimal non-overlapping testing regimes, we model the problem as a

[3]By computationally intractable we mean that no polynomial time algorithm exists, a property also known as NP-hardness.

mixed-integer linear program which is tractable in our simulations. We choose a population size, pool size constraints and testing budgets that mirror realities at IPICYT. Our results indicate that our algorithm computes near-optimal testing regimes, and vastly outperforms the mixed-integer linear program with respect to running time.

We also provide empirical evidence in support of our approach to welfare-maximizing pooled testing in a resource-constrained environment by evaluating the randomized controlled trial at IPICYT. For the pilot study, we developed a web application that formed the center point for participants, administrators and the LANBAMA testing laboratory, and implemented our algorithm to compute near-optimal non-overlapping group testing regimes.[4] Our trial results suggest that, compared to a best-case scenario of free mobility and full access to institutional resources (but not testing resources), our testing approach is just as efficient in terms of performance, learning, and mental health outcomes. At the same time, our protocol, which ensures that only negatively qPCR-tested individuals have in-person access, and safeguards the population's health within the institution, unlike a full reopening without testing, and at a fraction of the cost of an individual qPCR testing regime.[5]

Related work. Pooled testing dates back to the seminal work of Dorfman (1943), and has since become a mature field in its own right with a rich literature of protocols. In contrast to our test allocation mechanism, a large part of this literature aims to ascertain the infection status of *all* individuals in a population with a minimal number of tests.[6] It has been applied to combat various diseases in the past, especially HIV/AIDS (Tu et al., 1995; Wein and Zenios, 1996; Emmanuel et al., 1988). From the outset of the COVID-19 pandemic, it became clear that testing resource constraints would be a large issue for many countries, and hence pooled testing became a viable option for combating the virus, especially as it was shown that qPCR tests can be sensitive enough to pool samples in a pooled test (Sanghani et al., 2021; Mutesa et al., 2021; Nalbantoglu, 2020).

In the economics literature, various recent contributions aim to characterize optimal testing allocations. Lipnowski and Ravid (2021) study welfare-maximizing optimal pooled testing in a population heterogeneous in individual health probabilities. They describe a unique optimal test allocation exhibiting 'assortative batching', meaning that individuals with different health probabilities are never pooled together. This highlights an important difference to our work, as in our more general setup individuals of varying health probabilities may also differ in the utility that materializes when tested negatively. Secondly, in the model of Lipnowski and Ravid (2021), individuals may be quarantined regardless of their test result, depending on the benefit and cost of release and their posterior probability of being infected. Consequently, the available tests are assigned to individuals whose quarantine decision prior to testing is most uncertain. We take a more cautious approach to quarantining: only negatively tested individuals are released. Lipnowski and Ravid (2021) also demonstrate that individuals with lower health probabilities are tested in smaller pools due to smaller informational externalities. While this intuition translates to our model (see Example 1.1), the optimal allocation is complicated by the fact that different utilities may weight externalities differently. By further contrast with our work, Lipnowski and Ravid (2021) consider *continuous* populations and focus exclusively on non-overlapping testing regimes. Our focus is on quantifying the welfare discrepancy in overlapping and non-overlapping testing and on practical algorithms to compute an optimal testing allocation. We consider

[4]The web app code is available as open source at `https://github.com/edwinlock/csef.git`. A demo of the app can be accessed at `https://demo.c-sef.com`.

[5]At the time of reopening, San Luis Potosí had 221,870 cumulative COVID-19 cases, of which 615 were active. KN95 Masks were mandatory for everyone returning to IPICYT.

[6]The unfortunate reality is that many resource-constrained populations are in a situation where their testing budget falls far below the information theoretic lower bounds required to precisely ascertain the health profile of all individuals. Moreover, complex pooled testing regimes can be difficult to implement logistically at scale with limited laboratory personnel and workflow infrastructure (Cleary et al., 2021).

heterogeneous utilities to be indispensable in practical mechanisms as they may help improve efficiency as well as potential inequalities, taking into account for example individual productivity or vulnerabilities.

Ely et al. (2021) study a model where a policymaker can employ tests of different types, each with differential costs and sensitivities. The policymaker has an overall budget, and test allocations are measured with respect to the rate at which they correctly classify individuals as infected or healthy. Brault et al. (2021) focus on limited pooled tests for early screening at a non-diagnostic level with high penalties associated with false negatives. Gollier and Gossner (2020) study pooled tests as a means to estimate infection prevalence and to find healthy individuals in a population. The main differences between their work and ours is that we consider a heterogeneous population as well as upper bounds on pool sizes imposed by lab constraints. Bobkova et al. (2023) characterize the optimal test allocation in a continuous population that is heterogeneous in infection probabilities. They focus on the case where the population can be divided into a low-risk and a high-risk group, and consider pooled testing where each individual can be tested at most twice. Augenblick et al. (2020) study testing frequency as a crucial factor to limiting viral spread in a pandemic, and how pooled testing can increase the reach of a rapid frequency testing regime when tests are limited.[7]

Although not cast as a pooled testing paper, the results of Goldberg and Rudolf (2020) can be interpreted as computing the optimal allocation of a single (arbitrarily large) pooled test to a heterogeneous population as in our model setting. The authors show that computing an optimal single test allocation cannot be achieved in polynomial time, but they provide a procedure[8] for finding an approximately optimal single test allocation; we use their procedure as a component of our new algorithm.

The remainder of the paper is organized as follows. In Section 2, we describe our population and testing model. Section 3 presents the results on the optimality of non-overlapping compared to overlapping test allocations. In Section 4, we showcase our algorithms, including the Greedy heuristic and the MILP formulation. Section 5 details the implementation of our testing framework at IPICYT, our simulations on real-work data, as well an evaluation of the social and economical benefits of the test allocation mechanism. Section 6 is a discussion and outlook. All proofs are presented in the Appendix.

2 Model

Let $[n] \coloneqq \{1, \dots, n\}$ denote a collection of n individuals and $B \in \mathbb{N}$ be the testing budget. A *population* J of size n is a tuple $(p_1, \dots p_n, u_1, \dots, u_n)$ which assigns each individual $i \in [n]$ an independent probability of infection $p_i \in [0, 1]$ and a utility $u_i \geq 0$ that captures their gain of returning to in-person activities.[9] We also let $q_i = 1 - p_i$ denote the probability that an individual is healthy. We denote the universe of all populations by $\mathcal{J} \coloneqq \Delta^n \times \mathbb{R}^n_+$.

Fix a population J. A single test consists of samples of a subset of the individuals, which we identify with a set $t \subseteq [n]$ of the individuals whose samples are included in the test. Test sizes are bounded by a pool size constraint G, so $|t| \leq G$ for all tests t. We are primarily interested in pool sizes $G < n$.[10] For convenience, we introduce the notation $q_S = \prod_{i \in S} q_i$, for any $S \subseteq [n]$, to express the probability that all individuals in S are healthy; hence, q_t is the probability that test t is negative. A *test allocation* $T = (t_1, \dots, t_B)$ is an ordered collection of B tests satisfying $|t_j| \leq G$ for all $j \in [B] \coloneqq \{1, \dots, B\}$.

[7] See also Larremore et al. (2021).

[8] This procedure is a fully polynomial-time approximation scheme (FPTAS).

[9] Utility might reflect people's socioeconomic status, the type of occupation, or mental health considerations. See Section 5.2 for details on the utilities in our pilot.

[10] Pool sizes in pooled tests are limited due to biological constraints. Our partners in Mexico have replicated techniques from Sanghani et al. (2021) to achieve a maximal pool size of 5 with saliva samples.

For a given test allocation T, let P_i^T denote the probability that $i \in [n]$ is included in at least one negative test $t_j \in T$. A test allocation only earns utility from individuals who return to in-person activities as a result of being in a negative test. We let $u(T)$ denote the *welfare*, i.e. the aggregate expected utility, earned under test allocation T.[11] Linearity of expectation allows us to express the welfare of T as $u(T) = \sum_{i\in[n]} u_i \cdot P_i^T$. In addition, we let $u(t) := u(\{t\}) = q_t \left(\sum_{i\in t} u_i\right)$ for a single pooled test t. A test allocation T is *optimal* (for a given population) if it maximizes welfare. Without loss of generality, we assume that $B < n$. If this is not the case, testing every person in the population individually is optimal. When the population J is not fixed, we denote $u(T, J)$ the welfare of a test allocation T in population J.

Non-overlapping test allocations. As discussed in the Introduction, we are particularly interested in non-overlapping test allocations, which include each individual in at most one test. Formally, a test allocation T is *non-overlapping* if $t \cap t' = \emptyset$ for all distinct tests $t, t' \in T$. In general, P_i^T can be a complicated expression due to correlation between overlapping tests. In a non-overlapping test allocation, T, on the contrary, test results are independent of each other, and the welfare of T is given by $u(T) = \sum_{t\in T} u(t)$.

Independence of Infections. In general it may be the case that infections in a population are correlated. However, we emphasize that our testing model is intended for a regime wherein all individuals in the given population are assumed to be in full lockdown, hence social interactions at the workplace do not contribute to potential correlation of infection for two key reasons: Either individuals who would potentially interact are forcibly at home, and hence no longer interact, or if the individuals are interacting at the workplace, it is because they are both in a negative test and hence cannot infect each other.

3 Performance of non-overlapping testing

In general, overlapping testing can achieve higher welfare than test allocations that are restricted to not overlap, as demonstrated in Example 1.1. However, non-overlapping test allocations are often strongly preferred for logistical reasons. This was the case with our partner institution, which was running a small testing lab in which the assignment of individual samples to more than one test was close to infeasible. A natural question is to identify how much welfare may be lost by restricting to non-overlapping tests. If the difference in welfare achievable with overlapping and non-overlapping testing is not too large, even institutions with the logistical capacity to run overlapping tests may choose the latter. Our goal in this section is to provide robust upper bounds on the increase in welfare achievable when we allow tests to overlap.

Given a population J and budget B, we define the *overlap welfare ratio* $\mathscr{R}(B, J)$ *for budget* B *and population* J as the ratio of the welfare of an optimal test allocation over the welfare of an optimal non-overlapping test allocation. Formally, we let $\mathcal{T}^*(B)$ and $\mathcal{T}(B)$ respectively denote the space of all test allocations and all non-overlapping test allocations with testing budget B. We define optimal welfare from overlapping and non-overlapping test allocations as $\mathrm{OPT}^*(B, J) := \max_{T^*\in\mathcal{T}^*(B)} u(T^*, J)$ and $\mathrm{OPT}(B, J) := \max_{T\in\mathcal{T}(B)} u(T, J)$, respectively. The overlap welfare ratio for budget B and population J is given by $\mathscr{R}(B, J) = \frac{\mathrm{OPT}^*(B,J)}{\mathrm{OPT}(B,J)}$.

Definition 3.1. The *overlap welfare ratio* for budget B is an upper bound on the welfare increase from allowing overlapping test allocations across all possible populations J. Formally,

$$\mathscr{R}(B) = \sup_{J\in\mathcal{J}} \mathscr{R}(B, J).$$

[11] We omit the term 'expected' for brevity and assume that all welfares and utilities are determined in expectation.

Recalling Example 1.1 with a population J of three individuals and a budget B of 2 from the introduction, we saw that the maximum welfare achievable by testing with and without overlaps is $\text{OPT}^*(2, J) = \frac{7}{4}$ and $\text{OPT}(2, J) = \frac{3}{2}$, respectively. Hence the overlap welfare ratio for budget 2 and the given population is $\mathscr{R}(2, J) = \frac{7}{6}$, demonstrating that the ratio can be greater than 1. The example immediately implies Proposition 3.2.

Proposition 3.2. *The overlap welfare ratio $\mathscr{R}(2)$ for budget 2 is at least $\frac{7}{6}$.*

We next show in Proposition 3.3 that Example 1.1 also represents a worst case when we have a budget of $B = 2$. That is, no population achieves a higher overlap welfare ratio with 2 tests than $\frac{7}{6}$ achieved with the population from Example 1.1. In the example, welfare was maximized by including the healthy individual in both tests. Similarly, in the proof of Proposition 3.3 we use a property that the overlap welfare ratio is maximized when the health probabilities of the individuals included in both tests are 1. The proof of Proposition 3.3 works by dissecting an optimal overlapping two-test allocation $T^* = (t_1^*, t_2^*)$ for an arbitrary population into the individuals A contained only in the first test, the individuals B contained only in the second test, and the individuals C in both tests. We then consider the four non-overlapping two-test allocations $T^1 = (A \cup C, B)$, $T^2 = (A, C)$, $T^3 = (B, C)$ and $T^4 = (A \cup B, C)$ and show that at least one of these allocations achieves a welfare that is greater or equal to $\frac{6}{7}$ of the welfare achieved by T^*.

Proposition 3.3. *The overlap welfare ratio $\mathscr{R}(2)$ for budget 2 is at most $\frac{7}{6}$.*

Stated differently, with a budget of two tests, allowing for overlaps increases welfare by at most 16.7%. Similar results can be derived for testing budgets of 3 and 4, with larger upper bounds. The proof of Proposition 3.4 considers all subsets of individuals who are included in exactly k sets, for all $1 \leq k \leq B$. These subsets are non-overlapping by construction, and we show that the B subsets with the largest utility achieve an overlap welfare ratio no greater than $\frac{7}{3}$ for $B = 3$ and $\frac{15}{4}$ for $B = 4$.

Proposition 3.4. *The overlap welfare ratios for budgets $B \in \{3, 4\}$ are bounded by $\mathscr{R}(3) \leq \frac{7}{3}$ and $\mathscr{R}(4) \leq \frac{15}{4}$.*

Proposition 3.4 states that allowing overlaps cannot increase welfare by more than 133.3% with three tests, and by more than 275% with four tests. Given these sharp increases in upper bounds of the overlap welfare ratio, one might expect further increases from adding even more tests to the budget, or the ratio being even unbounded for $B \in \mathbb{Z}_+$. Surprisingly, however, we show that allowing overlaps increases welfare by at most 300% for *any testing budget*, i.e. even if the budget is arbitrarily high, for any populations and any pool sizes.

Theorem 3.5. *For any budget B, the overlap welfare ratio $\mathscr{R}(B)$ is at most 4.*

We emphasize that the proof of Proposition 3.4 does not provide any indication that there exist populations that actually achieve a gain from allowing overlaps of $\frac{7}{3}$ and $\frac{15}{4}$ with budgets 3 and 4, and Theorem 3.5 does also not give any indication that some population may actually be 300% better off with overlapping testing. In fact, we do not know of any populations that achieve an overlap welfare ratio greater than the moderate gain of $\frac{7}{6}$. Extensive computational searches we performed have failed to identify any such populations. We conjecture that the overlap welfare ratio is close to $\frac{7}{6}$ even for budgets greater than 2. That is, our upper bounds are not tight, albeit surprising in light of the richness of our model.

We conclude this section by sketching the main arguments of the proof of Theorem 3.5. The full proof is given in Appendix A.2. Given an optimal overlapping test allocation T^* with budget B for some population, the proof works by constructing two non-overlapping test allocations T^1 and T^2 whose tests are no larger in size than the tests in T^*, and showing that T^1 or T^2 achieves

an overlap welfare ratio no larger than 4. As this construction works for any population, testing budget and pool size, this proves Theorem 3.5.

We now provide more detail on how T^1 and T^2 are constructed and how we prove the overlap welfare ratio bound of 4. First we construct an intermediary non-overlapping test allocation $T = \{t_1, \ldots, t_B\}$ from T^* that removes each tested individual from all but the first test of T^* in which the individual appears. (As a result, some tests in T may be empty.)

For each test t_j of T, we choose a subset $S_j \subseteq t_j$ of smallest cardinality so that the probability of all individuals in S_j being healthy is strictly less than $\frac{1}{2}$. This means that the probability of any strict subset of individuals in S_j being healthy is at least $\frac{1}{2}$. If such a subset does not exist, we let S_j be the empty set. The two allocations T^1 and T^2 are defined as $T^1 \coloneqq (S_1, \ldots, S_B)$ and $T^2 \coloneqq (t_1 \setminus S_1, \ldots, t_B \setminus S_B)$. By construction, the individuals included in allocation T are split between T^1 and T^2.

The key technical step in the proof consists of Lemma A.4 which, broadly speaking, tells us that no test in T can be partitioned into two subsets which both have probability less than $\frac{1}{2}$ of being healthy.[12] For the proof of Lemma A.4, we devise a novel 'pivoting' technique that allows us to express the overall welfare of a test allocation as the sum of the expected marginal utilities gained from each consecutive test if we assume (for the sake of proof) that the tests are applied sequentially in order (see Observation A.2 and surrounding discussion for details).

In particular, as the probability q_{S_j} is less than $\frac{1}{2}$ due to our choice of S_j, Lemma A.4 tells us that the probability $q_{t_j \setminus S_j}$ of all individuals in $t_j \setminus S_j$ being healthy is at least $\frac{1}{2}$. The construction of the sets S_j thus ensures that $q_t \geq q_i \cdot \frac{1}{2}$ for any test t in T^1 or T^2 and any individual $i \in t$. We use this, together with the fact that T^1 and T^2 are non-overlapping, to argue that the welfare achieved by T^1 is $u(T^1) = \frac{1}{2}\sum_{i \in T^1} q_i u_i$, and likewise for T^2. It follows that $\max\{u(T^1), u(T^2)\} \geq \frac{1}{4}\sum_{i \in T} q_i u_i$. As the welfare of allocation T^* is at most $u(T^*) \leq \sum_{i \in T} q_i u_i$, the result follows.

4 Finding near-optimal test allocations

We now discuss how to determine approximately optimal test allocations. As we show, computing an exact optimal test allocation is computationally intractable even when the budget consists of a single test, as well as in practically relevant applications. Thus, our main goal is to establish a simple but effective heuristic for this problem. We prove that our heuristic produces a test allocation with expected welfare that is at least 20% of expected welfare in the optimal non-overlapping test allocation *for any population, and regardless of the pool size and testing budget*. While it is in itself surprising that such a theoretical guarantee exists, our computational experiments also show that, in practice, the algorithm we design performs extremely close to optimal, i.e. the welfare it achieves is within 99.5% of optimal non-overlapping welfare (see Tables 1 and 3 for outcomes of numerical experiments on pilot data). In special cases, we also show better theoretical guarantees: when the number of individuals is large relative to the testing budget, our heuristic finds an optimal test allocation; and when utilities are homogeneous, a variation of our heuristic achieves an e-approximation, i.e. at least roughly 37% of optimal non-overlapping welfare. Our heuristic is also the only computational method we know of that can find a near-optimal non-overlapping test allocation in a reasonable amount of time for realistic problem sizes such as the population encountered in our pilot.

First, consider the problem of computing an optimal test allocation with a budget of $B = 1$, where the non-overlapping and overlapping testing regimes coincide. For unbounded pool sizes, the single-test problem is NP-hard. This was shown by Goldberg and Rudolf (2020) in their work on the maximum expected value all-or-nothing subset problem. When the pool size is bounded by a constant $G < n$, an exhaustive search procedure that explores all possible test allocations checks

[12] This lemma is stated and proved in the appendix, and is a non-trivial generalization of Lemma 6 in (Goldberg and Rudolf, 2020). A variant of this lemma is also used in the proof of Theorem 4.1, our main result in Section 4.

$\sum_{k\in[G]} \binom{n}{k} = O(n^G)$ possible test constellations. Even though this exhaustive search algorithm runs in polynomial time for a fixed pool size, reasonable choices of pool sizes such as $G = 5$ or $G = 10$ for saliva testing (as implemented by our partner institution), and pool sizes of $G = 32$, $G = 48$, and $G = 57$ for nasal swab testing (Yelin et al., 2020; Shental et al., 2020; Theagarajan, 2020) make this procedure prohibitively expensive in practice.

In Section 4.2, we describe two methods for computing a single-test allocation. Both methods return a single test with a welfare that is within a factor of $1 - \varepsilon$ of optimal welfare, with $\varepsilon > 0$ denoting a tiny error tolerance that can be adjusted when using either method. Our heuristic relies on the repeated application of one of these single-test allocation procedures. In order to evaluate its performance in terms of optimal welfare and running time, we must also solve the optimal welfare problem with non-overlapping tests exactly. We tackle this problem with a convex programming formulation. For computational reasons, we again develop an approximation via a mixed-integer linear programming formulation (MILP), and we establish convergence to, and an approximation guarantee very close to, the optimal welfare. The MILP formulation is discussed in Section 4.3. Our numerical experiments show that our heuristic is significantly faster than solving the MILP.

4.1 A test-by-test heuristic

Our heuristic consists of a serial application of single-test allocations. We describe two single-test allocation algorithms below in Section 4.2, one based on a fully polynomial-time approximation scheme (FPTAS) and another formulating the problem as a mixed-integer conic optimization problem (MICP). For now, fix a single-test allocation mechanism. Our algorithm uses this mechanism to find an (approximately) optimal test among the population, and adds this test to the overall test allocation. Disregarding all individuals that have already been included in tests, the single-test mechanism is used again to find the next (approximately) optimal test allocation among the remaining individuals. The procedure continues in the same fashion until the testing budget or the population is exhausted. This results in a non-overlapping test allocation, as we never consider individuals that have already been included in a test. We refer to this algorithm as Greedy. The performance of Greedy depends, unsurprisingly, on the performance of the single-test allocation mechanism. Given some population J, suppose $\mathrm{OPT}(1, J)$ denotes the welfare achieved by an optimal single-test allocation mechanism, and let $\widetilde{\mathrm{OPT}}(1, J)$ denote the welfare achieved by a single-test approximation algorithm. For both our single-test allocation mechanisms it holds that $\frac{\widetilde{\mathrm{OPT}}(1,J)}{\mathrm{OPT}(1,J)} = 1 - \varepsilon$, where the approximation error ε is 10^{-7} for the conic program, and can be set arbitrarily small for the FPTAS.

Theorem 4.1. *For any population and testing budget,* Greedy *achieves at least* $\frac{1-\varepsilon}{5}$ *of the welfare of the optimal non-overlapping test allocation, where* ε *is the error tolerance of the single-test allocation mechanism.*

To provide an overview of the proof of Theorem 4.1, suppose that T^* is an optimal non-overlapping test allocation and that T is the test allocation returned by Greedy which tests individuals N'. Without loss of generality, we can assume that $N' = \{1, \dots, n'\}$ for some $n' \leq n$. The first step of the proof demonstrates that T obtains $(1 - \varepsilon)$ of the welfare achieved by T^* from individuals not in N', so $u(T) \geq (1 - \varepsilon) \sum_{j\in B} q_{t_j^*} \cdot \left(\sum_{i\in t_j^* \setminus N'} u_i\right)$. This follows from the fact that Greedy picks t_j, and not $t_j^* \setminus N'$, as the j-th test. Indeed, the fact that it chooses t_j means that t_j contributes a utility of at least $(1 - \varepsilon)$ of the utility of test $t_j^* \setminus N'$; and this, in turn, is at least the utility T^* obtains from individuals in $t_j^* \setminus N'$. The second part of the proof shows that T obtains at least $\frac{(1-\varepsilon)^2}{4}$ of the total expected utility of healthy individuals contained in N', which is given by $\sum_{i\in N'} q_i \cdot u_i$. The latter is also an upper bound on the utility that allocation T^* obtains from individuals in N'. Hence, T^* achieves a utility of at most $\frac{4}{(1-\varepsilon)^2} u(T)$ utility

from individuals in N' and at most $\frac{1}{(1-\varepsilon)^2}u(T)$ from individuals in $[n] \setminus N'$. Summing these two utilities shows us that T^* achieves a utility no higher than $\frac{5}{(1-\varepsilon')}u(T)$ for $\varepsilon' \geq 2 \cdot \varepsilon$.

Note that we can also combine Theorem 4.1 with Theorem 3.5 to show a constant-factor bound on how well Greedy performs in comparison to overlapping testing.

Homogeneous utilities. Next we consider the special case in which all the individuals have the same utility. The following example demonstrates that Greedy may produce suboptimal allocations (though necessarily bounded in their suboptimality via Theorem 4.1). In the example, all individuals have the same utility and health probability; the health probability is chosen such that Greedy will exhaust the entire population with a single test, rendering the remaining $B-1$ tests useless. In the limit, as the population size goes to infinity, the example shows that Greedy can perform a factor of e worse than optimal overlapping testing.

Example 4.2. Consider a population of n individuals with homogeneous utilities $u = 1$ and probabilities $q = \frac{n-1}{n}$, and a testing budget of n. Clearly, the optimal testing strategy is to test every person individually, which yields an expected welfare of $nq = n-1$. The Greedy algorithm tests the entire population with its first test, as $nq^n \geq kq^k$ for all $k \leq n$. Hence Greedy achieves an expected welfare of $nq^n = n(\frac{n-1}{n})^n$, implying a welfare ratio of $(\frac{n-1}{n})^{1-n} \to e$ as $n \to \infty$.

We design a variant of Greedy which sorts the individuals in decreasing order with respect to the probability of being healthy and, in each step, adds individuals to the current test as long as the expected utility of the test increases. In Appendix B.2.2, we show that this OrderedGreedy algorithm returns an e-approximate non-overlapping test allocation. When applying OrderedGreedy to Example 4.2, all individuals are also pooled into a single test, hence this shows that the approximation ratio achieved by OrderedGreedy is tight.

Proposition 4.3. *If all individuals have the same utility,* OrderedGreedy *returns an e-approximate non-overlapping test allocation.*

Suppose that T is a test allocation returned by OrderedGreedy which tests individuals $[n'] \subseteq [n]$. The proof of approximate optimality of T rests on showing that if an individual i is included in test t_j of size $|t_j| = k$, then $q_i \geq (k-1)/k$. This implies that $q_{t_j} \geq q_i \left(\frac{k-1}{k}\right)^{k-1} \geq q_i \frac{1}{e}$. It follows that $u(T) \geq \frac{1}{e}\sum_{i\in[n']} q_i \cdot u_i$. On the other hand, the welfare of T^* is upper-bounded by $u(T^*) \leq \sum_{i\in[n']} q_i \cdot u_i$, as this is the expected utility of healthy individuals tested in T^*.

Optimal performance with clusters. Finally we consider the scenario in which individuals can only exhibit utilities and probabilities from a finite set of values. Specifically, we assume that the population at hand can be partitioned into m disjoint clusters, $C_1, \ldots C_m$, where C_ℓ, has n_ℓ individuals with identical utility, u_ℓ, and probability of infection p_ℓ (probability of health q_ℓ). Since individuals are indistinguishable within a cluster, we can identify a test t with the number of individuals included from each cluster. We let $t(\ell)$ denote the number of individuals from cluster C_ℓ included in t. Suppose that t^* is a single test that achieves optimal utility and that t is a near-optimal test with $u(t) \geq (1-\varepsilon)u(t^*)$. We can show that if the population at hand is such that t can be repeated B times, then not only does Greedy return this allocation, but it is also obtains $(1-\varepsilon)$ welfare of what is obtained by an optimal overlapping test allocation.

Proposition 4.4. *Suppose that t is a $(1-\varepsilon)$-optimal test and $B \cdot t(\ell) \leq n_\ell$ holds for each $\ell \in [m]$. Then* Greedy *returns an optimal allocation that applies B tests with the same composition as t. This allocation obtains $(1-\varepsilon)$ welfare of what is obtained in an optimal overlapping test allocation.*

The intuition behind this result is simple. Suppose that $T^* = (t_1^*, \ldots, t_B^*)$ is an optimal overlapping test allocation and that t^* is an optimal single pooled test for the population. Using Observation A.2, we can decompose $u(T^*)$ as the sum of marginal utility obtained from each

test when T^* is assumed, for the sake of the proof, to be applied sequentially. We can show that each marginal utility is bounded by $u(t^*)$, and so $u(T^*) \leq B \cdot u(t^*)$. On the other hand, $u(t) \geq (1-\varepsilon)u(t^*)$ and the fact that $B \cdot t(\ell) \leq n_\ell$ holds for each $\ell \in [m]$ means that t can be applied B times in the population. This what GREEDY does, returning an allocation T with welfare $u(T) \geq (1-\varepsilon)u(T^*)$.

4.2 Optimally allocating a single test

4.2.1 Designing an FPTAS

Goldberg and Rudolf (2020) introduce a type of algorithm commonly known as a fully polynomial-time approximation scheme (FPTAS) that runs in time $O\left(n^5/\epsilon\right)$.[13] This FPTAS returns a test with an almost-optimal welfare; the test returned achieves a factor of $1-\varepsilon$ of the optimum achievable welfare. Here $\varepsilon > 0$ is a positive value that can be chosen arbitrarily small. The algorithm of Goldberg and Rudolf (2020) returns a test with an arbitrary pool size. In the appendix, we show how their algorithm can be adapted to return an almost-optimal single-test allocation when a constant pool size constraint is imposed.

To provide intuition for the FPTAS, we first describe at a high level a dynamic program that exactly solves for an optimal test. For $i \in [n]$, let $P(i, C, L) = \max\{q_S \mid S \subseteq [i], |S| = L, \text{ and } \sum_{\ell \in S} u_\ell = C\}$ denote the largest negative probability of a subset of $[i]$ of cardinality L and total utility C. Note that $\bar{C} = \sum_{i \in [n]} u_i$ is an upper bound on the sum of utilities of an optimal test, and pooled tests are bounded by G. We can compute $P(i, C, L)$ for $C \in [\bar{C}]$ and $L \in [G]$ via dynamic programming in $O(nG\bar{C}) = O(n^2\bar{C})$ time. The test with maximal utility is the test that maximizes $C \cdot P(i, C, L)$. Our approach is similar to the dynamic program used in Goldberg and Rudolf (2020), but differs in that we keep track of pool sizes using parameter L.

This dynamic program does not run in polynomial time due to the time dependence on $\bar{C}$, which is exponential in the binary representation of the population data. Instead, our polynomial-time algorithm uses a similar dynamic program as described above, but replaces each agent's utility by $\lfloor u_i/\kappa \rfloor$.[14] Larger values of κ give rise to better running times (as the resulting problem is smaller), but result in worse approximation guarantees. Choosing the value of κ carefully as a function of the desired approximation parameter ϵ, we achieve a polynomial running time of $O(n^5/\epsilon)$ while retaining an approximation factor of $1-\varepsilon$.

Proposition 4.5. *The FPTAS presented in Appendix B.3 finds an almost-optimal single-test allocation with pool size at most G in time $O(n^5/\epsilon)$.*

4.2.2 Formulating a conic optimization problem

The problem of allocating a single test can also be formulated as a mixed-integer conic optimization program (MICP), and solved using a commercial conic solver. This implementation is used in the web app accompanying our pilot study. We formulate the MICP as follows.

Define the indicator vector $\boldsymbol{x} \in \{0,1\}^n$ with $x_i = 1$ if individual i is included in the (single) test, and $x_i = 0$ otherwise. Our objective is to maximize the expected utility from the test given by $\sum_{i \in [n]} u_i x_i \cdot \prod_{i \in [n]} q_i^{x_i}$. We impose a pool size between 1 and G with constraint $1 \leq \sum_{i \in [n]} x_i \leq G$. In order to isolate the non-linear element of the optimization problem, we maximize the logarithm of the original objective, and introduce the variables $z = \sum_{i \in [n]} u_i x_i$ and $y = \log z$. Note that we can we relax the equality in the only remaining non-linear constraint $y = \log z$ to $y \leq \log z$ without affecting the outcome. The resulting optimization problem is given by

[13] An FPTAS is an algorithm that achieves $(1-\varepsilon)$ optimality, where for a given ε, the running time of the algorithm is polynomial in the problem instance (in this case n) as well as $1/\varepsilon$.

[14] This approach is commonly used to design FPTASs for knapsack problems in the computer science literature.

$$\max \quad y + \sum_{i\in[n]} x_i \log q_i \tag{1a}$$

$$\text{s.t.} \quad y \le \log(z), \tag{1b}$$

$$z = \sum_{i\in[n]} u_i x_i, \tag{1c}$$

$$1 \le \sum_{i\in[n]} x_i \le G, \tag{1d}$$

$$x_i \in \{0,1\}, \quad \forall i \in [n]. \tag{1e}$$

Moreover, Constraint 1b can be formulated as the conic constraint $(z, 1, y) \in K_{\text{exp}}$. Here K_{exp} is the exponential cone defined as

$$K_{\text{exp}} = \{(x_1, x_2, x_3) \mid x_1 \ge x_2 e^{x_3/x_2}, x_2 > 0\} \cup \{(x_1, 0, x_3) \mid x_1 \ge 0, x_3 \le 0\}.$$

We show in our numerical experiments that the resulting MICP can be solved rapidly with the conic solver MOSEK (ApS, 2023). Example running times can be found in Table 1 (Section 5.3) and Table 4 (Appendix C).

4.3 Optimal test allocation of any number of tests

When multiple tests are to be allocated, the optimal allocation problem no longer admits a conic formulation. Instead, we can formulate a mixed-integer convex program. Unfortunately, the time it takes to solve this program directly is prohibitive. Thus, we approximate the convex program with a mixed-integer linear program (MILP), resulting in a near-optimal non-overlapping solution. The MILP formulation approximates exponential constraints of the convex program with piecewise-linear functions that can be formulated as a collection of mixed integer linear constraints. The accuracy of this approximation can be adjusted by tuning the number K of segments of the piecewise-linear function, at the cost of introducing more (integer) variables and thus the time to solve the program. We provide practical (additive) approximation guarantees for solutions computed by the MILP as a function of parameter K. The MILP formulation is competitive when test budgets are low but too computationally intensive when budgets increase.

For the MILP program it can also be valuable to cluster identical individuals in the population. A formulation of the MILP with clusters, which may speed up computations, is provided in Appendix B.4.

4.3.1 Formulating a mixed-integer linear program

We can assume that the testing budget B is at most the population size n, and so pool sizes lie between 1 and G. We proceed similarly to our formulation of the conic program. For each test $j \in [B]$, we introduce an indicator vector $\boldsymbol{x}^j \in \{0,1\}^n$ with $x_i^j = 1$ if individual i is included in j and $x_i^j = 0$ otherwise, and let variable w^j denote its expected utility $w^j = u \cdot x^j \prod_{i\in[n]} q_i^{x_i^j}$. We impose pool sizes between 1 and G with constraints $1 \le \sum_{i\in[n]} x_i^j \le G$ for all $j \in [B]$, and non-overlapping testing with constraints $\sum_{j\in[B]} x_i^j \le 1$ for all $i \in [n]$. Our objective is to maximize welfare $\sum_{j\in[B]} w^j$. In order to isolate the non-linear elements of the optimization problem, we reformulate the problem with additional variables: variables l^j denote the logarithm of w^j, and variables y^j and z^j allow us to isolate the non-linear elements of each w^j into constraints (2b) and (2d).

$$\max \quad \sum_{j\in[B]} w^j \tag{2a}$$

$$\text{s.t.} \qquad w^j = \exp l^j, \qquad \forall j \in [B], \tag{2b}$$

$$l^j = y^j + \sum_{i \in [n]} x_i^j \log q_i, \quad \forall j \in [B], \tag{2c}$$

$$y^j = \log z^j, \qquad \forall j \in [B], \tag{2d}$$

$$z^j = u \cdot x^j, \qquad \forall j \in [B], \tag{2e}$$

$$\sum_{j \in [B]} x_i^j \leq 1, \qquad \forall i \in [n], \tag{2f}$$

$$1 \leq \sum_{i \in [n]} x_i^j \leq G, \qquad \forall j \in [B], \tag{2g}$$

$$x_i^j \in \{0, 1\}, \qquad \forall i \in [n], \forall j \in [B] \tag{2h}$$

In order to make the problem tractable, we assume that the utility vector u is integral and non-negative. This assumption is benign, as the problem is invariant to scaling of utilities. We describe in Appendix B.4 how the non-linear constraints (2b) and (2d) can respectively be captured approximately and exactly by integer linear constraints. The exponential constraints are approximated with piecewise-linear functions that can be formulated as a collection of mixed integer linear constraints. The accuracy of this approximation can be adjusted by tuning the number K of segments of the piecewise-linear function, at the cost of introducing more (integer) variables and thus the time to solve the program. As described in the appendix, this allows us to compute (additive) approximation guarantees for the MILP as a function of K.

5 Our algorithms in practice

On 9 June 2021, Mexico performed 0.07 COVID-19 tests per 100,000 inhabitants, whereas, e.g., the UK performed 13.48 tests per 100,000 people.[15] In light of this discrepancy, variable contagion rates and vaccine hesitancy, there was an urgent need for alternative strategies to manage the pandemic. While laboratories with qPCR testing capabilities were few and lateral flow tests prohibitively expensive, private and public institutions had some liberty in developing institutional health policies. Thus, in late 2021, we partnered with the Potosinian Institute for Scientific and Technological Research (IPICYT), a higher education and research institute in San Luis Potosí, Mexico. At the time, academics, administrative staff, and students performed all their tasks remotely and the IPICYT administration was eager to find a safe way to allow their staff and students to return to campus. Within this context, our utility-maximizing Greedy algorithm for pooled testing was accepted as a phase-in experiment. In preparation for this pilot study and during the development of our algorithmic framework, numerous consultations with the IPICYT administration, faculty, and lead scientists from their in-house LANBAMA laboratory informed our algorithms and their implementation. Some constraints, including the testing budget of $B = 30$ per week, a pool size of $G = 5$, and a preference for non-overlapping tests were set by the institution and the lab.[16]

Our goals as researchers for the implementation of our algorithm were two-fold: first, we wanted to capitalize on the opportunity to obtain real-world data that could provide valuable input for simulation experiments. Second, we wanted to evaluate the impact of our testing framework as a whole on IPICYT's population. In this section, we describe the transfer and deployment of (some) of our algorithms in practice and how we evaluated their performance and impact.

In Section 5.1, we outline the design of our algorithmic implementation and the pilot study. Section 5.2 details the crucial building block of probabilities and utilities as input to our algorithms.

[15] Numbers were retrieved from Our World in Data.

[16] LANBAMA had validated and implemented pooled testing with saliva samples for pool sizes up to 5.

In Section 5.3 we showcase our simulations on the performance of our algorithms based on real-world data input. Finally, Section 5.4 presents the impact evaluation of our GREEDY algorithm and testing framework at IPICYT.

5.1 Experimental design

We implemented our utility-based non-overlapping pooled testing regime based on GREEDY in a two-group randomized controlled trial at IPICYT in September 2022. A heterogeneous population of 130 individuals participated, including students, academics, and administrative staff.[17] Shortly before the trial commenced, IPICYT resumed full access for all its members. Having the academic community return to in-person activities allowed us to define a treatment and control group, each of them clustered by field and working group.[18] The treatment group followed a safety protocol, including scheduled testing and onsite work conditional on a negative result. The control group followed no protocol and returned to onsite activities. The control was in essence a 'first-best' benchmark, generating the outcomes one would expect without any COVID-19 restrictions and impositions from testing requirements - but without the guarantee of being in a infection-free environment. Our outcomes of interest were the staff and students' productivity, performance, learning, stress score, and subjective well-being.

All elements of the pilot study, including consent, a baseline and an endline survey, email invitations for testing, data processing, and computing test allocations, were coordinated in a web app developed specifically for the trial. A demo version is available at `https://demo.c-sef.com`.

At the beginning of each week, we computed an optimal test allocation among the treatment group for each day of the week, given a budget of 30 tests, and invited individuals to submit their saliva samples for testing at the LANBAMA facility. Among submitted samples, we determined once more the optimal test allocation on a daily basis. The treatment group was allowed access to university facilities after a negative test result, for at most 72 hours, and otherwise was required to work remotely. Participants were assigned to an experimental condition with peers from their working group, so that the majority of their social institutional interaction was contained in their experimental condition.[19]

5.2 Determining population data: utilities and health probabilities

The crucial input for our algorithms are the individuals' utilities and health probabilities.

Our rationale for developing a three-dimensional utility measure lies in documented effects of COVID-19 social-distancing policies. Firstly, an individual's need for in-person work or study depends on the nature of their work: e.g., an experimentalist in a lab must attend more frequently than a theoretician must visit their campus office. This specific need for in-person access was inferred from questions about use of digital media, which were designed, ordered and worded to make it difficult to judge how to answer a question in order to be prioritized for testing. Secondly, we considered evidence that the closure of learning environments disproportionately affects vulnerable individuals, e.g. with low-income.[20] We queried vulnerability with data on participants' self-perceived socio-economic status, a characteristic known to correlate with most

[17] At the end of the pilot, we collected between 118 and 122 complete data points, depending on the outcome of analysis. More on attrition in Appendix D.4.

[18] In practice, field and working group were analogous, as only one working group from each participating field volunteered to be a part of the experiment.

[19] If non-treated participants had encountered treated participants, possible contagions would be contained within our health protocol due to the 72-hour non-contagious access window. Although not strictly enforced, participants were encouraged not to socialize with non-treatment participants as to avoid psychological spillovers.

[20] See e.g. (Azevedo et al., 2021; Gorgen and McAleavy, 2020; Goudeau et al., 2021; Hossain, 2021). These studies of heterogeneous effects on vulnerable populations are primarily conducted with students ranging from K-12 to Higher Education; however, they also document similar issues for teaching staff and generally academic environments.

forms of protected vulnerability attributes in Mexico.[21] Thirdly, mental health is known to be negatively affected by pandemic-induced remote learning and work for younger and and older individuals,[22] and students and employees in Mexico have struggled with mental health problems associated with COVID-19 institutional closures[23]. Mental health status was queried with standard, validated survey questions (cf. Appendix D.3).

We denote by u_i^{pr} the utility subject i gains from increased productivity, by u_i^{psy} the benefit on i's mental health from attending in-person, and by u_i^{se} a bonus for socio-economically disadvantaged individuals, who are likely to be more affected by working remotely. The overall utility is a weighted sum $u_i = \sum_{k\in\{pr,psy,se\}} w^k u_i^k$.[24] We define the composition of u_i^k for category $k \in \{pr, psy, se\}$. Let $P_{i,z}^k$ denote the number of points achieved by the answer of subject i to question z, and Z^k the number of questions relevant in category k.[25] For each category, the score is $u_i^k = \frac{1}{Z^k}\sum_z P_{i,z}^k$.

Health probabilities were estimated for age and gender categories using Bayesian updates of local public health data. We computed the probability of being infected conditional on being in one of the following 6 groups: {male, female} × {age 15-29, 30-59, ≥60}. The baseline probability of infection for a given age group is determined using Bayesian updates of local public health data, under the guidance of local epidemiologists. More specifically, we used publicly available epidemiological models from the Institute of Health Metrics and Evaluation (IHME) to estimate baseline infection rates in San Luis Potosí.[26] These estimates provided us with values for Pr[infection] for all individuals in the population, irrespective of their category. Furthermore, we estimated the probability that an individual belongs to a given category given an infection via official national data on testing results.[27] These estimates provide us with values for Pr[category | infected] for each category. Finally, we used census data to compute the probability of membership to a given category at the state/national level.[28] This provides us with an estimate for Pr[category] for the population. With Bayes' rule, we compute the desired probability of infection per category as follows: $\Pr[\text{infection} \mid \text{category}] = \frac{\Pr[\text{category}|\text{infection}]\Pr[\text{infection}]}{\Pr[\text{category}]}$. The probabilities of being healthy were approximately 99.5% for each group. The probabilities stayed constant throughout this trial, as it only ran for 4 weeks. If applied over a longer period, the health probabilities may also be updated.[29]

5.3 Performance evaluation of Greedy and the MILP

We evaluate the accuracy and running times of the Greedy algorithm and the MILP on populations reflecting real-world scenarios. In our first numerical experiment, we test both algorithms on population data from the pilot study with testing budgets B up to 34 and pool size constraints G of 5 and 10.[30] These simulations illustrate the efficacy of Greedy with low

[21] Low income is highly correlated with belonging to an ethnic minority, the elderly, or being female (Ordóñez Barba, 2018).

[22] E.g. (Asanov et al., 2021; Bertoni et al., 2022)

[23] See e.g. (Limón-Vázquez et al., 2020; Martinez Arriaga et al., 2021)

[24] The weights were $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ in our trial. We refrained from any value judgment on importance of category.

[25] Relevance of a question to a specific category is marked in the survey in the Online Appendix (Section 4) by the corresponding abbreviation just after the question numbering. Not all questions are relevant for the construction of utilities.

[26] Estimated infection rates for SLP with IHME models can be found at their dashboard for different public behavior regimes: https://covid19.healthdata.org/mexico/san-luis-potosi?view=infections-testing&tab=trend&test=infections.

[27] National testing aggregates can be found at https://datos.covid-19.conacyt.mx and https://covid19.healthdata.org/mexico/san-luis-potosi?view=infections-testing&tab=trend&test=infections.

[28] Census data can be found at https://www.inegi.org.mx/programas/ccpv/2020/.

[29] If the random element of time spent offsite can be controlled for, Bayesian updating using test results may be preferred.

[30] As the MILP is designed to admit integral utilities only, and the problem of computing test allocations is invariant to scaling utilities, we first scale up the utilities of all individuals in the population by a factor of 50, and

disease incidence. In order to study how well Greedy performs when faced with higher infection rates, we also ran simulations on synthetic data in which health probabilities range from 0.5 to 1, and explore test budgets B up to 12. Moreover, we also study outcomes when pool sizes increase to $G = 10$ (a typical pool size for nasopharyngeal samples).

In our synthetic experiments, we showcase the average-case behavior of Greedy and the MILP by generating random populations of size $n = 150$. Health probabilities are drawn independently and uniformly at random from the interval $[0.5, 1]$, and utilities are drawn from a normal distribution that was fitted to the utilities observed in the pilot study. We then run Greedy and the MILP on each population for both pool sizes $G \in \{5, 10\}$ and for all testing budgets $B \in \{2, 4, \ldots, 12\}$, recording the welfare achieved for both algorithms, as well as their running times (in milliseconds).[31]

For all experiments, we document the true welfare achieved by the test allocations returned by both algorithms, and not the objective values of the underlying MILP and conic optimization problems. For the MILP algorithm, we tune the parameter K of the formulation so that the additive approximation guarantee (cf. Appendix B.4) is small ($K = 25$ for the simulations on pilot data, and $K = 20$ for the simulations on synthetic data). The code used to run our simulations can be found at `https://github.com/edwinlock/pooled-testing.git`.

Results. Table 1 lists the welfares achieved by Greedy and the MILP on the pilot data for pool size $G = 5$, as well as the running times for both algorithms, the approximation guarantee achieved by the MILP, and an upper bound on how much better the MILP does compared to Greedy in terms of a welfare ratio. This upper bound is determined by summing the welfare and the approximation guarantee of the MILP to upper bound the optimum achievable welfare, and then dividing by the welfare achieved by Greedy. Table 3 in Appendix C shows analogous results for pool size constraint $G = 10$. We observe that Greedy achieves near-optimal welfare for budgets up to 10 for both pool sizes. Moreover, the running time of the MILP increases significantly faster with the test budget B than Greedy, and the latter is extremely fast (even for larger populations and testing budgets). This makes Greedy attractive for implementations that rely on a quick turnaround, run on 'budget hardware' or wish to avoid costly cloud computing services.

In our synthetic simulations with lower health probabilities, Greedy performs as well as the MILP when $G = 5$, and remains highly competitive also when $G = 10$. Figures 1 and 2 in Appendix C plot the mean welfare achieved by both algorithms for pool size constraints $G = 5$ and $G = 10$, as well as the welfare ratios. For the latter, we divide the welfare achieved by the MILP by the welfare of Greedy for each population, and depict the resulting ratios as black dots. In Appendix C, Tables 4 and 5 list the mean welfare and running times of both algorithms, as well as the approximation guarantee of the MILP, for $G \in \{5, 10\}$.

Comparing the outcomes between different pool sizes $G \in \{5, 10\}$, we see that increasing pool sizes from 5 to 10 significantly increases mean welfare if health probabilities are very high (cf. Tables 1 and 3). This effect is less pronounced in the experiment with synthetic data, in which participants have lower health probabilities on average (cf. Tables 4 and 5). These results suggest that the pool size limit of 5 imposed by saliva sampling, as opposed to the typical limit of 10 for nasopharyngeal samples, may be considered a limitation in some scenarios, and institutions may wish to weigh the positives and negatives of saliva and nasopharyngeal sampling carefully.

then round the resulting number to the nearest integer. Choosing a larger scaling factor increases the running time, as the number of variables in the MILP increases (cf. Appendix B.4).

[31] The experiments were run on an AWS EC2 instance type 'c6g.8xlarge' with 32 vCPUs and 64GiB memory. Gurobi 9.5.0 was used for the MILP algorithm, and MOSEK 10 for the MICP.

	MILP			Greedy		
Budget	Welfare	Guarantee	Time	Welfare	Apx To Optimal	Time
2	461.24	0.18	292 ms	461.22	1.000028	49 ms
6	1292.04	0.53	1911 ms	1291.91	1.000097	48 ms
10	2070.82	0.89	5345 ms	2070.58	1.000115	71 ms
14	2814.82	1.25	2550656 ms	2814.50	1.000115	400 ms
18	3524.78	1.61	1523318 ms	3524.24	1.000154	2718 ms
22	4189.57	1.97	569366 ms	4188.85	1.000170	42869 ms
26	4790.46	2.32	12160741ms	4789.37	1.000227	747631 ms
30	4805.24	2.68	69541821 ms	4789.37	1.003313	750206 ms
34	4816.37	3.04	408803099 ms	4789.37	1.005637	757912 ms

Table 1: Summary showing welfare and computation time for the MILP and GREEDY on the pilot data with a population of $n = 130$ and pool size constraint $G = 5$, with testing budgets $B \in \{2, 6, \ldots, 34\}$. We also state the additive approximation guarantee of MILP (compared to optimal non-overlapping welfare).

5.4 Pilot Study

In this section, we report the outcomes of the implementation of our test allocation framework in the pilot at IPICYT.

Our algorithm in practice. The GREEDY algorithm, as described in Section 4.1, demonstrated favorable trade-offs between speed and accuracy (cf. Section 5.3). For that reason, we implemented a version of GREEDY in our web application for computing test allocations. For the purpose of the trial, we allowed individuals to express onsite work preferences for two-day windows through the allocation of a virtual token budget in the web app. This helped us avoid scheduling individuals for testing on days they did not wish to access IPICYT facilities in the first place, and allocate more tests to particular days that were more popular. Moreover, our partner institute observed – in an independent pool testing trial – that a small fraction of participants invited to submit a saliva sample for testing fail to do so. In order to optimize pooling in this setting, we perform a second optimization round, in which we compute an (approximately) optimal pooling among the samples that have been submitted. It is immediate that the second optimization round cannot decrease the expected welfare achieved.

Evaluation and methods. In the trial we measured subjects' stress levels and subjective well-being (life satisfaction), as well as self-assessed performance, productivity, and learning. We obtain these measures through survey questions that subjects are invited to answer before (baseline) and after (endline) the trial period. A detailed description of these variables is given in Appendix D.3. The treatment effect is estimated with bivariate linear regressions[32], using the above-mentioned outcomes as dependent variables. The main regressor is a binary treatment variable, which takes on the value one if the subject is in the treatment group and zero otherwise. We estimate level effects on endline outcomes as well as the effect on the difference in outcomes (delta models) between our two points of measurement before and after the trial. We further collected a number of covariates for robustness checks of our estimations.

[32]Note that we also include equivalence tests in Table 6, and multinomial logistic regression models for non-normally distributed outcomes discussed in Appendix D.6 and the Online Appendix (Section 3, Tables 5 and 6). These robustness checks corroborate our results from our preferred model specification.

Results. We present the results on performance from the linear models in Table 2. Further results on performance and mental health are shown in Tables 8 and 9 in Appendix D.6. For the levels models, treatment group participants exhibit, on average, higher scores in self-perceived performance, productivity, and learning.[33] These group differences are not statistically significant. Similarly, participants in the treatment group report, on average, higher levels of stress and higher levels of subjective well-being (life satisfaction). These differences are not statistically significant either. Despite following different protocols, participants in both experimental groups react similarly to partially or fully lifting lockdown restrictions.

Importantly, throughout the month of the trial, only one pooled test returned positive. As per the protocol, treatment participants in the positive pool were asked to work remotely until such a time where they were again scheduled to submit a sample. The LANBAMA lab ran individual tests to identify the infected individual(s), to make sure that regardless of a show or no show of symptoms, the infected individual(s) had certainty of their status and reacted accordingly.

	Dependent variable					
	Performance	Productivity	Learning	Δ Performance	Δ Productivity	Δ Learning
Treatment	0.120	0.076	0.175	−0.053	−0.256·	0.086
	(0.143)	(0.133)	(0.287)	(0.143)	(0.130)	(0.323)
Constant	1.984***	2.097***	8.194***	0.000	0.081	0.177
	(0.086)	(0.082)	(0.206)	(0.110)	(0.090)	(0.209)
Observations	119	120	119	118	119	119
R^2	0.006	0.003	0.003	0.001	0.032	0.001
Adjusted R^2	−0.002	−0.006	−0.005	−0.007	0.024	−0.008

Sig. p codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '·' 0.1 ' '

Table 2: Linear regression models of performance, productivity, and learning outcomes. Note that regression coefficients are expressed in the unit of the score. HC1 standard errors are in parentheses.

We also estimate delta models, or first-differences in scores (between endline and baseline), to correct for probable time-dependent confounders. The positive trend in increased stress in the treatment group disappears, while the trend in life satisfaction increases. All treatment effects related to performance, productivity, and learning are corrected downwards, but remain statistically insignificant, with the exception of productivity, where we report a small and borderline statistically significant negative.[34] This may stem from treatment group individuals having to exert additional effort to schedule their week according to their sample submission date, and actually submitting their sample at the testing facilities. These extra tasks can result in a loss in productivity as they take up working time and constitute a higher cognitive load when mental bandwidth is limited. Individuals in positive pools who are required to work from home may also face productivity constraints, but during our trial only one pooled test returned positive and, within the pool, only one individual was identified as infected.

In summary, we find no statistical evidence that our test allocation strategy has a negative effect on participants' work/study performance, learning, or mental health, despite the increased effort in coordination it demands from them compared to a full reopening (the protocol followed by the control group). At the same time, our strategy ensures greater safety for all participating individuals compared to a full reopening without any safety mechanisms in place. We conjecture that accounting for welfare is the crucial ingredient in our mechanism, enabling in-person access for those who need and benefit from it the most.

[33]This also holds for the measures of achieving their own and their supervisors' goals, see Appendix D.6.

[34]The estimated p-value is 0.051, exactly on the cutoff of statistical significance. We consider statistical significance for all values $p < 0.05$, but not for values on or above that cutoff (Zhu, 2016).

6 Discussion

This work introduces a novel utility-based approach to pooled testing in resource-constrained environments. In a setting where the population exhibits heterogeneity in health probabilities *and* utilities, we provide strong theoretical and empirical performance guarantees of our near-optimal test allocation mechanism, which justify the implementation of non-overlapping testing regimes beyond their essential logistical simplicity. We test a version of our test allocation procedure in a real-world experiment at the higher education research institute IPICYT, in Mexico. We use the trial data to evaluate two algorithms, Greedy and a MILP formulation, through simulations. Our simulations with real-world and synthetic data demonstrate that Greedy performs almost optimally and is significantly faster than our alternative MILP implementation. Our randomized controlled trial also provides evidence that our test allocation performs no worse than a 'first-best' benchmark of allowing full institutional access for all individuals, where performance is measured with respect to participants' work and study performance and productivity, as well as subjective well-being and mental health (proxied by stress levels).

Our pooled-testing protocol ensures that everyone released for onsite work is guaranteed to be non-infectious. Thus, it provides a safe work environment during high, variable, and low infection rate periods, while imposing no penalties on well-being and productivity in periods with low infection rates. Resource-saving alternatives such as self-testing, by contrast, cannot provide the same level of safety and may expose a community to asymptomatic but contagious carriers. Our testing regime provides institutions with an economical and safe solution to keep their entire community safe.

There are many directions for future work. On a theoretical level, there is a gap between our upper bound of 4 and lower bound of 7/6 on the overlap welfare ratio, and an upper bound of 5 on the approximation factor of the Greedy algorithm. Beyond the tight approximation bound of e for Greedy that we establish for the case with identical utilities for individuals, we expect that tighter bounds are achievable when utilities take a fixed number of values (e.g. for dichotomous or trichotomous populations). On a more practical level, the overall testing and re-integration policy we propose is static in nature, as we consider the one-shot setting where a testing budget is to be fully utilized by a policymaker. Testing could also be dynamic, with allocations chosen adaptively as a function of previous test results, and it is valuable to understand what potential benefits this extended functionality may bring. Additionally, policymakers potentially have access to different types of tests, each with different associated costs and performance (i.e., pool size and sensitivity), and providing optimal budget-constrained allocations in this heterogeneous test setting is a key open question.

Most importantly, we hope that the insights in performance and efficacy of our welfare-maximizing test allocations can help better protect resource-constrained communities when confronted with an outbreak of an infectious disease. Our pooled testing framework may find further, equally important applications in mass screening for HIV/AIDS and pooled frameworks for organ donation.

References

A. Abera, H. Belay, A. Zewude, B. Gidey, D. Nega, B. Dufera, A. Abebe, T. Endriyas, B. Getachew, H. Birhanu, et al. Establishment of COVID-19 testing laboratory in resource-limited settings: challenges and prospects reported from ethiopia. *Global Health Action*, 13(1):1841963, 2020.

P. D. Allison. Change scores as dependent variables in regression analysis. *Sociological Methodology*, 20:93–114, 1990. ISSN 00811750, 14679531. URL http://www.jstor.org/stable/271083.

D. G. Altman. Comparability of randomised groups. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 34(1):125–136, 1985.

M. ApS. *MOSEK Optimization Suite 10.0*, 2023. URL https://www.mosek.com.

I. Asanov, F. Flores, D. McKenzie, M. Mensmann, and M. Schulte. Remote-learning, time-use, and mental health of ecuadorian high-school students during the covid-19 quarantine. *World development*, 138:105225, 2021.

N. Augenblick, J. T. Kolstad, Z. Obermeyer, and A. Wang. Group testing in a pandemic: The role of frequent testing, correlated risk, and machine learning. Working Paper 27457, National Bureau of Economic Research, July 2020. URL http://www.nber.org/papers/w27457.

J. P. Azevedo, A. Hasan, D. Goldemberg, K. Geven, and S. A. Iqbal. Simulating the potential impacts of covid-19 school closures on schooling and learning outcomes: A set of global estimates. *The World Bank Research Observer*, 36(1):1–40, 2021.

L. Becchetti, L. Corrado, and P. Conzo. Sociability, altruism and well-being. *Cambridge Journal of Economics*, 41(2):441–486, 2017.

M. Bertoni, D. Cavapozzi, G. Pasini, and C. Pavese. Remote working and mental health during the first wave of the covid-19 pandemic. Available at SSRN: https://ssrn.com/abstract=4111999 or http://dx.doi.org/10.2139/ssrn.4111999, 2022.

N. Bobkova, Y. Chen, and H. Eraslan. Optimal group testing with heterogeneous risks. *Economic Theory*, 2023. doi: https://doi.org/10.1007/s00199-023-01502-3.

V. Brault, B. Mallein, and J.-F. Rupprecht. Group testing as a strategy for COVID-19 epidemiological monitoring and community surveillance. *PLoS computational biology*, 17(3):e1008726, 2021.

M. Bruhn and D. McKenzie. In pursuit of balance: Randomization in practice in development field experiments. *American economic journal: applied economics*, 1(4):200–232, 2009.

G. Camera and A. Gioffré. The economic impact of lockdowns: A theoretical assessment. *Journal of Mathematical Economics*, 97:102552, 2021.

B. Cleary, J. A. Hay, B. Blumenstiel, M. Harden, M. Cipicchio, J. Bezney, B. Simonton, D. Hong, M. Senghore, A. K. Sesay, et al. Using viral load and epidemic dynamics to optimize pooled testing in resource-constrained settings. *Science translational medicine*, 13(589):eabf1568, 2021.

S. Cohen, T. Kamarck, R. Mermelstein, et al. Perceived stress scale. *Measuring stress: A guide for health and social scientists*, 10(2):1–2, 1994.

P. Deb, D. Furceri, J. D. Ostry, and N. Tawk. The economic effects of covid-19 containment measures. *Open Economies Review*, 33(1):1–32, 2022.

G. N. Dhabaan, W. A. Al-Soneidar, and N. N. Al-Hebshi. Challenges to testing COVID-19 in conflict zones: Yemen as an example. *Journal of global health*, 10(1), 2020.

A. Donner. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review / Revue Internationale de Statistique*, 54(1):67–82, 1986. ISSN 03067734, 17515823. URL http://www.jstor.org/stable/1403259.

R. Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.

J. Ely, A. Galeotti, O. Jann, and J. Steiner. Optimal test allocation. *Journal of Economic Theory*, 193:105236, 2021.

J. Emmanuel, M. Bassett, H. Smith, and J. Jacobs. Pooling of sera for human immunodeficiency virus (hiv) testing: an economical method for use in developing countries. *Journal of clinical pathology*, 41(5):582–585, 1988.

N. Goldberg and G. Rudolf. On the complexity and approximation of the maximum expected value all-or-nothing subset. *Discrete Applied Mathematics*, 283:1–10, 2020.

C. Gollier and O. Gossner. Group testing against Covid-19. Technical report, EconPol Policy Brief, 2020.

K. Gorgen and T. McAleavy. Best practice in pedagogy for remote teaching (p. 10). *EdTech Hub Report*, 2020.

S. Goudeau, C. Sanrey, A. Stanczak, A. Manstead, and C. Darnon. Why lockdown and distance learning during the covid-19 pandemic are likely to increase the social class achievement gap. *Nature Human Behaviour*, 5(10):1273–1281, 2021.

M. Hossain. Unequal experience of covid-induced remote schooling in four developing countries. *International Journal of Educational Development*, 85:102446, 2021.

M. M. Kavanagh, N. A. Erondu, O. Tomori, V. J. Dzau, E. A. Okiro, A. Maleche, I. C. Aniebo, U. Rugege, C. B. Holmes, and L. O. Gostin. Access to lifesaving medical resources for african countries: COVID-19 testing and response, ethics, and politics. *The Lancet*, 395(10238): 1735–1738, 2020.

M. M. King and M. E. Frederickson. The pandemic penalty: The gendered effects of covid-19 on scientific productivity. *Socius*, 7:23780231211006977, 2021.

T. K. Koo and M. Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

D. Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4):355–362, 2017.

D. Lakens. The practical alternative to the p value is the correctly used p value. *Perspectives on psychological science*, 16(3):639–648, 2021.

D. B. Larremore, B. Wilder, E. Lester, S. Shehata, J. M. Burke, J. A. Hay, M. Tambe, M. J. Mina, and R. Parker. Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. *Science advances*, 7(1):eabd5393, 2021.

A. K. Limón-Vázquez, G. Guillén-Ruiz, and E. V. Herrera-Huerta. The social isolation triggered by covid-19: Effects on mental health and education in mexico. *Health and Academic Achievement-New Findings*, 2020.

E. Lipnowski and D. Ravid. Pooled testing for quarantine decisions. *Journal of Economic Theory*, 198:105372, 2021. ISSN 0022-0531. doi: https://doi.org/10.1016/j.jet.2021.105372. URL https://www.sciencedirect.com/science/article/pii/S0022053121001897.

S. L. Lohr. *Sampling: design and analysis*. CRC press, 2021.

R. J. Martinez Arriaga, L. P. Gonzalez Ramirez, J. M. de la Roca-Chiapas, and M. Hernández-González. Psychological distress of covid-19 pandemic and associated psychosocial factors among mexican students: An exploratory study. *Psychology in the Schools*, 58(9):1844–1857, 2021.

L. Mutesa, P. Ndishimye, Y. Butera, J. Souopgui, A. Uwineza, R. Rutayisire, E. L. Ndoricimpaye, E. Musoni, N. Rujeni, T. Nyatanyi, et al. A pooled testing strategy for identifying sars-cov-2 at low prevalence. *Nature*, 589(7841):276–280, 2021.

O. U. Nalbantoglu. Group testing performance evaluation for sars-cov-2 massive scale screening and testing. *BMC Medical Research Methodology*, 20(1):1–11, 2020.

OECD. *OECD guidelines on measuring subjective well-being*. OECD, 2013.

G. Ordóñez Barba. Discrimination, poverty and vulnerability: the intricacies of social inequality in mexico. *Región y sociedad*, 30(71):0–0, 2018.

H. R. Sanghani, D. A. Nawrot, F. Marmolejo-Cossío, J. M. Taylor, J. Craft, E. Kalimeris, M. I. Andersson, and S. R. Vasudevan. Concentrating Pooled COVID-19 Patient Lysates to Improve Reverse Transcription Quantitative PCR Sensitivity and Efficiency. *Clinical Chemistry*, 67(5): 797–798, 2021. ISSN 0009-9147. doi: 10.1093/clinchem/hvab035.

B. Shan, X. Liu, A. Gu, and R. Zhao. The effect of occupational health risk perception on job satisfaction. *International Journal of Environmental Research and Public Health*, 19(4):2111, 2022.

N. Shental, S. Levy, V. Wuvshet, S. Skorniakov, B. Shalem, A. Ottolenghi, Y. Greenshpan, R. Steinberg, A. Edri, R. Gillis, M. Goldhirsh, K. Moscovici, S. Sachren, L. M. Friedman, L. Nesher, Y. Shemer-Avni, A. Porgador, and T. Hertz. Efficient high-throughput sars-cov-2 testing to detect asymptomatic carriers. *Science Advances*, 6(37):eabc5961, 2020. doi: 10.1126/sciadv.abc5961. URL https://www.science.org/doi/abs/10.1126/sciadv.abc5961.

L. N. Theagarajan. Group testing for covid-19: How to stop worrying and test more, 2020.

X. M. Tu, E. Litvak, and M. Pagano. On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application to hiv screening. *Biometrika*, 82(2):287–297, 1995.

M. Versteeg and P. Steendijk. Putting post-decision wagering to the test: a measure of self-perceived knowledge in basic sciences? *Perspectives on Medical Education*, 8(1):9–16, 2019.

L. M. Wein and S. A. Zenios. Pooled testing for hiv screening: capturing the dilution effect. *Operations Research*, 44(4):543–569, 1996.

I. Yelin, N. Aharony, E. S. Tamar, A. Argoetti, E. Messer, D. Berenbaum, E. Shafran, A. Kuzli, N. Gandali, O. Shkedi, T. Hashimshony, Y. Mandel-Gutfreund, M. Halberthal, Y. Geffen, M. Szwarcwort-Cohen, and R. Kishony. Evaluation of COVID-19 RT-qPCR Test in Multi sample Pools. *Clinical Infectious Diseases*, 71(16):2073–2078, 05 2020. ISSN 1058-4838. doi: 10.1093/cid/ciaa531. URL https://doi.org/10.1093/cid/ciaa531.

W. Zhu. p< 0.05,< 0.01,< 0.001,< 0.0001,< 0.00001,< 0.000001, or< 0.0000001.... *Journal of sport and health science*, 5(1):77, 2016.

A Proofs for Section 3 (Performance of Non-Overlapping Testing)

A.1 Proof of Proposition 3.3

Fix some population J, and suppose $T^* = (t_1^*, t_2^*)$ is an optimal two-test allocation for population J in the overlapping testing regime. We assume that $t_1^* \cap t_2^* \neq \emptyset$ (otherwise the overlap welfare ratio is 1, and we are done) and let $A = t_1^* \setminus t_2^*$, $B = t_2^* \setminus t_1^*$ and $C = t_1^* \cap t_2^*$. Note that $t_1^* \cup t_2^* = A \cup B \cup C$.

Without loss of generality, we assume that $q_A \geq q_B$ and denote $u_A = \sum_{i \in A} u_i$, $u_B = \sum_{i \in B} u_i$, and $u_C = \sum_{i \in C} u_i$. In order to prove Proposition 3.3, we consider the four non-overlapping test allocations $T^1 = (A \cup C, B)$, $T^2 = (A, C)$, $T^3 = (B, C)$ and $T^4 = (A \cup B, C)$, and make use of the following lemma.

Lemma A.1. *For any $T \in \{T^1, T^2, T^3, T^4\}$, the ratio $\frac{u(T^*)}{u(T)}$ is maximized when $q_C = 1$.*

Proof. Consider $T = T^1$. We need to show that

$$\begin{aligned}\frac{u(T^*)}{u(T^1)} &= \frac{q_C\left(q_A \cdot u_A + q_B \cdot u_B + (q_A + (1-q_A)\cdot q_B)\cdot u_C\right)}{q_C \cdot q_A \cdot u_A + q_B \cdot u_B + q_C \cdot q_A \cdot u_C} \\ &\leq \frac{q_A \cdot u_A + q_B \cdot u_B + (q_A + (1-q_A)\cdot q_B)\cdot u_C}{q_A \cdot u_A + q_B \cdot u_B + q_A \cdot u_C},\end{aligned}$$

which is equivalent to

$$q_C \leq \frac{q_C \cdot q_A \cdot u_A + q_B \cdot u_B + q_C \cdot q_A \cdot u_C}{q_A \cdot u_A + q_B \cdot u_B + q_A \cdot u_C}.$$

But this is true because

$$q_C = q_C \cdot \frac{q_A \cdot u_A + q_B \cdot u_B + q_A \cdot u_C}{q_A \cdot u_A + q_B \cdot u_B + q_A \cdot u_C} \leq \frac{q_C \cdot q_A \cdot u_A + q_B \cdot u_B + q_C \cdot q_A \cdot u_C}{q_A \cdot u_A + q_B \cdot u_B + q_A \cdot u_C}.$$

Next, consider the case that $T = T^2$. Here, we need to show that

$$\begin{aligned}\frac{u(T^*)}{u(T^2)} &= \frac{q_C\left(q_A \cdot u_A + q_B \cdot u_B + (q_A + (1-q_A)\cdot q_B)\cdot u_C\right)}{q_A \cdot u_A + q_C \cdot u_C} \\ &\leq \frac{q_A \cdot u_A + q_B \cdot u_B + (q_A + (1-q_A)\cdot q_B)\cdot u_C}{q_A \cdot u_A + u_C} \\ \Leftrightarrow q_C &\leq \frac{q_A \cdot u_A + q_C \cdot u_C}{q_A \cdot u_A + u_C}.\end{aligned}$$

This is true because

$$q_C = q_C \cdot \frac{q_A \cdot u_A + u_C}{q_A \cdot u_A + u_C} \leq \frac{q_A \cdot u_A + q_C \cdot u_C}{q_A \cdot u_A + u_C}.$$

Analogously, the ratio $\frac{u(T^*)}{u(T^3)}$ is maximized when $q_C = 1$. Lastly, we consider the case $T = T^4$.

$$\begin{aligned}\frac{u(T^*)}{u(T^4)} &= \frac{q_C\left(q_A \cdot u_A + q_B \cdot u_B + (q_A + (1-q_A)\cdot q_B)\cdot u_C\right)}{q_A \cdot q_B \cdot (u_A + u_B) + q_C \cdot u_C} \\ &\leq \frac{q_A \cdot u_A + q_B \cdot u_B + (q_A + (1-q_A)\cdot q_B)\cdot u_C}{q_A \cdot q_B \cdot (u_A + u_B) + u_C} \\ \Leftrightarrow q_C &\leq \frac{q_A \cdot q_B \cdot (u_A + u_B) + q_C \cdot u_C}{q_A \cdot q_B \cdot (u_A + u_B) + u_C}.\end{aligned}$$

This holds because

$$q_C = q_C \cdot \frac{q_A \cdot q_B \cdot (u_A + u_B) + u_C}{q_A \cdot q_B \cdot (u_A + u_B) + u_C} \leq \frac{q_A \cdot q_B \cdot (u_A + u_B) + q_C \cdot u_C}{q_A \cdot q_B \cdot (u_A + u_B) + u_C}.$$

□

We are now ready to complete the proof of Proposition 3.3.

Proof of Proposition 3.3. Recall the optimal overlapping two-test allocation T^*, its sub-populations A, B and C, and the four non-overlapping allocations T_1, T_2, T_3 and T_4 from above. Our goal is to show that $\frac{u(T^*)}{u(T)} \leq \frac{7}{6}$ at least one $T = T_i$. Without loss of generality, we can assume that $q_C = 1$ by Lemma A.1. Suppose first that $q_A \geq 5/6$. Note that

$$\frac{(q_A + (1-q_A)\cdot q_B)\cdot u_C}{q_A \cdot u_C} \leq \frac{(q_A + (1-q_A)\cdot q_A)\cdot u_C}{q_A \cdot u_C} = 2 - q_A \leq \frac{7}{6},$$

where the first inequality holds since we assume $q_A \geq q_B$ and the last inequality follows from $q_A \geq 5/6$. Hence, we see that

$$\frac{u(T^*)}{u(T^1)} = \frac{q_A \cdot u_A + q_B \cdot u_B + (q_A + (1-q_A)\cdot q_A)\cdot u_C}{q_A \cdot u_A + q_B \cdot u_B + q_A \cdot u_C} \leq \frac{7}{6}.$$

Now assume that $q_A < 5/6$. For the sake of contradiction, suppose that $\frac{u(T^*)}{u(T)} > \frac{7}{6}$ for all $T \in \{T^1, T^2, T^3, T^4\}$. Thus

$$\frac{u(T^*)}{u(T^1)} = \frac{q_A \cdot u_A + q_B \cdot u_B + (q_A + (1-q_A)\cdot q_B)\cdot u_C}{q_A \cdot u_A + q_B \cdot u_B + q_A \cdot u_C} > \frac{7}{6} \tag{3}$$

implies

$$(6\cdot(q_A + (1-q_A)\cdot q_B) - 7q_A)\cdot u_C > q_A \cdot u_A + q_B \cdot u_B. \tag{4}$$

Similarly,

$$\frac{u(T^*)}{u(T^2)} = \frac{q_A \cdot u_A + q_B \cdot u_B + (q_A + (1-q_A)\cdot q_B)\cdot u_C}{q_A \cdot u_A + u_C} > \frac{7}{6}$$

implies

$$q_B \cdot u_B > \frac{1}{6} q_A \cdot u_A + \left(\frac{7}{6} - (q_A + (1-q_A)\cdot q_A)\right)\cdot u_C, \tag{5}$$

where we used the assumption that $q_A \geq q_B$. Analogously, $\frac{u(T^*)}{u(T^3)} > \frac{7}{6}$ implies

$$q_A \cdot u_A > \frac{1}{6} q_B \cdot u_B + \left(\frac{7}{6} - (q_A + (1-q_A)\cdot q_A)\right)\cdot u_C. \tag{6}$$

Equations (5) and (6) together imply

$$q_A \cdot u_A + q_B \cdot u_B > \frac{6}{5}\cdot 2 \cdot \left(\frac{7}{6} - (q_A + (1-q_A)\cdot q_A)\right)\cdot u_C, \tag{7}$$

and from Equation (4) we conclude that

$$(6\cdot(q_A + (1-q_A)\cdot q_B) - 7q_A) > \frac{6}{5}\cdot 2 \cdot \left(\frac{7}{6} - (q_A + (1-q_A)\cdot q_A)\right), \tag{8}$$

which is true when $\frac{1}{2} < q_A < \frac{2}{3}$. Hence, from now on we assume that $\frac{1}{2} < q_A < \frac{2}{3}$.

Equations (3) and (4) together imply

$$\frac{(6\cdot(q_A + (1-q_A)\cdot q_B) - 7q_A)\cdot u_C + (q_A + (1-q_A)\cdot q_B)\cdot u_C}{q_A \cdot u_A + u_C} > \frac{7}{6},$$

and so

$$(6\cdot(1-q_A)\cdot q_B - 1)\cdot u_C > q_A u_A. \tag{9}$$

Analogously, $\frac{u(T^*)}{u(T^3)} > \frac{7}{6}$ implies

$$(6\cdot(1-q_A)\cdot q_B - 1)\cdot u_C > q_B \cdot u_B \tag{10}$$

Lastly,

$$\frac{u(T^*)}{u(T^4)} = \frac{q_A \cdot u_A + q_B \cdot u_B + (q_A + (1-q_A)\cdot q_B)\cdot u_C}{q_A \cdot q_B \cdot (u_A + u_B) + u_C} > \frac{7}{6}$$

implies

$$q_A \cdot u_A + q_B \cdot u_B > \frac{(7 - 6(q_A + (1-q_A)\cdot q_B))\cdot u_C}{6 - 7\cdot q_B}, \tag{11}$$

where the last inequality follows from the fact that $q_B \le q_A < \frac{5}{6} < \frac{6}{7}$ by assumption. Equations (9) to (11) together tell us that

$$\begin{aligned}\frac{u(T^*)}{u(T^1)} &= \frac{q_A \cdot u_A + q_B \cdot u_B + (q_A + (1-q_A)\cdot q_B)\cdot u_C}{q_A \cdot u_A + q_B \cdot u_B + q_A \cdot u_C} \\ &\le \frac{2\cdot(6\cdot(1-q_A)\cdot q_B - 1)\cdot u_C + (q_A + (1-q_A)\cdot q_B)\cdot u_C}{\frac{7-6(q_A+(1-q_A)\cdot q_B)\cdot u_C}{6-7\cdot q_B} + q_A \cdot u_C}\end{aligned}$$

which is maximized when $q_A = q_B = 1/2$, given that $1/2 < q_A < 2/3$. But this implies that $\frac{u(T^*)}{u(T^1)} \le \frac{7}{6}$, a contradiction. □

Proof of Proposition 3.4. Consider first the case $B = 3$. Suppose T^* is an optimal overlapping test allocation for some population. We partition the individuals that are tested under T^* into seven sets as following. The first three sets, denoted S_1, S_2 and S_3, consist of the individuals that are contained only in the first, the second and the third test, respectively. The next three sets, denoted S_4, S_5 and S_6 consist of individuals that are included only in the first and second test, in the first and third test, and in the second and third test, respectively; the last set, denoted by S_7, consists of individuals that are included in all three tests. Note that $u(T^*) \le \sum_{j\in[7]} q_{S_j} \cdot \sum_{i\in S_j} u_i$. Without loss of generality, assume that the tests in T^* are ordered such that $q_{S_j} \sum_{i\in S_j} u_i \ge q_{S_{j+1}} \sum_{i\in S_{j+1}} u_i$ for every $j \in [6]$. Finally, we define the non-overlapping test allocation $T = \{S_1, S_2, S_3\}$, so that $u(T) = \sum_{j\in[3]} q_{S_j} \cdot \sum_{i\in S_j} u_i$, and $\frac{u(T^*)}{u(T)} \le \frac{7}{3}$.

Now consider the case $B = 4$. We partition the individuals into 15 sets similarly: the first fours sets consist of individuals included in exactly one test, the next six tests consist of individuals included in exactly two tests, the next four tests consist of individuals included in exactly three tests and the last set consists of individuals that included in all four tests. We then construct a non-overlapping test allocation with four tests that pools individuals from the four sets with the highest utility to get an approximation of $\frac{15}{4}$. □

A.2 Proof of Theorem 3.5

We now prepare for the proof of Theorem 3.5. We first introduce some notation. Given a test allocation $T = (t_1, \ldots, t_B)$ and individual $i \in [n]$, we let $T(i) = \{t_j \in T \mid i \in t_j\}$ be the tests in which individual i is included. Furthermore, $T(i;j) = \{t_{j'} \in T(i) \mid j' < j\}$ denotes the tests with index less than j that contain i. We say that test t_j is *pivotal for individual* i if it is the negative test of smallest index in $T(i)$. That is, we have $t_j \in T_i$, the outcome of t_j is negative, and all tests in $T(i;j)$ are positive. It is immediate that at most one test can be pivotal for each individual. We let $P^T_{i,j}$ denote the probability that t_j is pivotal for individual i under random infection realizations:

$$P^T_{i,j} = \begin{cases} \Pr[t_j \text{ is negative, and } \forall t_{j'} \in T(i;j) : t_{j'} \text{ is positive}] & \text{if } t_j \in T(i), \\ 0 & \text{otherwise.} \end{cases}$$

Observation A.2. The notion of pivotal tests permits a convenient equivalent formulation of the expected utility of a test allocation. Note that individual i is in a negative test if and only if exactly one test, $t_j \in T$, is pivotal for i, hence $P^T_i = \sum_{j\in[B]} P^T_{i,j}$. Recalling that $u(T) = \sum_{i\in[n]} u_i \cdot P^T_i$,

we can expand the expression with pivotal test probabilities and switch the order of summation to obtain

$$u(T) = \sum_{j\in[B]} \left(\sum_{i\in[n]} u_i \cdot P_{i,j}^T \right).$$

For each test $t_j \in T$, we can interpret $\sum_{i\in[n]} u_i \cdot P_{i,j}^T$ as the expected marginal utility gained from applying test t_j after tests $t_1, \ldots, t_{j-1}$ have been applied.

Suppose T is an optimal overlapping test allocation, and i is an individual in the population. In order to prove Theorem 3.5, we first rewrite the probability $P_{i,j}^T$ that the test t_j containing i is pivotal for individual i using conditional probabilities.

$$\begin{aligned} P_{i,j}^T &= \Pr[t_j \text{ is negative and } \forall t_k \in T(i;j), t_k \text{ is positive}] \\ &= \Pr[t_j \text{ is negative}] \cdot \Pr[\forall t_k \in T(i;j), t_k \text{ is positive} \mid t_j \text{ is negative}] \\ &= \Pr[t_j \text{ is negative}] \cdot \Pr[\forall t_k \in T(i;j), t_k \setminus t_j \text{ is positive}] \\ &= q_{t_j} \cdot \Pr[\forall t_k \in T(i;j), t_k \setminus t_j \text{ is positive}] \end{aligned} \tag{12}$$

where the third equality follows since t_k is positive if and only if some individual in $t_k \setminus t_j$ is infected (as t_j is negative by assumption), and the health probabilities of individuals are independent by assumption. We now introduce two lemmas that are used in the proof of Theorem 3.5.

Lemma A.3. *There exists an optimal overlapping test allocation $T = (t_1, \ldots, t_B)$ such that for every $t_j \in T(i)$ we have $P_{i,j}^T > 0$.*

Proof. Suppose T is an 'inclusion-wise minimal' optimal overlapping test allocation in the sense that no test can be reduced in size while keeping the other tests the same, without reducing the welfare obtained. Fix individual i and $t_j \in T(i)$. Using the reformulation of Eq. (12), we now show that $P_{i,j}^T = q_{t_j} \cdot \Pr[\forall t_k \in T(i;j), t_k \setminus t_j \text{ is positive}]$ is strictly positive.

First note that $q_{t_j} > 0$. Indeed, if this were not the case, there would be some individual $i' \in t_j$ with $q_{i'} = 0$, and it is straightforward to see that it is sub-optimal to include such an individual in any test allocation. It remains to show that the second term is non-zero. This holds if there exists an $i' \in t_k \setminus t_j$ with $q_{i'} < 1$ for each $t_k \in T(i;j)$. We now prove this property.

If the difference $t_k \setminus t_j$ is empty, or contains only healthy individuals, then the probability of being in a healthy test either stays the same or increases for every individual. Hence the test allocation obtained from T by reducing t_j to $t_j \setminus t_k$ achieves the same or better welfare. But this contradicts our assumption that T is an inclusion-wise optimal overlapping test allocation. □

Lemma A.4. *Suppose that $T^* = (t_1^*, \ldots, t_B^*)$ is either an optimal overlapping or an optimal non-overlapping test allocation and that $\alpha \in (0,1)$. For any t_j^* and any $S \subset t_j^*$, if $q_S < \alpha$, then $q_{t_j^* \setminus S} \geq 1 - \alpha$.*

Proof. Assume for the sake of contradiction that $q_S < \alpha$ and $q_{t_B^* \setminus S} < 1 - \alpha$ for test t_B^* and subset $S \subseteq t_B^*$. The choice of test t_B is without loss of generality, as we can relabel the tests if necessary. By Lemma A.3, we know that $P_{i,B}^{T^*} > 0$ for any individual i contained in t_B^*. It follows that

$$\begin{aligned} u(T^*) &= \sum_{i\in[n]} P_i^T \cdot u_i = \sum_{i\in[n], j\in[B]} P_{i,j}^{T^*} \cdot u_i \\ &= \sum_{j\in[B-1], i\in[n]} P_{i,j}^{T^*} \cdot u_i + \sum_{i\in[n]} I_{i,B}^{T^*} \cdot q_{t_B^*} \cdot \Pr[\forall t_j^* \in T^*(i;B), t_j^* \setminus t_B^* \text{ is positive}] \cdot u_i \\ &= \sum_{j\in[B-1], i\in[n]} P_{i,j}^{T^*} \cdot u_i + q_S \cdot q_{t_B^* \setminus S} \cdot \sum_{i\in t_B^*} \Pr[\forall t_j^* \in T^*(i;B), t_j^* \setminus t_B^* \text{ is positive}] \cdot u_i \end{aligned}$$

$$
\begin{aligned}
&= \sum_{j\in[B-1],i\in[n]} P^{T^*}_{i,j}\cdot u_i + q_{t^*_B\setminus S}\cdot\left(q_S\cdot\sum_{i\in S}\Pr[\forall t^*_j\in T^*(i;B), t^*_j\setminus t^*_B \text{ is positive}]\cdot u_i\right)\\
&\quad + q_S\left(q_{t^*_B\setminus S}\cdot\sum_{i\in t^*_B\setminus S}\Pr[\forall t^*_j\in T^*(i;B), t^*_j\setminus t^*_B \text{ is positive}]\cdot u_i\right)\\
&\leq \sum_{j\in[B-1],i\in[n]} P^{T^*}_{i,j}\cdot u_i + q_{t^*_B\setminus S}\cdot\left(q_S\cdot\sum_{i\in S}\Pr[\forall t^*_j\in T^*(i;B), t^*_j\setminus S \text{ is positive}]\cdot u_i\right)\\
&\quad + q_S\left(q_{t^*_B\setminus S}\cdot\sum_{i\in t^*_B\setminus S}\Pr[\forall t^*_j\in T^*(i;B), t^*_j\setminus (t^*_B\setminus S) \text{ is positive}]\cdot u_i\right)\\
&< \sum_{j\in[B-1],i\in[n]} P^{T^*}_{i,j}\cdot u_i + \max\left\{q_S\cdot\sum_{i\in S}\Pr[\forall t^*_j\in T^*(i;B), t^*_j\setminus S \text{ is positive}]\cdot u_i,\right.\\
&\qquad \left. q_{t^*_B\setminus S}\cdot\sum_{i\in t^*_j\setminus S}\Pr[\forall t^*_j\in T^*(i;B), t^*_j\setminus (t^*_B\setminus S) \text{ is positive}]\cdot u_i\right\}
\end{aligned}
$$

To justify the first inequality, we begin by showing that, for all $i\in S$,

$$\Pr[\forall t^*_j\in T^*(i;j),\ t^*_j\setminus t^*_B \text{ is positive}]\leq \Pr[\forall t^*_j\in T^*(i;j),\ t^*_j\setminus S \text{ is positive}].$$

Since $S\subseteq t^*_B$, it follows that $t^*_j\setminus t^*_B\subseteq t^*_j\setminus S$ for every $t^*_j\in T^*(i;j)$. This in turn implies that if $t^*_j\setminus t^*_B$ is positive, then $t^*_j\setminus S$ is positive. It follows that the event $[\forall t^*_j\in T^*(i:j),\ t^*_j\setminus t^*_B$ is positive$]$ implies the event $[\forall t^*_j\in T^*(i:j),\ t^*_j\setminus S$ is positive$]$, and so the inequality follows. Furthermore, the argumentation above can be replicated with $t^*_B\setminus S$ rather than S to fully justify the inequality from the main derivation. The second inequality follows from our assumptions $q_S<\alpha$ and $q_{t^*_B\setminus S}<1-\alpha$.

To reach a contradiction, let T be a test allocation with $t_j=t^*_j$ for every $j\in[B-1]$ and $t_B=S$ if

$$
\begin{aligned}
&q_S\cdot\sum_{i\in S}\Pr[\forall t^*_j\in T^*(i;B), t^*_j\setminus S \text{ is positive}]\cdot u_i\\
&\geq q_{T^*_B\setminus S}\cdot\sum_{i\in T^*_B\setminus S}\Pr[\forall t^*_j\in T^*(i;B), t^*_j\setminus (t^*_B\setminus S) \text{ is positive}]\cdot u_i
\end{aligned}
$$

and $t_B=t^*_B\setminus S$, otherwise. Then,

$$
\begin{aligned}
u(T) &= \sum_{i\in[n],j\in[B]} P^T_{i,j}\cdot u_i\\
&= \sum_{j\in[B-1],i\in[n]} P^{T^*}_{i,j}\cdot u_i + \sum_{i\in t_B} q_{t_B}\cdot\Pr[\forall t^*_j\in T^*(i;B), t^*_j\setminus t_B \text{ is positive}]\cdot u_i\\
&= \sum_{j\in[B-1],i\in N} P^{T^*}_{i,j}\cdot u_i + \max\left\{q_S\cdot\sum_{i\in S}\Pr[\forall t^*_j\in T^*(i,B), t^*_j\setminus S \text{ is positive}]\cdot u_i,\right.\\
&\qquad \left. q_{t^*_B\setminus S}\cdot\sum_{i\in T_B\setminus S}\Pr[\forall t^*_j\in T^*(i;B), t^*_j\setminus (t^*_B\setminus S) \text{ is positive}]\cdot u_i\right\}\\
&> u(T^*),
\end{aligned}
$$

contradicting the optimality of T^*. Notice that if T^* is non-overlapping, then so is T, which is why optimality is contradicted in both the scenario where T^* is assumed to be optimal non-overlapping as well as when T^* is assumed to be optimal overlapping. □

We are now ready to prove Theorem 3.5.

Proof of Theorem 3.5. We begin by constructing an intermediate non-overlapping test allocation T from T^* as follows. Every individual who is tested under T^* is contained in one or more tests. Let T be the allocation obtained by removing each such individual i from all but the first test in T^* in which i appears. Note that some tests in T may be empty. Thus we have $t_j \subseteq t_j^*$ and so $q_{t_j} \geq q_{t_j^*}$ for every $j \in [B]$. We let S_j be the smallest subset of t_j such that $q_{S_j} < 1/2$ (if $q_{t_j} \geq 1/2$, then $S_j = \emptyset$). Note that this implies $q_{S_j \setminus \{i\}} \geq 1/2$ for every $i \in S_j$. We can also show that $q_{t_j \setminus S_j} \geq 1/2$. To see this, note that $S_j \subseteq t_j \subseteq t_j^*$ and $q_{S_j} < 1/2$. By Lemma A.4, $q_{t_j^* \setminus S_j} \geq 1/2$, and it follows that $q_{t_j \setminus S_j} \geq q_{t_j^* \setminus S_j} \geq 1/2$ since $t_j \setminus S_j \subseteq t_j^* \setminus S_j$.

Next, consider the two disjoint test allocations T^1 and T^2 given by $t_j^1 = S_j$ and $t_j^2 = t_j \setminus S_j$. Using the properties established in the previous paragraph, we now show that $P_i^{T^\ell} \geq q_i \cdot 1/2$ for every $i \in t_j^\ell$ where $\ell \in \{1, 2\}$. For $\ell = 1$, we see that $P_i^{T^1} = q_{t_j^1} = q_i \cdot q_{t_j^1 \setminus \{i\}} \geq q_i \cdot 1/2$, as $q_{t_j^1 \setminus \{i\}} \geq 1/2$. For $\ell = 2$, we see similarly that $P_i^{T^2} = q_{t_j^2} = q_i \cdot q_{t_j^2 \setminus \{i\}} \geq q_i \cdot 1/2$, as our choice of S_j ensures that $q_{t_j^2 \setminus \{i\}} \geq q_{t_j^2} = q_{t_j \setminus S_j} \geq 1/2$.

Note that the same individuals are tested under T^* and under the intermediate test allocation T, as every individual included in some test under T^* is also included in some test under T, and $t_j \subseteq t_j^*$ for every $j \in [B]$. We can write the welfare of T^* as

$$u(T^*) = \sum_{i \in [n]} P_i^{T^*} \cdot u_i = \sum_{j \in [B], i \in S_j} P_i^{T^*} \cdot u_i + \sum_{j \in [B], i \in T_j \setminus S_j} P_i^{T^*} \cdot u_i.$$

Suppose that $\sum_{j \in [B], i \in S_j} P_i^{T^*} \cdot u_i \geq \sum_{j \in [B], i \in t_j \setminus S_j} P_i^{T^*} \cdot u_i$. Then

$$u(T^*) \leq 2 \cdot \sum_{j \in [B], i \in S_j} P_i^{T^*} \cdot u_i \leq 2 \cdot \sum_{j \in [B], i \in S_j} q_i \cdot u_i$$
$$\text{and} \quad u(T^1) = \sum_{j \in [B], i \in S_j} P_i^{T^1} \cdot u_i \geq \sum_{j \in [B], i \in S_j} 1/2 \cdot q_i \cdot u_i,$$

so $\frac{u(T^*)}{u(T^1)} \leq 4$. Finally, suppose that $\sum_{j \in [B], i \in S_j} P_i^{T^*} \cdot u_i < \sum_{j \in [B], i \in t_j \setminus S_j} P_i^{T^*} \cdot u_i$. An analogous argument shows that $\frac{u(T^*)}{u(T^2)} \leq 4$. Hence either T^1 or T^2 achieves the required overlap welfare ratio of at most 4. □

B Proofs for Section 4 (Finding Near-Optimal Test Allocations)

B.1 Correctness of Greedy

Proof of Theorem 4.1. Let T be the test allocation that is returned by Greedy, and T^* be an optimal non-overlapping test allocation. Without loss of generality, let $N' = \{1, \ldots, n'\}$ be the individuals included in T. We can write out the welfares of T and T^* as

$$u(T^*) = \sum_{j \in [B]} u(t_j^*) = \sum_{j \in [B]} q_{t_j^*} \cdot \left(\sum_{i \in N'} I_{i,j}^{t_j^*} \cdot u_i \right) + \sum_{j \in [B]} q_{t_j^*} \cdot \left(\sum_{i \in [n] \setminus N'} I_{i,j}^{t_j^*} \cdot u_i \right),$$

and

$$u(T) = \sum_{j \in [B]} u(t_j) = \sum_{j \in [B]} q_{t_j} \cdot \left(\sum_{i \in N'} I_{i,j}^{T} \cdot u_i \right) = \sum_{j \in [B]} q_{t_j} \cdot \left(\sum_{i \in t_j} u_i \right).$$

Now, let $T' = (t_1', \ldots, t_B')$ be the test allocation created from T^* by removing any individuals in N', so $t_j' = t_j^* \setminus N'$ for every $j \in [B]$. Note that since every t_j' consists of individuals that

are not included in any test in T, all the individuals in t'_j are available at the j-th iteration of GREEDY. Thus for each $j \in [B]$, we have $u(t_j) \geq (1-\varepsilon)u(t'_j)$, as otherwise GREEDY would have chosen t'_j instead of t_j at the j-th iteration. Thus, we get that

$$u(T) = \sum_{j\in[B]} u(t_j) \geq (1-\varepsilon) \sum_{j\in[B]} u(t'_j) = (1-\varepsilon) \cdot u(T').$$

Note also that

$$u(T') = \sum_{j\in[B]} q_{t'_j} \cdot \left(\sum_{i\in[n]\setminus N'} I^{T'}_{i,j} \cdot u_i \right) = \sum_{j\in[B]} q_{t'_j} \cdot \left(\sum_{i\in[n]\setminus N'} I^{T^*}_{i,j} \cdot u_i \right) \geq \sum_{j\in[B]} q_{t^*_j} \cdot \left(\sum_{i\in[n]\setminus N'} I^{T^*}_{i,j} \cdot u_i \right)$$

where the second equality follows from the fact that $I^{T'}_{i,j} = I^{T^*}_{i,j}$ for any $i \in [n] \setminus N'$ and $j \in [B]$, and the last inequality follows from the fact that $q_{t'_j} \geq q_{t^*_j}$ since $t'_j \subseteq t^*_j$ for any $j \in [B]$. Thus,

$$u(T) \geq (1-\varepsilon)^2 \cdot \sum_{j\in[B]} q_{t^*_j} \cdot \left(\sum_{i\in[n]\setminus N'} I^{T^*}_{i,j} \cdot u_i \right).$$

From all the above we have

$$\begin{aligned}
\frac{u(T^*)}{u(T)} &= \frac{\sum_{j\in[B]} q_{t^*_j} \cdot \left(\sum_{i\in N'} I^{T^*}_{i,j} \cdot u_i \right) + \sum_{j\in[B]} q_{t^*_j} \cdot \left(\sum_{i\in[n]\setminus N'} I^{T^*}_{i,j} \cdot u_i \right)}{u(T)} \\
&\leq \frac{\sum_{j\in[B]} q_{t^*_j} \cdot \left(\sum_{i\in N'} I^{T^*}_{i,j} \cdot u_i \right) + \frac{u(T)}{(1-\varepsilon)^2}}{u(T)} \\
&= \frac{\sum_{j\in[B]} q_{t^*_j} \cdot \left(\sum_{i\in t^*_j\cap N'} u_i \right)}{u(T)} + \frac{1}{(1-\varepsilon)^2} \\
&\leq \frac{\sum_{j\in[B]} \left(\sum_{i\in t^*_j\cap N'} q_i \cdot u_i \right)}{u(T)} + \frac{1}{(1-\varepsilon)^2} \\
&= \frac{\sum_{i\in N'} q_i \cdot u_i}{u(T)} + \frac{1}{(1-\varepsilon)^2} \qquad (13)
\end{aligned}$$

where the second inequality follows since $q_i \geq q_{t_j}$ when i is included in t_j.

In what follows, we will show that each test $t_j \in T$ obtains at least a $\frac{(1-\varepsilon)^2}{4}$ ratio of the maximal possible utility to be gained from individuals in t_j. In other words, we show that the following holds: $u(t_j) = q_{t_j} \cdot \sum_{i\in t_j} u_i > \frac{(1-\varepsilon)^2}{4} \sum_{i\in t_j} q_i u_i$. To do so, we first show that, in a similar nature to Lemma A.4, for any $t_j \in T$ there cannot exist $S \subseteq t_j$ such that $q_S + q_{t_j\setminus S} < 1-\epsilon$. Suppose that this is not the case, and we have such a subset S. From the definition of GREEDY, we know that $u(t_j) \geq (1-\varepsilon)u(S)$ and $u(t_j) \geq (1-\varepsilon)u(t_j \setminus S)$, otherwise the algorithm would return S or $(t_j \setminus S)$ respectively at the j-th step of its execution. It follows that $u(t_j) \geq (1-\varepsilon)\max\{u(S), u(t_j \setminus S)\}$. At the same time, we can also express the welfare of t_j as $u(t_j) = q_{t_j\setminus S}\left(q_S \cdot \sum_{i\in S} u_i\right) + q_S\left(q_{t_j\setminus S} \cdot \sum_{i\in t_j\setminus S} u_i\right)$. From this we obtain $u(t_j) = q_{t_j\setminus S} \cdot u(S) + q_S \cdot u(t_j \setminus S) \leq (q_{t_j\setminus S} + q_S)\max\{u(S), u(t_j \setminus S)\}$. Under the assumption that $q_S + q_{t_j\setminus S} < 1-\varepsilon$, it follows that $u(t_j) < (1-\varepsilon)\max\{u(S), u(t_j \setminus S)\}$, which is a contradiction, hence no such $S \subseteq t_j$ can exist.

Now let us consider the case where there exists $i' \in t_j$ such that $q_{i'} < (1-\varepsilon)/2$. From the definition of the GREEDY algorithm, we know that

$$(1-\varepsilon) \cdot q_{t_j\setminus\{i'\}} \cdot \sum_{i\in t_j\setminus\{i'\}} u_i \leq q_{t_j} \cdot \sum_{i\in t_j} u_i$$

otherwise the algorithm would return $t_j \setminus \{i'\}$ instead of t_j at step j and also,

$$(1-\varepsilon) \cdot q_{i'} \cdot u_{i'} \leq q_{t_j} \cdot \sum_{i \in t_j} u_i$$

as otherwise the algorithm would return $\{i'\}$ instead of t_j at step j. Moreover, we know that $q_{t_j \setminus \{i'\}} \geq (1-\varepsilon)/2$ since it must be the case that $q_{i'} + q_{t_j \setminus \{i'\}} \geq 1-\varepsilon$. Thus, we get that

$$\begin{aligned} q_{t_j} \cdot \sum_{i \in t_j} u_i \geq \frac{(1-\varepsilon)}{2}(q_{t_j \setminus \{i'\}} \cdot \sum_{i \in t_j \setminus \{i\}} u_i + q_{i'} \cdot u_{i'}) &\geq \frac{(1-\varepsilon)}{2}\left(\frac{1-\varepsilon}{2} \cdot \sum_{i \in t_j \setminus \{i\}} u_i + q_{i'} \cdot u_{i'}\right) \\ &= \frac{(1-\varepsilon)^2}{4}\left(\sum_{i \in t_j \setminus \{i'\}} u_i + q_{i'} \cdot u_{i'}\right) \\ &\geq \frac{(1-\varepsilon)^2}{4} \sum_{i \in t_j} q_i u_i. \end{aligned}$$

As a second case, assume that for any $\ell \in t_j$, $q_\ell \geq (1-\varepsilon)/2$. We show that for any $i \in t_j$, $q_{t_j \setminus \{i\}} \geq (1-\varepsilon)^2/4$. If $q_{t_j} \geq (1-\varepsilon)/2$, then indeed $q_{t_j \setminus \{i\}} \geq q_{t_j} > (1-\varepsilon)^2/4$. Hence we focus on the scenario where $q_{t_j} < (1-\varepsilon)/2$ and consider an arbitrary $i \in t_j$. Let S be a non-empty subset $S \subseteq t_j \setminus \{i\}$ such that $q_S \geq (1-\varepsilon)/2$ and $q_{S \cup \{i\}} < (1-\varepsilon)/2$. Since we know that $q_\ell \geq (1-\varepsilon)/2$ for each $\ell \in t_j$ and that $q_{t_j} < (1-\varepsilon)/2$, there must exist such an S. Since t_j can be decomposed into $S \cup \{i\}$ and $t_j \setminus (S \cup \{i\})$, it follows that $q_{S \cup \{i\}} + q_{t_j \setminus (S \cup \{i\})} \geq 1-\varepsilon$, hence $q_{t_j \setminus (S \cup \{i\})} \geq (1-\varepsilon)/2$. Putting everything together, $q_{S \setminus \{i\}} = q_{t_j \setminus (S \cup \{i\})} \cdot q_S \geq (1-\varepsilon)^2/4$ as desired. Now we have

$$q_{t_j} \cdot \sum_{i \in t_j} u_i = \sum_{i \in t_j} q_{t_j \setminus \{i\}} \cdot q_i \cdot u_i \geq \frac{(1-\varepsilon)^2}{4} \cdot \sum_{i \in t_j} q_i \cdot u_i.$$

Overall, we see that it is always the case that $u(t_j) = q_{t_j} \cdot \sum_{i \in t_j} \geq \frac{(1-\varepsilon)^2}{4} \cdot \sum_{i \in t_j} q_i \cdot u_i$, hence:

$$u(T) = \sum_{j \in [B]} q_{t_j} \cdot \left(\sum_{i \in t_j} u_i\right) > \sum_{j \in [B]} \left(\frac{(1-\varepsilon)^2}{4} \cdot \sum_{i \in t_j} q_i \cdot u_i\right) = \frac{(1-\varepsilon)^2}{4} \sum_{i \in N'} q_i u_i$$

Along with Equation (13), we get that $u(T^*)/u(T) \leq 5/(1-\varepsilon)^2 \leq 5/(1-\varepsilon')$ for any $0 < \varepsilon' < 1$ with $\varepsilon' \geq 2\varepsilon$ and the theorem follows. □

B.2 Homogeneous Utilities

In this section, we consider the case in which population utilities are homogeneous, so $u_i = u_{i'}$ for all $i, i' \in [n]$. Without loss of generality, assume that $u_i = 1$ for any $i \in [n]$ and $q_i \geq q_{i+1}$ for any $i \in [n-1]$.

B.2.1 Structural Lemmas for Optimal Test Allocations

In the context of homogeneous utilities, we say that a test allocation $T = (t_1, \ldots, t_B)$ is *proper* if the following hold:

- $|t_1| \geq |t_2| \geq \ldots |t_B|$
- for all $j, j' \in [B]$ such that $j < j'$, if $i \in t_j$ and $i' \in t_{j'}$, then $i < i'$.

The following crucial lemma shows that there exist proper optimal test allocations.

Lemma B.1. *Let $T^* = (t_1^*, \ldots, t_B^*)$ such that $|t_j^*| \geq |t_{j+1}^*|$ for any $j \in [B-1]$. Then, there exists a proper optimal test allocation T' such that for each $j \in [B]$, $|t_j'| = |t_j^*|$.*

Proof. First we show that without loss of generality, we can assume that T^* tests precisely $[k] \subseteq [n]$ for some $k \leq n$. Suppose that this is not the case, and there exist $i, i' \in [n]$ such that $i < i'$ where i is not tested in T^* yet i' is tested t_j. It is straightforward to see that the test allocation which replaces i' with i in t_j obtains at least as much utility as T^* and is hence optimal by assumption.

We prove the lemma by induction on the number of tests. We start from the case where $B = 2$. Let T^* be an optimal test allocation with $|t_1^*| \geq |t_2^*|$. Assume that $t_1^* = S_1 \cup S_1'$, where $S_1 \subset \{1, \ldots, |t_1^*|\}$, and $S_1' \subset \{|t_1^*| + 1, \ldots, k\}$ and $t_2^* = S_2 \cup S_2'$, where $S_2 \subset \{1, \ldots, |t_1^*|\}$ and $S_2' \subset \{|t_1^*| + 1, \ldots, k\}$. Since $t_1^* \cup t_2^* = [k]$, we get that $|S_1'| = |S_2|$ and since $|t_1^*| \geq |t_2^*|$, we get that $|S_1| \geq |S_2'|$. Now, consider the test allocation T such that $t_1 = S_1 \cup S_2$ and $t_2 = S_1' \cup S_2'$. Notice that $t_1 \cup t_2 = [k]$, $|t_1| = |t_1^*|$ and $|t_2| = |t_2^*|$. Then, we have

$$u(T^*) = q_{S_1} \cdot q_{S_1'} \cdot |t_1^*| + q_{S_2} \cdot q_{S_2'} \cdot |t_2^*|$$

and

$$u(T) = q_{S_1} \cdot q_{S_2} \cdot |t_1^*| + q_{S_1'} \cdot q_{S_2'} \cdot |t_2^*|.$$

and hence,

$$u(T) - u(T^*) = \left(q_{S_2} - q_{S_1'}\right) \cdot \left(q_{S_1} \cdot |t_1^*| - q_{S_2'} \cdot |t_2^*|\right). \tag{14}$$

Due to optimality of T^*, we have that for any $\hat{S}_1 \subseteq S_1$

$$q_{S_1} \cdot q_{S_1'} \cdot |t_1^*| \geq q_{\hat{S}_1} \cdot q_{S_1'} \cdot (|\hat{S}_1| + |S_1'|)$$

as otherwise if T' is a test allocation with $t_1' = \hat{S}_1 \cup S_1'$ and $t_2' = t_2^*$, then it would hold that $u(T') > u(T^*)$ which is a contradiction to the optimality of T^*. Now, choose an arbitrary $\hat{S}_1 \subseteq S_1$ such that $|\hat{S}_1| = |S_2'|$. We know that this is feasible since $|S_1| \geq |S_2'|$. We know that $q_{S_1} \cdot q_{S_1'} \cdot |t_1^*| \geq q_{\hat{S}_1} \cdot q_{S_1'} \cdot |\hat{S}_1 \cup S_1'| = q_{\hat{S}_1} \cdot q_{S_1'} \cdot |t_2^*|$, where we also used the fact that $|\hat{S}_1| = |S_2'|$. Finally, by construction $q_{\hat{S}_1} \geq q_{S_2'}$, hence it follows that $q_{S_1} \cdot q_{S_1'} \cdot |t_1^*| \geq q_{S_2'} \cdot q_{S_1'} \cdot |t_2^*|$. We thus obtain

$$q_{S_1} \cdot |t_1^*| - q_{S_2'} \cdot |t_2^*| \geq 0.$$

Now from Eq. (14), we have that $u(T) \geq u(T^*)$ since $\left(q_{S_2} - q_{S_1'}\right) \geq 0$ as for each $i \in S_2$ and each $i' \in S_1'$ it holds that $q_i \geq q_{i'}$ and $|S_1'| = |S_2|$. Thus, we conclude that if T^* is optimal, so is T, which is in turn a proper test allocation as desired.

Now, suppose that the claim holds for $B - 1$. We will show that it holds for B. Let $T^* = (t_1^*, \ldots, t_B^*)$ be an optimal test allocation with $|t_j^*| \geq |t_{j+1}^*|$ for any $j \in [B-1]$. It must be the case that the allocation $(t_1^*, \ldots, t_{B-1}^*)$ is optimal for the population given by $\cup_{j=1}^{B-1} t_j^*$, hence we can apply our inductive assumption to obtain a proper optimal test allocation $T' = (t_1', \ldots, t_{B-1}')$ for $\cup_{j=1}^{B-1} t_j^*$ such that $|t_j'| = |t_j^*|$. Let $T^0 = (t_1', \ldots, t_{B-1}', t_B^*)$.

We proceed to create a proper test allocation for all of $[n]$ in $B-1$ rounds. In the ℓ-th round we begin with an optimal test allocation $T^{\ell-1} = (t_1^{\ell-1}, \ldots, t_B^{\ell-1})$ such that $|t_j^{\ell-1}| = |t_j^*|$ for all $j \in [B]$ and $(t_1^{\ell-1}, \ldots, t_{\ell-1}^{\ell-1})$ is an optimal proper test allocation for $\cup_{j=1}^{\ell-1} t_j^{\ell-1} = \{1, \ldots, \sum_{j=1}^{\ell-1} |t_j^*|\}$. We consider the allocation $(t_\ell^{\ell-1}, t_B^{\ell-1})$ which must be optimal for the population given by $t_\ell^{\ell-1} \cup t_B^{\ell-1}$. We can apply the lemma for the case of $B = 2$ to obtain an optimal proper test allocation for $t_\ell^{\ell-1} \cup t_B^{\ell-1}$ which is given by (t_ℓ^ℓ, t_B^ℓ) such that $|t_\ell^{\ell-1}| = |t_\ell^\ell|$ and $|t_B^{\ell-1}| = |t_B^\ell|$. For $j \neq \ell, B$, we let

$t_j^\ell = t_j^{\ell-1}$. It follows that T^ℓ is such that $|t_j^\ell| = |t_j^*|$ for all $j \in [B]$ and $(t_1^\ell, \ldots, t_\ell^\ell)$ is an optimal proper test allocation for $\cup_{j=1}^{\ell} t_j^\ell = \{1, \ldots, \sum_{j=1}^{\ell} |t_j^*|\}$. Given the construction, it follows that T^ℓ for $\ell = B$ is a proper optimal test allocation for the entire population such that $|t_j^\ell| = |t_j^*|$ for all $j \in [B]$ as desired. □

Observation B.2. Using Lemma B.1, we can find an optimal test allocation as follows: for all $k \in [n]$ and $k_1 \geq k_2 \ldots \geq k_B$ with $\sum_{\ell \in [B]} k_\ell = k$, we calculate the welfare of the test allocation, $T = (t_1, \ldots, t_B)$, such that

$$t_j = \Big\{ \sum_{\ell \in [j-1]} k_\ell + 1, \ldots \sum_{\ell \in [j-1]} k_\ell + k_j \Big\}.$$

The allocation which returns the highest welfare amongst the $O(n^{B+1}/B!)$ choices of k and $k_1, \ldots, k_B$ values must necessarily be optimal.

B.2.2 The OrderedGreedy Algorithm

Here, we show that when the utilities are identical, we can find an e-approximate test allocation with respect to the optimal non-overlapping test allocation, for any value B. This result is tight, as shown in Example 4.2. Once more, we assume a population, $[n]$, such that $u_i = 1$ for all i and $q_1 \geq q_2 \geq \cdots \geq q_n$. Specifically, we consider a variation of the Greedy algorithm introduced in Section 4.2 which we call OrderedGreedy. The algorithm proceeds in B rounds. In the j-th round it computes a test t_j composed of individuals not yet tested in previous rounds as follows: it sequentially adds untested individuals of highest health probability to t_j until adding any further untested individual reduces the welfare of t_j. Note that OrderedGreedy always returns a proper test allocation, T, which tests $[n'] \subseteq [n]$ within the population. We start with the following lemma.

Lemma B.3. *Suppose that $T^* = (t_1^*, \ldots, t_B^*)$ is proper optimal test allocation and that T is a proper test allocation computed by* OrderedGreedy. *If T^* tests $[n'] \subseteq [n]$ within the population and T tests $[n''] \subseteq [n]$ within the population, then $n'' \leq n'$.*

Proof. Since T^* is proper, suppose that for $j \in [B]$, $t_j^* = \{i_{j-1}^* + 1, \ldots, i_j^*\}$ with $i_0^* = 0$ and $i_{j-1}^* < i_j^*$. Since T is also proper, for $j \in [B]$ we let $t_j = \{i_{j-1} + 1, \ldots, i_j\}$, with $i_0 = 0$ and $i_{j-1} < i_j$. We show that for each $j \in [B]$, $i_j^* \leq i_j$. Suppose for contradiction that t_j^* is the first test such that $i_j^* > i_j$. Due to the structure of T^* and T, this means that $\cup_{j' \in [j-1]} t_{j'}^* \subseteq \cup_{j' \in [j-1]} t_{j'}$, and hence $i_{j-1}^* \leq i_{j-1}$. Given that OrderedGreedy does not pool $i_j + 1$ in t_j, we have that

$$\begin{aligned} q_{i_{j-1}+1} \cdot \ldots \cdot q_{i_j} \cdot |t_j| &> q_{i_{j-1}+1} \cdot \ldots \cdot q_{i_j} \cdot q_{i_j+1} \cdot (|t_j| + 1) \\ \Rightarrow \frac{|t_j|}{|t_j| + 1} &> q_{i_j+1}. \end{aligned}$$

as otherwise, from the definition of OrderedGreedy, $i_j + 1$ would have been included in t_j.

We also note that,

$$q_{i_j^*} < \frac{|t_j|}{|t_j| + 1} \leq \frac{|t_j^*| - 1}{|t_j^*|}.$$

The first inequality holds due to the fact that $i_j^* \geq i_j + 1$ (resulting from the integrality of the indices) and the fact that we have assumed health probabilities are in decreasing order, hence $q_{i_j^*} \leq q_{i_j+1} < \frac{|t_j|}{|t_j|+1}$. As for the second inequality, we note that by assumption $i_{j-1}^* \leq i_j$ and $i_j^* > i_j$, hence $|t_j^*| > |t_j|$, and from integrality of indices, we get $|t_j^*| - 1 \geq |t_j|$. The function $f(x) = \frac{x-1}{x}$ is increasing in x, hence $\frac{|t_j|}{|t_j|+1} \leq \frac{|t_j^*|-1}{|t_j^*|}$. Thus, we have that,

$$q_{t_j^* \setminus \{i_j^*\}} \cdot (|t_j^*| - 1) > q_{t_j^* \setminus \{i_j^*\}} \cdot q_{i_j^*} \cdot |t_j^*|.$$

This means that $u(t_j^* \setminus \{i_j^*\}) > u(t_j^*)$ and hence, we have that if T' is the test allocation with $t'_{j'} = t^*_{j'}$ for each $j' \neq j$ and $t'_j = t^*_j \setminus \{i^*_j\}$, then $u(T') > u(T^*)$ which is a contradiction. We conclude that for each $j \in [B]$, $i^*_j \leq i_j$, and the statement follows. □

Now, we are ready to show that OrderedGreedy returns an e-approximate test allocation for any population. This is complemented by a matching lower bound shown in Example 4.2 below.

Theorem B.4. OrderedGreedy *returns an e-approximate test allocation.*

Proof. Let T be the test allocation that is returned by OrderedGreedy which tests the first $n' \leq n$ individuals. We start by showing that for each i, $P_i^T \geq q_i \cdot \frac{1}{e}$. Consider a test t_j in T of size k and let i' be the final individual included in t_j. It must be the case that $q_{i'} \geq (k-1)/k$, for otherwise

$$q_{i'} \prod_{i'' \in t_j \setminus \{i'\}} q_{i''} \cdot k < \prod_{i'' \in t_j \setminus \{i'\}} q_{i''} \cdot (k-1)$$

which is a contradiction. Since i' was assumed to be the final individual added to t_j in OrderedGreedy, it follows that for all $i \in t_j$, $q_i \geq q'_i \geq (k-1)/k$. With this in hand, we get that

$$P_i^T = q_{t_j} = q_i \cdot \prod_{i' \in t_j \setminus \{i\}} q_{i'} \geq q_i \cdot \left(\frac{k-1}{k}\right)^{k-1} \geq q_i \cdot \frac{1}{e}. \tag{15}$$

From Lemma B.3, we know that it exists an optimal non-overlapping test allocation T^* that pools the first n'' individuals with $n'' \leq n'$. Then, we have

$$\frac{u(T^*)}{u(T)} = \frac{\sum_{i \in [n'']} P_i^T \cdot u_i}{\sum_{i \in [n']} P_i^T \cdot u_i} \leq \frac{\sum_{i \in [n']} q_i \cdot u_i}{\sum_{i \in [n']} q_i \cdot \frac{1}{e} \cdot u_i} \leq e,$$

where the third inequality follows from Equation (15). □

B.2.3 Optimality of Greedy for Clustered Populations

Proof of Proposition 4.4. Suppose that $T^* = (t_1^*, \ldots, t_B^*)$ is an optimal overlapping test allocation and that t^* is an optimal single group test for the population. By Observation A.2, the welfare of T^* is

$$u(T^*) = \sum_{j \in [B]} \sum_{i \in t_j^*} u_i \cdot P_{i,j}^{T^*},$$

and $\sum_{i \in t_j^*} u_i \cdot P_{i,j}^{T^*}$ is the marginal utility of t_j^*. Since $P_{i,j}^{T^*} \leq q_{t_j^*}$, it follows that the marginal utility of t_j^* is at most $q_{t_j^*} \sum_{i \in t_j^*} u_i = u(t_j^*) \leq u(t^*)$. It follows that $u(T^*) \leq B \cdot u(t^*)$. On the other hand, $u(t) \geq (1-\varepsilon)u(t^*)$. The fact that $B \cdot t(\ell) \leq n_\ell$ holds for each $\ell \in [m]$ means that t can be applied B times in the population, which is what Greedy does, returning allocation T. It follows that $u(T) = B \cdot u(t) \geq B(1-\varepsilon)u(t^*) \geq (1-\varepsilon)u(T^*)$. □

B.3 An FPTAS for single-test allocations

We describe a fully polynomial-time approximation scheme (FPTAS) for optimally allocating a single test with pool size constraint $G < n$. This FPTAS is obtained by modifying the FPTAS of Goldberg and Rudolf (2020), which finds an approximately optimal single test of arbitrary size. We use similar notation as Goldberg and Rudolf (2020) where possible.

For any $i \in [n]$, let $P(i, C, L) = \{q_S \mid S \subseteq [i], |S| = L, \sum_{i \in S} u_i = C\}$ denote the maximum probability of a subset of $[i]$ to be negative with utility sum C and size L. Analogous to Eq. (6)

of (Goldberg and Rudolf, 2020), we consider the dynamic program (DP) given by the recurrence relations

$$P(i,C,L) = \begin{cases} \max\{P(i-1,C,L), q_i \cdot P(i-1,C-u_i,L-1)\} & i \geq 2 \text{ and } u_i < C, \\ P(i-1,C,L) & i \geq 2 \text{ and } u_i \geq C, \\ q_i & i = 1 \text{ and } u_1 = C, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

With a slight abuse of notation, we denote by $t(P(i,C,L))$ a test that forms a subset of $[i]$ and satisfies $q_{t(i,C,L)} = P(i,C,L)$, $\sum_{\ell \in [t(i,C,L)]} u_\ell = C$ and $|t(i,C,L)| = L$.

Let $\bar{C}$ be an upper bound on the sum of utilities of an optimal test ($\bar{C} = \sum_{i\in[n]} u_i$ suffices). It is straightforward that the test $t(n,C^*,L^*)$ with $C^*, L^* = \arg\max_{C\in[\bar{C}],L\in[G]} C \cdot P(n,C,L)$ is welfare-maximizing and has pool size at most G. The running time of this dynamic program is given by $O(nG\bar{C}) \leq O(n^2\bar{C})$, since $G \leq n$. This is pseudo-polynomial in the input size as $\bar{C}$ is potentially exponential in the size of its representation.

In what follows, we describe a modification of the utilities in order to approximately solve the dynamic programming with a polynomial running time. In order to achieve this, we need to scale down and round the utility coefficients whose magnitudes, through the upper bound $\bar{C}$, determine the running time of the program. We can achieve this using identical arguments as in Section 3.2 of Goldberg and Rudolf (2020). We present the full proof here for completeness.

We scale down utilities using factor κ and round the result, by setting $\hat{u}_i = \lfloor u_i/\kappa \rfloor$ for every $i \in [n]$. Before choosing κ, we introduce notation. Let $N_{1/2} = \{i \in [n] : q_i \geq \frac{1}{2}\}$ be the set of individuals with health probability at least $\frac{1}{2}$. Without loss of generality, we assume that $N_{1/2} = [h]$ and $[n] \setminus N_{1/2} = \{h+1,\ldots,n\}$, relabeling individuals if necessary. Let $\hat{P}(i,C,L)$ denote the DP in Equation (16) by replacing u_i with $\hat{u}_i$. Moreover, we assume that there exists a dummy individual $n+1$ with $\hat{u}_{n+1} = 0$ and $q_{n+1} = 1$. Then, for $i \in [n]$ and $j > i$ the scaled DP problem is defined as

$$\hat{z}_\kappa(i,j) = \max_{C\in[\bar{C}(i)],L\in[G]} (\kappa \cdot C + u_j) \cdot \hat{P}(i,C,L) \cdot q_j, \tag{17}$$

where $\bar{C}(i) = \sum_{i'\in[i]} u_{i'}$. Let

$$\hat{C}_{i,j}, \hat{L}_{i,j} = \arg\max_{C\in[\bar{C}(i)],L\in[G]} (\kappa \cdot C + u_j) \cdot \hat{P}(i,C,L) \cdot q_j.$$

Note that $\hat{t} = t(\hat{P}(i,\hat{C}_{i,j},\hat{L}_{i,j}) \cup \{j\})$ returns an optimal test by replacing u_i with $\hat{u}_i$ and adding the constraint that for any $\ell \in [i+1,\ldots,n] \setminus \{j\}$, ℓ is not pooled into the test, while j is pooled into it. From Lemma A.4, we know that it suffices to evaluate $\hat{z}(i,j)$ for $i \in [h]$ and $j \in \{h+1,\ldots,n\}$ in order to evaluate $\hat{z}_\kappa(n,n+1)$, as at most one individual from $[n] \setminus N_{1/2}$ may be included in $\hat{t}$. The specific reason for this being that if $j,j' \in \{h+1,\ldots,n\}$ form a part of $\hat{t}$, then $q_j \leq 1/2$ and $q_{\hat{t}\setminus j} < q_{j'} \leq 1/2$. The following lemma establishes an upper bound on a value of κ that suffices to bound the relative error of solutions of $\hat{z}_\kappa$ in approximating the optimal test within a given $\varepsilon > 0$.

Lemma B.5. *Let t^* be an optimal single test with $\sum_{\ell\in t^*} u_\ell = C^*$ and $|t^*| = L^*$. For a given $\varepsilon > 0$, there exist $i \in [h]$ and $j \in \{h,\ldots,n+1\}$, such that if $\bar{t} = t^* \setminus \{j\}$ and $\kappa \leq \frac{\varepsilon \max_{i\in\bar{t}} q_i \cdot u_i}{n}$ then*

$$\hat{z}_\kappa(i,j) \geq (1-\varepsilon) \cdot C^* \cdot P(n,C^*,L^*) = (1-\varepsilon) \cdot u(t^*).$$

Proof. First note

$$\sum_{i\in t^*} u_i - \kappa \sum_{i\in t^*} \hat{u}_i = \sum_{i\in t^*} u_i - \kappa \sum_{i\in t^*} \lfloor u_i/\kappa \rfloor \leq \sum_{i\in t^*} u_i - \kappa \sum_{i\in t^*} (u_i/\kappa - 1) \leq \kappa n,$$

Algorithm 1 The FPTAS for computing an ε-optimal single-test allocation.

$\kappa \leftarrow (\varepsilon \cdot 1/2 \cdot \max_{i \in N_{1/2}} u_i)/n$
$z^* \leftarrow 0; t \leftarrow \emptyset$
for $j = h+1, \dots n$ **do**
  if $\hat{z}(h,j) < q_j \cdot u_j$ **then**
    if $q_j \cdot u_j > z^*$ **then**
      $z^* \leftarrow q_j \cdot u_j$
      $t \leftarrow \{j\}$
    end if
  else
    if $\hat{z}(h,j) > z^*$ **then**
      $z^* \leftarrow \hat{z}(h,j)$
      $t \leftarrow t(\hat{P}(h, \hat{C}_{i,j}, \hat{L}_{i,j}))$
    end if
  end if
end for
return t

where the last inequality follows since $|t^*| \leq n$.

Let j be the individual with the smallest probability of being healthy in t^* by breaking ties with respect to individuals that have higher index, and let i be the individual with the highest index in $t^* \setminus \{j\}$. We denote by $\hat{t} \subseteq [i]$ the set that maximizes Equation (17). We have

$$\begin{aligned}\hat{z}_\kappa(i,j) = q_j \left(\kappa \sum_{\ell \in \hat{t} \setminus \{j\}} \hat{u}_\ell + u_j\right) q_{\hat{t} \setminus \{j\}} &= q_j \left(\kappa \sum_{\ell \in t^* \setminus \{j\}} \hat{u}_\ell + u_j\right) q_{t^* \setminus \{j\}} \\ &\geq \left(1 - \frac{n\kappa}{\sum_{i \in t^*} u_i}\right) \cdot q_{t^*} \sum_{i \in t^*} u_i,\end{aligned}$$

where the first equality follows from optimality of $\hat{t}$ under the scaled utilities. The second equality follows from the fact that t^* must necessarily also be optimal for truncated utilities. Finally the inequality holds from the fact that we showed $\sum_{i \in t^*} u_i - \kappa \sum_{i \in t^*} \hat{u}_i \leq \kappa n$. Thus, to ensure an ε-approximate solution, we need

$$\frac{n\kappa}{\sum_{\ell \in t^*} u_\ell} \leq \varepsilon \Leftrightarrow \kappa \leq \frac{\varepsilon \cdot \sum_{\ell \in t^*} u_\ell}{n}.$$

Thus, it suffices to choose

$$\kappa \leq \frac{\varepsilon \cdot \max_{\ell \in [t^* \setminus \{j\}]} u_\ell \cdot q_\ell}{n} \leq \frac{\varepsilon \cdot q_{t^*} \cdot \sum_{i \in t^*} u_i}{n} \leq \frac{\varepsilon \cdot \sum_{\ell \in t^*} u_i}{n}.$$

□

Even though we do not know t^* a priori, we can choose a value for κ that satisfies the above lemma and hence guarantees that method of computing the optimal test for truncated utilities indeed gives rise to an FPTAS. We detail how such a value of κ can be set as per the initialization of Algorithm 1. Before continuing, we recall that from Lemma A.4, we know that $t^* \cap ([n] \setminus N_{1/2}) \leq 1$.

Algorithm 1 eventually fixes the single individual $j \in [n+1] \setminus N_{1/2}$ that is pooled into the optimal test t^* (where $j = n+1$ indicates the case that $t^* \cap ([n] \setminus N_{1/2}) = 0$). Hence, we can

apply Lemma B.5 with the given j by setting $i = h$ and thus $\bar{t} \subseteq N_{1/2}$. Since for each $\ell \in N_{1/2}$, $q_\ell \geq 1/2$, we have that $\max_{\ell \in \bar{t}} q_\ell \cdot u_\ell \geq 1/2 \max_{\ell \in \bar{t}} \cdot u_\ell$. Thus, we can choose κ such that

$$\kappa = \frac{\varepsilon \cdot 1/2 \cdot \max_{\ell \in [\bar{t}]} u_\ell}{n} \leq \frac{\varepsilon \cdot \max_{\ell \in [\bar{t}]} q_\ell \cdot u_\ell}{n} \leq \frac{\varepsilon \cdot q_{t^*} \cdot \sum_{i \in t^*} u_i}{n},$$

where the last inequality follows from optimality of t^*.

Now we show that Algorithm 1 is indeed an FPTAS for the optimal test. First note that if $|t^*| = 1$, then Algorithm 1 finds the optimal test in Lines 5-7. Hence, we focus on the case that $|t^*| > 1$. Using Lemma A.4, we distinguish between two cases.

Case I: $|t^* \setminus N_{1/2}| = 0$. For each given $\varepsilon > 0$, κ satisfies the supposition of Lemma B.5. So following Lemma B.5 with $\bar{C} = \bar{C}(h) = \sum_{\ell \in N_{1/2}} \geq \sum_{\ell \in t^*} \hat{u}_\ell$, for $\hat{C}_{h,n+1}$ and $\hat{L}_{h,n+1}$, we have

$$\begin{aligned} u(t) = \hat{z}_\kappa(h, n+1) = \kappa \cdot \hat{C}_{h,n+1} \cdot \hat{P}(h, \hat{C}_{h,n+1}, \hat{L}_{h,n+1}) &\geq (1-\varepsilon) \cdot C^* \cdot P(h, C^*, L^*) \\ &\geq (1-\varepsilon) \cdot C^* \cdot P(n, C^*, L^*) \\ &= (1-\varepsilon) u(t^*) \end{aligned}$$

Case II: $|t^* \setminus N_{1/2}| = 1$. Then, for each $\varepsilon > 0$, there is a $j \in [n] \setminus N_{1/2}$ such that the choice of κ satisfies

$$\hat{z}_\kappa(h, j) \geq (1-\varepsilon) \cdot C^* \cdot P(n, C^*, L^*) = (1-\varepsilon) u(t^*)$$

where the inequality follows from Lemma B.5. The algorithm must determine j since it enumerates all elements of $[n] \setminus N_{1/2}$ in the main loop. The complexity of the algorithm is determined by at most n evaluations of Equation (17). Hence the running time is

$$O\left(n^3 \bar{C}\right) \subseteq O\left(n^3 \sum_{\ell \in N_{1/2}} \frac{u_\ell}{\kappa}\right) \subseteq O\left(\frac{n^5}{\varepsilon}\right).$$

B.4 Details on the MILP formulation

B.4.1 Approximating the non-linear constraints

Here we describe how to approximate the exponential constraints and reformulate the logarithmic constraints from the mixed-integer program in Section 4.3.1.

Handling the logarithmic constraints. We can replace (2d) with integer linear constraints as follows. Fix some test $j \in [B]$. Note that z^j takes integral values in the range $[L, U]$, where $L = \min_i u_i$ and $U = G \max_i u_i$. We introduce an indicator vector $\gamma^j \in \{0,1\}^{[L,U]}$ indexed by $k \in [L, U]$ with constraints $\sum_{k \in [L,U]} \gamma^j_k = 1$ and $\sum_{k \in [L,U]} k \cdot \gamma^j_k = z^j$ to encode which value z holds, and ensure $y^j = \log(z^j)$ with the constraint $y^j = \sum_{k \in [L,U]} \log(k) \cdot \gamma^j_k$.

Approximating the exponential constraints. We now describe how to approximate (2b) from above by a piecewise-linear function f using integer linear constraints. Fix some test $j \in [B]$. Note first that we can relax the equality in (2b) to $w^j \leq \exp(l^j)$ without affecting the outcome. The variable l^j takes values between $A = \min_i(\log u_i) + G \min_i(\log q_i)$ and $B = \log(G \max_i u_i) + \max_i(\log q_i)$ (and these values will be generically non-integral). We approximate exp from above by a piecewise-linear function $f : [A, B] \to \mathbb{R}$ with K linear segments. (Here the parameter K is given exogenously.) Partitioning $[A, B]$ into K parts $[c_k, c_{k+1}], k \in [K]$, we define the k-th line segment as the linear function $f_k(x) = a_k x + b_k$ on domain $[c_k, c_{k+1}]$ with slope $a_k = \frac{\exp c_{k+1} - \exp c_k}{c_{k+1} - c_k}$ and residual $b_k = \exp c_{k+1} - a_k c_{k+1}$. Note that the number of

integer variables in the MILP increases with K, so this parameter must be chosen judiciously. Moreover, given a fixed number of segments K, we wish to determine a partitioning of $[A, B]$ that minimizes the approximation error $\varepsilon = \max_{x\in[A,B]}(f(x) - \exp(x))$. In our implementation, we apply binary search techniques to numerically determine the partition of $[A, B]$ such that the error $\max_{x\in[c_k-c_{k+1}]}(f_k(x) - \exp(x))$ is the same for all parts $[c_k, c_{k+1}]$, which minimizes ε. We introduce indicator vectors $\delta^j \in \{0, 1\}^K$ to encode in which part $[c_k, c_{k+1}]$ the value of l^j lies, as well as the vector $v^j \in \mathbb{R}^K$ whose k-th entry agrees with l^t if l^j lies in the k-th part, and is 0 otherwise. This is guaranteed by constraints $\sum_{k\in[K]} \delta^j_k = 1$, $l^j = \sum_{k\in[K]} v^j_k$ and $c_k \cdot \delta^j_k \leq v^j_k \leq c_{k+1} \cdot \delta^j_k, \forall k \in [K]$. Finally, we require that $w^j \leq f_k(l^j)$ for the k-th part $[c_k, c_{k+1}]$ that l^t lies in. This is expressed by constraint $w^j \leq \sum_{k\in[K]} a_k v^j_k + b_k \cdot \delta^j_k$.

Bounding the approximation error. Recall that the piecewise-function f with K segments approximates exp on domain $[A, B]$ from above with error ε. Let $\sigma(x) = \sum_{j\in[B]} \exp(l^t)$ and $\sigma'(x) = \sum_{j\in[B]} f(l^t)$ respectively denote the corresponding objective values of the convex program (2) and the MILP described above for testing x. Let x^* denote an optimal non-overlapping testing, so x^* maximizes σ, and x' be an optimal solution for the MILP. Clearly, x^* and x' are both feasible for both programs and satisfy $\sigma(x') \leq \sigma(x^*)$ as well as $\sigma'(x^*) \leq \sigma'(x')$. By construction of f, we have $\sigma(x) \leq \sigma'(x)$ and $\sigma(x) \geq \sigma'(x) - \varepsilon B$, which implies $\sigma(x^*) \leq \sigma'(x^*) \leq \sigma'(x') \leq \sigma(x') + T\varepsilon$. Here ε is the additive approximation error of f with regard to exp. Hence, $0 \leq \sigma(x^*) - \sigma(x') \leq \varepsilon B$. This allows us to compute a bound on the additive gap between the welfare achieved by the optimal solution of our MILP and the optimal non-overlapping testing.

C Additional figures and tables from experiments

Here we show figures and summary tables for our experiments comparing the MILP and Greedy on the pilot study data with pool size constraint $G = 10$, and on synthetic populations of size $n = 200$ and pool size constraints $G \in \{5, 10\}$. For details on the experiments, we refer to Section 5.3.

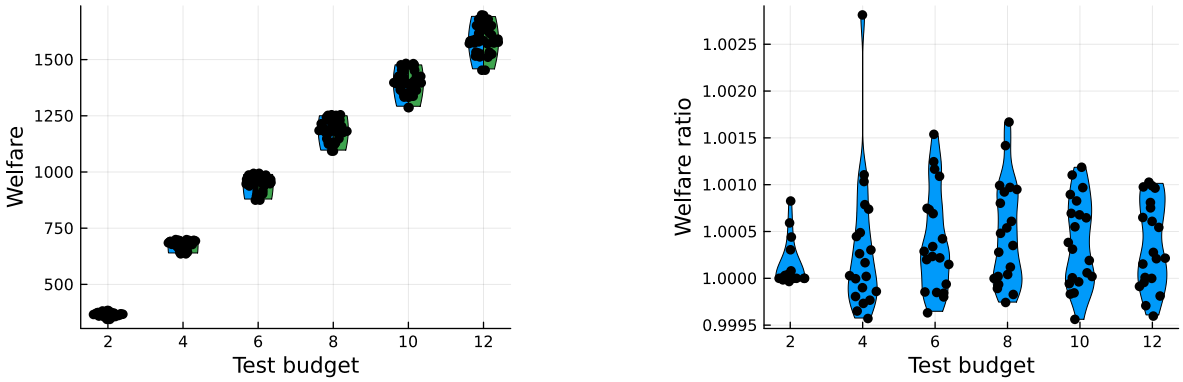

Figure 1: Outcomes of Greedy and the MILP on synthetic data with $n = 200$, pool size bound $G = 5$ and testing budgets $B \in \{2, 4, \ldots, 12\}$. Left: Welfares achieved by the MILP (left regions, blue) and Greedy (right regions, green). Right: Ratios between the welfares of the MILP and Greedy. In both figures, each black dot corresponds to one of the 20 randomly generated populations.

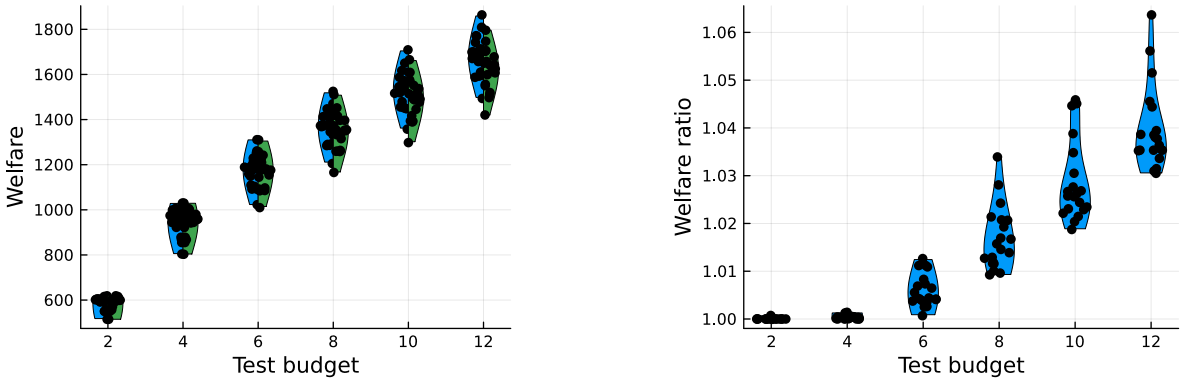

Figure 2: Outcomes of GREEDY and the MILP on synthetic data with $n = 200$, pool size bound $G = 10$ and testing budgets $B \in \{2, 4, \ldots, 12\}$. Left: Welfares achieved by the MILP (left regions, blue) and GREEDY (right regions, green). Right: Ratios between the welfares of the MILP and Greedy. In both figures, each black dot corresponds to one of the 20 randomly generated populations.

	MILP			Greedy		
Budget	Welfare	Guarantee	Time	Welfare	Apx To Optimal	Time
2	866.58	0.47	470 ms	866.47	1.000127	52 ms
6	2391.49	1.42	3504 ms	2390.99	1.000210	50 ms
10	3775.35	2.36	1781766 ms	3774.20	1.000305	87 ms
14	4696.30	3.31	19293052 ms	4678.52	1.003800	279 ms
18	4741.30	4.25	62784039 ms	4678.52	1.013417	672 ms
22	4770.00	5.19	88422845 ms	4678.52	1.019554	1086 ms
26	4790.29	6.13	133956040 ms	4678.52	1.023889	289 ms
30	4805.25	7.08	284989519 ms	4678.52	1.027088	1215 ms
34	4816.15	8.03	545021914 ms	4678.52	1.029417	1175 ms

Table 3: Summary showing welfare and computation time for the MILP and GREEDY on pilot study data (with a population of $n = 130$) and pool size constraint $G = 10$ with testing budgets $B \in \{2, 6, \ldots, 34\}$. We also state the additive approximation guarantee of MILP (compared to optimal non-overlapping welfare).

D The randomized controlled trial

D.1 Randomization

Sufficient separation of treatment and control groups is important for our protocol to work, primarily to disentangle psychological dynamics. If non-treated participants were to run into treated participants, possible health spillovers would be contained within our health protocol: participants have a non-infectious 72 hour window, starting from sample submission, in which they are allowed into the building after receiving a negative qPCR test result. In practice, this translated into a 48 hour window of access. Note that we cannot control offsite socialization.

Many individuals participating in our trial belong to a research 'discipline', and within that discipline, to a working group. Similarly, staff in the administration belong to specific departments, e.g. accounting, which we label as their working group. We followed a clustering approach, randomly assigning the working groups to treatment and control. Because only one group from each discipline volunteered to participate in the trial, discipline and working group

	MILP			Greedy	
Budget	Welfare	Guarantee	Time	Welfare	Time
2	366.90	0.56	285 ms	366.85	35 ms
4	675.51	1.11	837 ms	675.29	67 ms
6	947.86	1.67	2585 ms	947.48	98 ms
8	1188.80	2.23	9933 ms	1188.24	134 ms
10	1399.83	2.78	202447 ms	1399.30	175 ms
12	1584.99	3.34	949195 ms	1584.43	226 ms

Table 4: Experiment summary on synthetic data with pool size bound $G = 5$ and testing budgets $B \in \{2, 4, \ldots, 12\}$. Welfares and times are averaged over 20 randomly generated populations. We also state the additive approximation guarantee of the MILP (compared to optimal non-overlapping welfare).

	MILP			Greedy	
Budget	Welfare	Guarantee	Time	Welfare	Time
2	587.20	1.24	389 ms	587.16	33 ms
4	941.65	2.48	2089 ms	941.34	83 ms
6	1177.26	3.71	9297 ms	1170.46	171 ms
8	1368.17	4.95	33184 ms	1345.93	265 ms
10	1533.75	6.19	75437 ms	1491.34	379 ms
12	1680.95	7.43	350039 ms	1617.25	566 ms

Table 5: Experiment summary on synthetic data with pool size bound $G = 10$ and testing budgets $B \in \{2, 4, \ldots, 12\}$. Welfares and times are averaged over 20 randomly generated populations. We also state the additive approximation guarantee of the MILP (compared to optimal non-overlapping welfare).

are henceforth analogous.

Students, researchers and staff are clustered based on their discipline/working group and each cluster is randomly assigned to treatment or control groups. Treatment and control groups work were spatially separated by different floors or offices, or even buildings. Crucial for this approach to be effective is that individuals across working disciplines are comparable. Given IPICYT's reports about their staff, we know that staff, research students and researchers are assigned to work/study in each of the teams contingent only on their academic discipline, based on no individual characteristics. Hence, we may consider the assignment as good as random. Nevertheless, we further collect a number of covariates to conduct a balance analysis.

Covariate balance. In the Online Appendix (Section 3, Table 2) we provide a covariate balance analysis table. Due to our clustered randomization strategy, it is important to observe whether baseline covariates and outcome variables are balanced across experimental conditions, without accounting for clustering. Of the 15 observed variables, only 2 variables are unbalanced with negligible differences: firstly, the Digital Resources Score[35], is one of the three composite scores with which we construct individual utilities for onsite work and study. The imbalance in this score is also reflected in the distribution of individual utilities per experimental condition. In Figure 1 in Section 2 of the Online Appendix we visualize the overlap between these distributions, and the

[35]This variable is a computed measure of the need for digital and other resources found in the institute.

non-overlapping areas. We notice that the treatment group's distribution exhibits bunching at the right tail, meaning that there are more people in the treatment group that would benefit from working onsite. The inverse is also true, there are more people in the control group that derive less value from working onsite. Recall that the experimental protocol gives more freedom to control group participants and less so to treatment group participants; neutral results (i.e. no difference in endline outcomes) between experimental conditions might, in fact, reflect conservative results, since treatment participants could potentially benefit more from full mobility and no scheduling limitations.

Secondly, the outcome variable 'Productivity' is not balanced, with a higher average productivity in the treatment group. This is likely to have implications for the linear model analyses. Specifically, the models are more likely to be sensitive to positive changes experienced by the control group. Covariate-controlled regressions are included in Appendix D.7 to control for the observed imbalance, and the main results remain unchanged.

Based on the covariate balance discussion above, the randomization process was successful. Some minor imbalance is to be expected as a result of chance (Bruhn and McKenzie, 2009; Altman, 1985), and it is not a likely source of bias for an experimental analysis.

D.2 Scheduling preferences

Our mechanism allowed individuals in the treatment group to indicate their preferences over days on which they wished to access the institute. A negative pooled test on a given day allowed individuals in the pool to access the campus for 72 hours starting from sample submission, which meant 2 days after receiving their result in practice. In the web app, participants were given a set of 10 virtual tokens that they could distribute arbitrarily among all consecutive two-day windows (Monday & Tuesday, Tuesday & Wednesday, etc.) on which they wished to enter the institute. This distribution of tokens then expressed the agent's relative preferences. (Assigning more tokens to some two-day window indicated a stronger preference for these two days.) The individual's utility for each two-day block is then computed from baseline utilities as described in Section 5.2 and their relative preference for the block.

D.3 Outcomes and covariates

Mental health outcomes. Mental health problems related to social isolation as a consequence of the COVID-19 pandemic have been documented for students and the general population (Martinez Arriaga et al., 2021). We conjecture that putting in place a safe education protocol decreases stress levels among students, researchers, and staff, by increasing *safe* sociability (Becchetti et al., 2017) and modulating the perception of health risk in the institute (Shan et al., 2022). Consequently, subjective well-being may also be positively affected. Stress is measured via the validated 4-item Perceived Stress Scale by Sheldon Cohen (Cohen et al., 1994) and we use a variation of the European Quality of Life Survey measure of subjective well-being, using a 'life/subject evaluation' approach (OECD., 2013)[36]

Performance, productivity, and learning. The pandemic has disrupted learning processes and decreased productivity of Mexican students (Limón-Vázquez et al., 2020; Martinez Arriaga et al., 2021) and female researchers (King and Frederickson, 2021). A significant portion of this downfall in productivity may be due to remote work with limited access to the necessary resources for work, research, and learning. We conjecture that our testing protocol improves (self-assessed) productivity and performance (in learning environments), and self-assessed learning experience when compared to a remote work policy. In the presence of an alternative reopening strategy - as is our case - we expect to see no difference between groups. That is, two competing

[36]Note that we use baseline Stress and Subjective Well-being in our utilites' computation. On the other hand, we use endline Stress and Subjective Well-being to analyze between-group intervention effects.

opening strategies that allow all or some individuals in the population to socialize within the institutional premises should increase productivity. Since our mechanism imposes a greater logistical burden on subjects than the status-quo reopening policy of IPICYT, no difference in performance, productivity and learning is an indication that a utility-maximizing approach to partial reopening is a successful strategy.

We use a composite score for the evaluation of performance, productivity. Let $P_{i,z}^{ppa}$ denote the number of points 'achieved' by the answer of subject i to question z pertaining to 'Performance, productivity, and sense of achievement'. Let Z^{ppa} denote the number of relevant questions. Then the score is computed as $p_i = \frac{1}{Z^{ppa}} \sum_z P_{i,z}^{ppa}$. For learning, we use a self-assessment Likert scale that ranges between 1 and 10, where 1 is poor and 10 is excellent.

Covariates. Besides outcome variables, we have collected additional socioeconomic and psychosocial data of participants. These data are used in two ways. First, some of these features enter into the utility estimations needed for the testing algorithm. Second, we use relevant features/variables to check for group balance and, as needed for robustness checks as covariates in our proposed models. We collect the following covariates:

1. Socio-economic attributes: gender, age, ethnicity, educational affiliation, perceived socio-economic status, financial dependents.

2. Academic or job resources: internet access, access to job materials, need to collaborate in person, access to a dedicated working space outside of the office.

3. Psychosocial attributes: Sociability, fear of the virus, subjective well-being.

All covariates and their measurement strategy can be found in the baseline survey in the Online Appendix (Section 4).

D.4 Power and sample size

We estimate statistical power given the five outcome vectors outlined in Appendix D.3. All our outcome variables are continuous scores, where 'perceived stress' is a non-integer vector ranging from 1 to 4, and 'life satisfaction', 'learning', 'productivity', and 'performance' are integer vectors with ranges 1 to 10 for the first two vectors, and 1 to 5 for the remaining three. We perform two types of tests to determine power. We calculate the power of a post-hoc two one-sided (TOST) equivalence test, given that we are interested in observing no difference in outcomes between experimental groups. Equivalence tests are usually a good complement to a null hypothesis test to avoid the misinterpretation of p-values higher than α being considered as evidence of the absence of an effect (Lakens, 2017).

	Lower bound	Upper bound	Equivalence p	Result
Stress	-0.453	0.453	2.44e-02	reject null equivalence
Life satisfaction	-0.822	0.822	0.2.5e-02	reject null equivalence
Performance	-0.371	0.371	2.3e-02	reject null equivalence
Productivity	-0.403	0.403	1.61e-02	reject null equivalence
Learning	-0.742	0.742	2.38e-02	reject null equivalence

Table 6: TOST Equivalence test, using the R package 'TOSTER', assuming ICC ≈ 0.

Table 6 shows the lower and upper bounds for each outcome variable in our analysis. We set the bounds as follows: we use the lower bound of the realized confidence intervals from the null hypothesis tests as the equivalence test lower bound. This grounds our range in an empirically

	Value	Var	P-2.5%	P-97.5%
Stress				
Lohr ρ	-0.0009	0.0183	-0.3069	0.2449
Adj. R^2	0.0083	0.0180	-0.2943	0.2528
ANOVA ρ	0.0081	0.0197	-0.3122	0.2645
Life Satisfaction				
Lohr ρ	-0.1794	0.0146	-0.4573	0.0310
Adj. R^2	-0.1688	0.0142	-0.4414	0.0383
ANOVA ρ	-0.1790	0.0160	-0.4688	0.0399
Performance				
Lohr ρ	0.0243	0.0170	-0.2371	0.2721
Adj. R^2	0.0333	0.0166	-0.2237	0.2793
ANOVA ρ	0.0333	0.0182	-0.2371	0.2875
Productivity				
Lohr ρ	0.0378	0.0128	-0.1860	0.2511
Adj. R^2	0.0464	0.0125	-0.1713	0.2566
ANOVA ρ	0.0475	0.0135	-0.1809	0.2638
Learning				
Lohr ρ	0.0501	0.0162	-0.1959	0.2969
Adj. R^2	0.0587	0.0158	-0.1818	0.3041
ANOVA ρ	0.0603	0.0171	-0.1937	0.3143
Own Goals				
Lohr ρ	0.1472	0.0165	-0.0954	0.4021
Adj. R^2	0.1550	0.0163	-0.0846	0.4077
ANOVA ρ	0.1601	0.0172	-0.0917	0.4157
Supervisor Goals				
Lohr ρ	0.1075	0.0177	-0.1429	0.3688
Adj. R^2	0.1156	0.0175	-0.1312	0.3746
ANOVA ρ	0.1189	0.0186	-0.1411	0.3821

Table 7: Intracluster correlation coefficients for all outcomes computed with the R *fishmethods* package. We report the Pearson correlation coefficient between pairs (Lohr, 2021), an adjusted R^2, and the ANOVA ρ, a coefficient based on a one-way random effects model (Donner, 1986) (variance estimates are bootstrapped).

observed and statistically probable value. We set the upper bound as a full unit increase from the lower bound, as we expect any changes in means to be positive, i.e. our equivalence bounds are set more conservatively than what we observed in the null hypothesis test.[37] In all cases, the equivalence p-values are statistically significant at $p < 0.02$, and we can reject the null hypothesis of the TOST equivalence test.

The Online Appendix (Section 3, Table 1) includes the means and standard deviations per experimental condition and outcome vector that were used to compute the equivalence bounds.

Our randomization approach relied on the affiliation of experiment participants to a working group. Our post-hoc power calculations do not include an intra-cluster correlation coefficient (ICC) parameter based on the realized ICCs in our sample, shown in Appendix D.4. We calculated the ICC coefficient with the 'fishmethods' R package. The function computes the Pearson correlation coefficient between pairs (Lohr, 2021), an adjusted R^2, and a coefficient based on a one-way random effects model (Donner, 1986) (variance estimates are bootstrapped). The cluster variable used in the randomization (and for computation of the ICC) is each subject's working group. Appendix D.4 includes the estimated ρ value for all outcome vectors. All ICC scores are $\rho < 0.2$, too low to be considered reliable (Koo and Li, 2016).

Attrition. We observed two sources of missingness: non-systematic missing values across survey variables, for a total of 24 missing values spread across the dataframe at baseline, and 34 at endline. We further note that between baseline and endline, there was a total of 8 trial attriters, equivalent to 6 percent of our sample. We evaluated the relationship between attriters and experimental conditions using a logistic regression model. With $p = 0.065$, attrition is uncorrelated to treatment assignment. Note that here as for the rest of the analysis, statistical significance is only considered for $p < 0.05$.

[37] There is no theoretically-informed way to specify equivalence bounds, so we aimed to set realistic, robust, and empirically informed bounds.

D.5 Metrics and methods

We propose a two-group experimental design where $n = 130$ subjects are randomly assigned to either a treatment or a control group, conditional on some affiliation to a cluster. Group balance in observed and unobserved heterogeneity is a direct result of random assignment, allowing for treatment status to be the only source of exogenous variation. As such, the mean group difference in the outcomes of interest can be presented as the causal effect of our testing strategy in those those dimensions. We estimate the average treatment effect on the five, previously introduced, outcome variables. We are interested in non-significant differences between experimental conditions, given that the control group is *not* defined as participants working remotely, but participants working onsite with complete freedom of movement. We therefore refer to the p-values in bivariate regressions as evidence of no association.There is an ongoing debate over whether one can use insignificant p-values as evidence of no effect (Lakens, 2021). We resort to the equivalence tests in Table 6 as robustness checks for our findings.

We previously explain that, while we hope to deliver an ATE, we are likely to deliver ITT results, based on the assumption that some treatment participants may not fully comply with the protocol by, for instance, not attending an invitation for saliva sample submission. The protocol is designed such that opting out of sample submission does not affect results, as the grouping algorithm (for test pools) is run only on the subsample of compliers. The non-attendee is simply restricted from entering the premises until they get a negative test result, and they are not penalized when generating new sample submission invitations.

We denote our outcomes as $y_i \in \{s_i, w_i, l_i, p_i, pr_i\}$:

- The average stress level and subjective well-being, measured by each individual's stress score s_i and life satisfaction score w_i
- Subjects' self-assessed learning l_i, performance p_i and productivity scores pr_i.[38]

The treatment effect of endline outcomes is estimated using a linear model, with HC1 standard errors. Let Y denote the stacked vector of outcomes $(y_1, \ldots, y_n) \in \{s_i,\ w_i,\ l_i,\ p_i,\ pr_i\}$. Let β denote the vector of parameters to be estimated, and τ_i the treatment dummy. The independent variables are subsumed in $X = (1^n, \mathrm{T}, C)$, where 1^n is an $n-$vector of ones, T is a vector of treatment status τ_i, and C is a matrix of covariates[39]. Let ε denote the vector of error terms ε_i. We estimate the model

$$Y = X\beta + \varepsilon \tag{18}$$

for $Y \in \{S, W, P, Pr, L\}$ and test the hypothesis $\beta_1 \approx 0$. We collect baseline and endline data for the set of outcome vectors. Let $\Delta Y = Y_{endline} - Y_{baseline}$ denote the change in outcome Y from baseline to endline. We estimate the *delta model*

$$\Delta Y = X\beta' + \varepsilon' \tag{19}$$

to identify the effect of our intervention on the change in outcomes throughout the duration of the experiment. This analysis complements the analysis of endline outcomes: it eliminates all observed and unobserved confounds that are constant between our two points of measurement (Allison, 1990). This allows for the interpretation of results not only as static differences but also in the context of possible outcome trajectories, and adjusted for static unobservables.

[38] We adapt a measure based on the fit of 10 and 5 point likert scales, respectively, as per Versteeg and Steendijk (2019)

[39] We present covariate-controlled linear models in the Appendix as robustness checks. Theey are not, however, part of the main analysis.

D.6 Results

We present the results from the regression analyses based on Eqs. (18) and (19) in Tables 2, 8 and 9. On a high level, we are able to report for performance outcomes (in Table 2, in the main text) as well as mental health outcomes (Table 9) that our testing protocol has no negative effect, despite the increased effort it demands from participants.

To check the robustness of these outcomes, we estimate delta models for performance, productivity, and learning (also in Tables 2 and 8). All treatment effects are corrected downwards,[40] but remain non-significant with one exception. At $p = 0.05$, the change in treatment participants' productivity from t_0 to t_1 is at the border of statistical significance. It shows a downward effect of 0.25 score points. Mean self-reported productivity for treatment participants went from 2.27 at t_0, down to 2.17 at t_1. This small decrease of 0.1 may be due to the added coordination effort exerted by treatment participants, and time invested in familiarizing themselves with the protocol. On the other hand, control participants experienced an increase in mean self-reported productivity of 0.09 points; they went from 2.00 at t_0 up to 2.09 at t_1. This small increase in productivity may be a benefit from transitioning from remote work to full institutional access. Together, they explain the negative and borderline statistically (in)significant coefficient.

Mental health outcomes. Table 9 shows that there are no significant effects of the pooled-testing protocol on the subjects' stress level or subjective well-being (life satisfaction). On average, subjects in the treatment group report a stress score that is 0.186 points higher than for subjects in the control group. At $p = 0.17$, this difference is not statistically significant. When looking at the change in stress from baseline to endline in the delta bivariate model, the magnitude of the coefficient decreases to 0.009 at $p = 0.95$. That is, treatment status induces little to no variation in the change in stress between t_0 and t_1. Further, treatment participants report higher average life satisfaction scores. At endline, the difference in scores is small at 0.089, and insignificant ($p = 0.81$). However, the magnitude of the coefficient drastically increases for treatment participants by 0.284 points when looking at the change in scores pre and post trial. The change in life satisfaction score is, again, statistically insignificant ($p = 0.37$).

	Dependent variable:			
	Own goals	Supervisor goals	Δ Own goals	Δ Supervisor goals
Treatment	0.035	0.250	−0.307	−0.115
	(0.170)	(0.159)	(0.192)	(0.191)
Constant	2.344***	2.129***	−0.131	−0.113
	(0.107)	(0.099)	(0.101)	(0.098)
Observations	119	120	118	119
R^2	0.0004	0.021	0.022	0.003
Adjusted R^2	−0.008	0.012	0.014	−0.005

Sig. p codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Table 8: Linear model regressions of further performance outcomes.

[40] During the course of the month, treatment participants experience a small and statistically insignificant decrease in performance and goals; however, they still report higher learning, on average.

	Dependent variable			
	Overall stress	Life satisfaction	Δ Overall stress	Δ Life satisfaction
Treatment	0.186	0.089	0.009	0.373
	(0.135)	(0.371)	(0.146)	(0.414)
Constant	2.238***	7.556***	−0.813***	−0.270
	(0.093)	(0.259)	(0.096)	(0.276)
Observations	121	122	121	121
R^2	0.016	0.0005	0.00004	0.007
Adjusted R^2	0.007	−0.008	−0.008	−0.002

Sig. p codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Table 9: Linear model regressions of mental health outcomes.

D.7 Covariate analysis and secondary results

We also ran covariate-controlled regressions as robustness checks for our bivariate models (see the Online Appendix, Section 3, Tables 3 and 4), and find no contradictions to our main findings in our preferred model specification. All covariates are taken from the baseline survey, whereas the dependant variables are taken from the endline survey. We find that baseline utilities are strongly positively correlated with most measures of performance, except learning, suggesting that the higher the need to work onsite, the more productive the individual is (at endline). Interestingly, learning is negatively correlated with all performance measures. Further exploration is beyond the scope of this study.

D.7.1 Distributions of outcome variables

We chose to evaluate the testing allocation protocol with linear models, under the assumption that all outcomes were normally distributed and with the understanding that our interest was not in the magnitude of the coefficients but in mean equivalence between experimental conditions. Subjective well-being and stress are both continuous and normally distributed variables. Subjective well-being is coded as a likert scale that ranges from 1 to 10, and stress is a numeric score that goes from 1 to 5, including fractions. Learning is also normally distributed, and coded as a 1 to 10 likert scale. Performance and Productivity, however, are coded as 1 to 5 categorical scales. In practice, the last category (coded as 5) was never picked in the survey. Due to the integer restriction of these outcome variables, their distribution is not fully normal. To address this issue, we estimate multinomial logistic bivariate models (see the Online Appendix, Section 3, Tables 5 and 6) as robustness checks. The multinomial models corroborate our linear models' results for performance and productivity. We evaluated productivity and performance as aggregate scores instead of ordinal scores for comparability purposes across the set of outcomes, but nevertheless provide a model that better fits the outcomes variables' functional form. We note that we prefer a multinomial over an ordered model due to our interest in behavior relating to a reference point, the 'average', and not in the explicit increasing order in the categories.