

Patterns

Q-SID: Detecting collusion groups from exam question scores with error quantification

Highlights

- Q-SID detects likely collusion groups using graded numeric question scores
- Q-SID applies to multiple-choice and written exams and accommodates partial-credit scoring
- Empirical and synthetic nulls provide complementary estimates of false-positive rates
- Q-SID is more accurate and faster than existing collusion-detection methods

Authors

Guanao Yan, Jingyi Jessica Li,
Mark D. Biggin

Correspondence

jli@stat.ucla.edu (J.J.L.),
markbiggin@sbcglobal.net (M.D.B.)

In brief

Online exams make coordinated answer sharing difficult to detect, while existing collusion-detection methods typically focus on suspicious pairs of students and are often restricted to multiple-choice responses. Yan et al. introduce Q-SID, which detects likely collusion groups, including groups larger than two, directly from graded question scores, and provides estimated false-positive rates. Across diverse real exams, Q-SID improves accuracy and speed, giving instructors a scalable screening tool for identifying cases that warrant careful follow-up review.

Article

Q-SID: Detecting collusion groups from exam question scores with error quantification

Guanao Yan,^{1,2} Jingyi Jessica Li,^{1,3,*} and Mark D. Biggin^{4,5,*}

¹Department of Statistics and Data Science, University of California, Los Angeles, Los Angeles, CA, USA

²Department of Computational Mathematics Science and Engineering, Michigan State University, East Lansing, MI, USA

³Biostatistics Program, Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA

⁴Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

⁵Lead contact

*Correspondence: jli@stat.ucla.edu (J.J.L.), markbiggin@sbcglobal.net (M.D.B.)

<https://doi.org/10.1016/j.patter.2026.101642>

THE BIGGER PICTURE Online exams have expanded the access and flexibility of education, but it also makes it difficult to detect cases where students share answers in violation of testing rules. Existing collusion-detection approaches have important limitations, e.g., focusing only on suspicious student pairs or being useful only for multiple-choice exams.

More flexible and rigorous statistical approaches for detecting collusion, like the one introduced in this paper, may help educational institutions screen for possible instances of cheating. The authors caution, however, that groups identified during screening with such tools should undergo a careful review of written answers or other independent evidence before disciplinary decisions are made. Looking ahead, institutions could use tools like this one not only to investigate suspected cheating but also to routinely monitor collusion levels, target high-risk courses for improved assessment design, and communicate clear deterrent signals to students. Even as generative AI reshapes academic integrity challenges and enables new forms of shared external assistance, collusion detection algorithms could also help identify groups that rely on the same AI tools and, therefore, produce unusually similar answers.

SUMMARY

Collusion in online exams undermines the integrity of exam results. Existing detection methods identify suspicious student pairs and are often limited to multiple-choice formats. We introduce Question-Score Identity Detection (Q-SID), a statistical algorithm that efficiently detects likely collusion groups and quantifies uncertainty using only graded numeric question scores. This enables Q-SID to support a broad range of multiple-choice and non-multiple-choice exams. Q-SID reports two false-positive rates (FPRs) for each collusion group: (1) an *empirical FPR*, whose null data are from strictly proctored exam datasets, and (2) a *synthetic FPR*, whose null data are simulated from a copula-based model. Across 34 unproctored exam datasets, including two benchmark datasets with verified positives and negatives from textual analysis, Q-SID performs robustly across exam formats, class sizes, numbers of questions, and levels of complexity. It is more accurate and substantially faster than widely used pairwise statistics and group-level baselines, including answer similarity at group level (ASlg) and clique detection.

INTRODUCTION

Cheating on exams is a long-standing problem in academic testing^{1–3} and has increased with the rise in online testing because rigorous proctoring in these cases is difficult.^{4–8} Many methods have been proposed to detect such collusion and fall into two broad families^{9–25} (Table 1). The first are *pairwise methods* that compute a similarity statistic for each pair of students and flag unusually similar pairs, often using item response

theory (IRT) methods such as the generalized binomial test (GBT²⁵), M4 statistic,^{21,26} and ω and weighted- ω indices.^{27,28} Because collusion in practice often involves groups rather than isolated pairs, a second *groupwise* family aggregates pairwise evidence to infer collusion groups. Representative strategies in this family include clustering of pairwise similarity graphs (e.g., hierarchical or spectral clustering), testing for unusually dense cliques (clique detector [CD]²⁹), or computing group-level tail probabilities (answer similarity at group level [ASlg]³⁰).

Table 1. Summary of selected baseline methods and Q-SID

Method	Level	Statistic	Null/calibration	Input	Output
GBT	Pairwise	Identical option matches	NRM option probs; generalized binomial	MC options	Pair statistic; pair p value
M4	Pairwise	Matches on correct/incorrect	NRM; bivariate generalized trinomial	MC options	Pair statistic; pair p value
Ω	Pairwise	Directed LR (copier $\rightarrow$ source)	IRT likelihood	MC options	Directed pair score; pair p value
Weighted- ω	Pairwise	Weighted directed LR	IRT likelihood + item weights	MC options	Weighted directed score; pair p value
ASlg	Group	Group summary of pairwise stats	MVN approx; group tail probability	Pairwise stats	Group statistics; group p values
CD	Group	Group summary of similarity graph	Permutation	Similarity graph	Group statistics; group p values
Q-SID	Group & pairwise ^a	Collusion score	Empirical null; synthetic null	Graded question scores (MC & non-MC)	Group statistics; empirical FPR; synthetic FPR

GBT, generalized binomial test; ASlg, answer similarity at group level; CD, clique detector. NRM, nominal response model; MC, multiple choice; LR, likelihood ratio; IRT, item response theory; and MVN, multivariate normal.

^aQ-SID provides collusion pairs as intermediate outputs, with collusion groups as final outputs.

Despite their utility, prevailing methods share several limitations. First, many methods are developed for multiple-choice exams and often require binarization of question scores, which discards information in partial-credit, ordinal, or continuous question scores and can lead to miscalibration. Second, these methods' *true-positive rates* (TPRs), defined as the fraction of colluding students who are detected, are difficult to quantify because they were only tested on multiple-choice exams, for which solid evidence of collusion could not be obtained through the classic approach of textural analysis of written answers. Third, some methods report error rates for individual student pairs, not for groups assembled from those pairs, and are vulnerable to chain-linking, where long paths of modest pairwise similarities can appear as a coherent group. Fourth, reliable estimation of false-positive rates (FPRs) remains challenging because existing approaches often depend on parametric IRT assumptions and lack validation against independent no-collusion data. Fifth, computational scalability remains challenging because forming collusion groups from pairwise similarities can require combinatorial search (e.g., clustering or clique detection), which becomes costly as class size grows.

To address these limitations, we curated a collection of exam datasets and developed the *Question-Score Identity Detection* (Q-SID) framework (algorithm in the section "Q-SID algorithm" under "Methods"). Q-SID operates directly on graded question scores and is thus applicable to multiple-choice, partial-credit/ordinal, and free-response exams without binarization. It performs group-level inference by outputting collusion groups with associated uncertainty, and it can, when needed, summarize results at the pair level. For calibration, Q-SID reports two complementary, exam-specific FPRs: an *empirical FPR* estimated from strictly proctored in-person exams and a *synthetic FPR* obtained by simulating *in silico* examinees from an exam-specific multivariate null fitted to the question-score distribution (e.g., via copula or Gaussian

approximations). The agreement between these two FPRs supports stable false-positive control across exams with different sizes, numbers of questions, and score structures. In addition, by avoiding combinatorial searches over candidate collusion groups after computing pairwise student similarities, Q-SID scales to large cohorts and provides outputs that support transparent operational use.

We demonstrate Q-SID's performance through a real-exam benchmark and three real-world applications. In a benchmark on two exams with collusion groups independently confirmed from written work (see section "The data" under "Methods"), Q-SID recovered the verified collusion structure more accurately and with substantially lower runtime than pairwise methods GBT, M4, ω , weighted- ω , and groupwise methods ASlg and CD for aggregating pairwise similarity statistics into the group level. In addition, we (1) estimate collusion prevalence as the proportion of students assigned to groups whose collusion scores (CSs) exceed thresholds set by either an empirical or a synthetic FPR; (2) show that Q-SID is effective as an initial screening step for unproctored online exams, with flagged students typically corroborated by follow-up evidence (e.g., written-answer similarity, confessions, and disciplinary outcomes); and (3) present evidence of deterrence, with fewer students assigned to collusion groups on exams where students were warned about Q-SID than on comparable exams without such warnings. To facilitate adoption, we provide a free public web tool (<http://shiny2.stat.ucla.edu/Q-SID/>) that accepts question- and test-score data and returns a report listing detected groups, their CSs, and both empirical and synthetic FPRs. The R code is also openly available.

Finally, the rapid adoption of generative AI has further changed the academic-integrity landscape by enabling students to obtain unauthorized assistance without directly copying from their peers.^{31,32} We discuss the implications of the widespread use of AI in cheating for the ongoing utility of Q-SID.

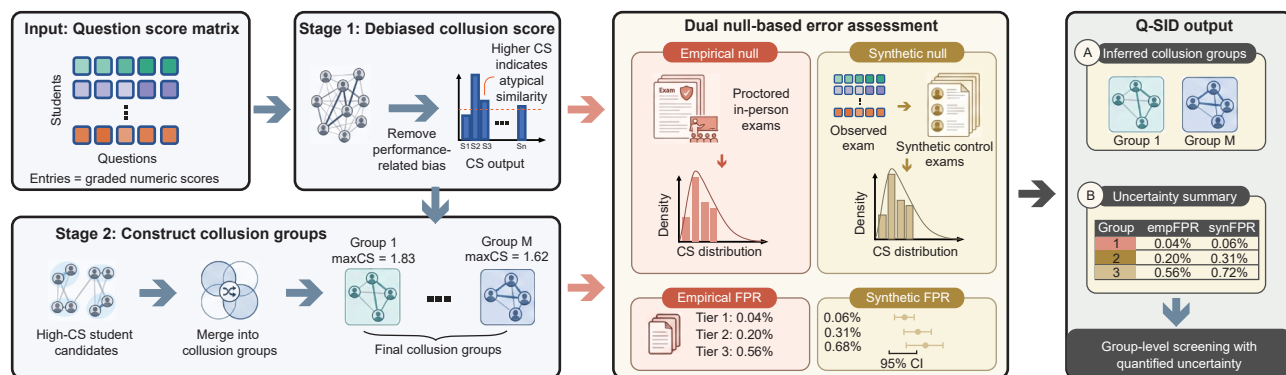


Figure 1. Q-SID analyzes a question score matrix (students by questions) in two steps

First, it assigns each student a collusion score that measures unusually high similarity after accounting for overall performance patterns. Second, it forms collusion groups from high-score signals and summarizes each group by its maximum collusion score. To quantify uncertainty, Q-SID uses two null references: an empirical null from strictly proctored in-person exams and an exam-specific synthetic null from synthetic control exams that preserve key properties of the observed score data under a no-collusion setting. Q-SID reports empirical and synthetic false-positive rates to support interpretation of the detected groups.

RESULTS

Overview of Q-SID for collusion-group detection with uncertainty estimates

We developed and evaluated Q-SID using data from 70 exams administered at UC Berkeley and UCLA (see section “the data” under “Methods,” Table S1, and Data S1 and S2). These exams serve complementary roles in method development and evaluation: (1) two *benchmark* unproctored online exams for which collusion status (true-positives and true-negatives) was independently established through manual textual analysis of written work, used in part to estimate the TPR; (2) 36 strictly *proctored* in-person exams in which collusion is not expected, used to characterize the empirical null behavior of Q-SID and to estimate empirical FPRs; and (3) 32 additional *unproctored* online exams collected during the COVID-19 pandemic, used to illustrate performance in operational settings where proctoring was ineffective and collusion likely.

Given an exam’s *question score matrix* $\mathbf{X} \in \mathbb{R}_{\geq 0}^{n \times p}$ (rows are students, columns are questions, and entries are graded numeric scores), Q-SID produces two primary outputs (Figure 1): (1) a collection of inferred *collusion groups* $\{\mathcal{G}_m\}_{m=1}^M$, each summarized by a group-level maximum CS $\max\text{CS}_m$, and (2) uncertainty measures for these group candidates in the form of *empirical* and *synthetic* FPRs (see section “Q-SID methodology” under “Methods” for details). Conceptually, Q-SID proceeds in two stages. First, it assigns each student a *debiased* CS that captures unusually high similarity to other students after accounting for systematic performance-related similarity. Second, it constructs collusion groups from high-CS students using two thresholds c_1 and c_2 ($c_2 > c_1 > 0$) and reports $\max\text{CS}_m$ as a group-level summary indicating severity of collusion. To support interpretation, Q-SID reports empirical FPRs estimated from proctored control exams and synthetic FPRs estimated from synthetic control exams that represent a no-collusion scenario while preserving key properties of the observed question-score data. Together, these outputs provide group-level error calibration.

Q-SID outperforms existing methods in collusion-group detection

To evaluate the performance of collusion group detection by Q-SID, we conducted a benchmark study using the two *benchmark datasets* (see section “benchmark datasets with true-positives and true-negatives” under “Methods”) for which collusion groups had been independently identified based on the written answers. Using the corresponding question-level score matrices, we compared Q-SID with three classical pairwise collusion statistics: the GBT,²⁵ Maynes’ M4 similarity statistic (M4^{21,26}), and a weighted- ω statistic,²⁸ an extension of the classical measurement ω .²⁷ These existing methods were originally developed for nominal or binary responses (Table 1), such as multiple-choice answers, rather than for numerical question scores. Nevertheless, we adapted them to the question-score setting to enable a direct comparison with Q-SID (see section “comparison with existing methods,” which is specifically designed for graded question scores). At the group level, we considered two representative groupwise detection methods: ASig³⁰ and a clique-detection baseline (CD²⁹), each built on the underlying pairwise statistics. Several group-level extensions have been proposed in the literature.³³ We chose CD and ASig because they represent two widely used and complementary paradigms for translating pairwise evidence into collusion groups while remaining comparable across underlying pairwise statistics: CD provides a simple, graph-theoretic baseline that flags unusually dense cliques in the pairwise similarity graph,³³ whereas ASig explicitly assigns a *group-level* significance measure and was shown to mitigate chain-linking-driven false-positives relative to more ad hoc clustering procedures.³⁰

To evaluate performance, we used three complementary metrics. First, we quantified the overlap between detected collusion groups and the known collusion groups using the Jaccard index, which measures the size of the intersection of two sets divided by the size of their union; a value of 1 indicates identical group membership, whereas a value of 0 indicates no shared students. Second, we assessed clustering agreement via the adjusted Rand index (ARI), which compares two partitions of all students

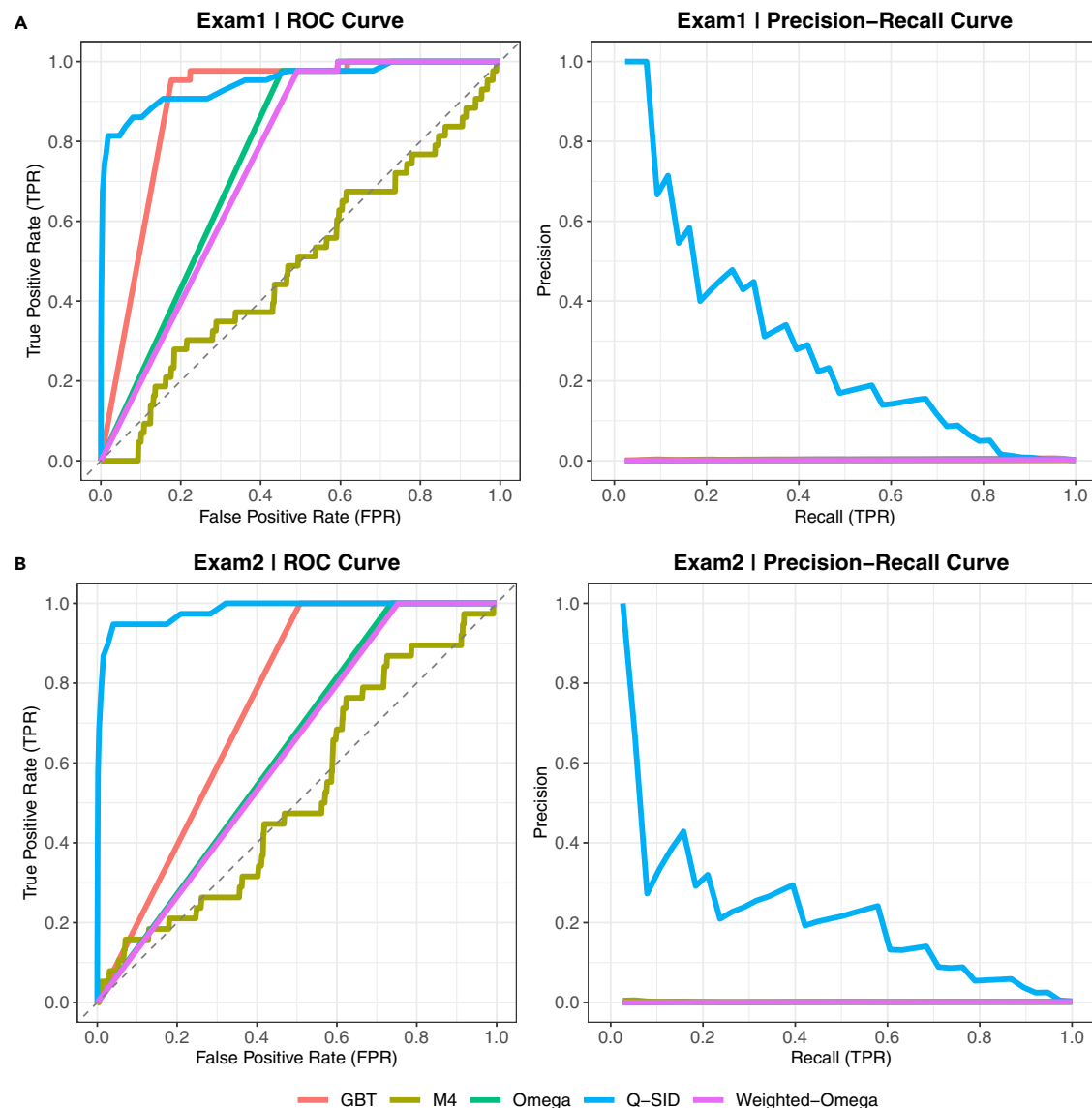


Figure 2. Comparison of methods' accuracy in identifying student pairs known to have colluded

(A) Pairwise collusion detection benchmark on benchmark exam 1 and (B) benchmark exam 2. Receiver operating characteristic (ROC) and precision-recall curves compare Q-SID with the generalized binomial test (GBT), M4, ω , and weighted- ω for detecting colluding student pairs. Higher ROC curves (toward the upper left) and higher precision-recall curves indicate better ability to rank true colluding pairs above non-colluding pairs. Q-SID attains higher true-positive rates for a given false-positive rate and maintains higher precision across a broad range of recall values than other methods.

(detected groups versus true groups) and is scaled so that 0 corresponds to agreement no better than random and 1 corresponds to perfect agreement. Third, we conducted a student pair-wise analysis by converting each true collusion group into all student pairs (e.g., a true group $\{1,2,3\}$ becomes pairs $\{(1,2),(1,3),(2,3)\}$) and ranking pairs for each method. For Q-SID, this ranking was based on the identity score; for the baseline methods, it was based on the respective test statistics. Taken together, these metrics provide a coherent assessment of how well each method recovers collusion groups.

Across both exams, Q-SID outperformed the baseline approaches at both the pairwise and groupwise levels. At the pair-

wise level (Figure 2), the receiver operating characteristic (ROC) curves indicate that Q-SID consistently attains higher TPRs than the GBT, M4, ω , and weighted- ω for any given FPR. The curves for the classical statistics lie close to the diagonal, whereas the curve for Q-SID is much closer to the upper-left corner, indicating superior discrimination. The precision-recall curves show a similar pattern: Q-SID maintains higher precision over the full range of recall values, whereas the baseline methods produce precision values that are indistinguishable from zero across the entire range, reflecting limited ability to recover true colluding pairs without incurring substantial numbers of false-positives.



Figure 3. Group-level detection performance on the two benchmark exams

The top panels report the adjusted Rand index (ARI), and the bottom panels report the Jaccard index, quantifying agreement between detected collusion groups and empirically verified groups. Q-SID achieves high ARI and Jaccard values in both exams, indicating strong recovery of the true collusion structure, whereas all ASlg- and CD-based implementations (ASlg-GBT, ASlg-M4, ASlg- ω , CD-GBT, CD-M4, and CD- ω) yield values close to zero.

At the groupwise level (Figure 3), Q-SID recovers collusion groups that agree well with the empirically verified groups. For both exams, the Jaccard index and ARI for Q-SID are high, indicating substantial overlap and clustering agreement between the detected groups and the known collusion structure. In contrast, ASlg- and CD-based implementations yield negligible overlap and agreement with the ground truth on the same question-score data (with only a small, nonzero overlap for ASlg-M4), resulting in Jaccard and ARI values that remain near zero across both exams.

Q-SID is more computationally efficient than the competing methods (Figure 4). For each exam, Q-SID required approximately 1 s of wall-clock time to compute the relevant statistics and form collusion outputs, whereas the pairwise methods GBT, ω , and weighted- ω each required about 150–160 s and M4 required over 550 s. For aggregating pairwise statistics into groupwise statistics, the ASlg-based procedures were slow, with total runtimes on the order of 1,000–1,040 s per exam. The CD-based procedures were comparably expensive (and in exam 1, slower), requiring on the order of 10^3 seconds per exam across the underlying pairwise statistics. These results

indicate that Q-SID offers orders-of-magnitude computational savings relative to existing pairwise and group-wise methods.

To summarize, in realistic exam settings with empirically verified cheating, Q-SID achieves higher sensitivity and closer agreement with the known collusion groups than existing pairwise similarity statistics and their ASlg and CD groupwise extensions, while also being more computationally efficient. Because it operates directly on numerical question scores and produces interpretable group-level error calibration at modest computational cost, Q-SID is well suited for application to large classes and routine operational use. We acknowledge that this implementation may not represent the optimal adaptation of these existing methods, which were designed for multiple-choice exams, to our setting. The results should therefore be interpreted only as evidence that applying existing methods to non-multiple-choice exams is not straightforward.

Proctored control exams support empirical false-positive rate

For any collusion detection method, a challenge is to obtain an accurate estimate of the FPR for each detected collusion group.

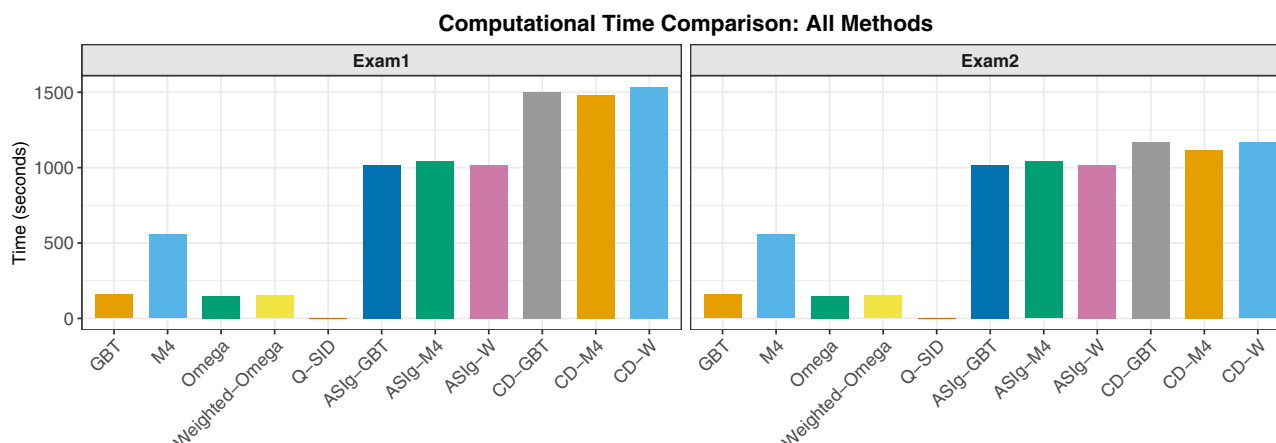


Figure 4. Computational time comparison for all methods on two benchmark exams

Bars show time (in seconds) required to compute the detection statistics and form collusion outputs for each method. Q-SID is faster than the classical pairwise methods (GBT, M4, ω , weighted- ω) and is substantially faster than the ASig- and CD-based group-level procedures (ASig-GBT, ASig-M4, ASig- ω , CD-GBT, CD-M4, and CD- ω).

If the estimate is poor or varies with the exam format, it would not be trusted by instructors, students, or academic administrators. To enhance the statistical rigor for collusion detection, Q-SID uses two types of null data (i.e., control exams) to evaluate the FPR of each detected collusion group: (1) empirical null data (i.e., proctored exams) and (2) synthetic null data (i.e., synthetic control exams). In this subsection, we report results for the empirical null and the resulting empirical FPR estimates.

Q-SID uses strictly proctored in-person exams (see “[proctored exam datasets](#)” section under “[Methods](#)”) as an empirical null reference. Because these exams were administered under conditions in which copying detailed written responses is unlikely, students flagged by Q-SID in the proctored exams can be treated as false-positives for estimation purposes. This assumption is supported by the observed score distributions: the CS distribution from the proctored control exams has a much lighter right tail than the distribution from an unproctored benchmark exam with verified collusion (Figure 5A). This contrast indicates that the proctored exams provide an appropriate reference for the null tail behavior of CS under non-colluding conditions.

Building on this empirical null reference, Q-SID assigns detected collusion groups to three error-risk tiers using group-level maxCS thresholds (see “[detection of collusion groups](#)” section under “[Methods](#)”). Across 36 proctored exams comprising 10,816 students, the estimated cumulative empirical FPRs for these tiers are 0.04%, 0.20%, and 0.56% (Table 2). These values imply that about 0.04% of students fall into the most stringent tier, an additional 0.16% fall into the intermediate tier, and an additional 0.36% fall into the least stringent tier, for a total of 0.56% across all three tiers. The corresponding 95% confidence intervals are shown in Table 2.

Taken together, these results show that proctored control exams provide an empirical basis for estimating Q-SID false-positive behavior and for assigning detected groups to interpretable FPR tiers. This empirical reference also establishes a baseline against which exam-specific synthetic FPR estimates can be compared in the next subsection.

Synthetic control exams provide exam-specific FPRs

Empirical FPR estimates from proctored exams provide a reference under non-colluding conditions, but they do not span the full heterogeneity of submitted exams, which vary in class size, number of questions, score scales, and total-score distributions. Q-SID, therefore, supplements the empirical null with an exam-specific null constructed from *synthetic control exams* generated from the submitted exam itself (see “[synthetic control data generation](#)” section under “[Methods](#)”). The validity of this synthetic null rests on the data-generating mechanism. In the synthetic generation step, students are treated as observations and questions as features, and synthetic students are sampled independently and identically from a fitted unimodal joint distribution of question scores (implemented within score-stratified cohorts). This construction preserves population-level score patterns, while removing student-specific dependence structures. In particular, it does not retain pairwise or groupwise similarities tied to specific examinees, which are the signal of interest for collusion detection. For this reason, the synthetic controls provide a null reference for false-positive assessment that is tailored to the submitted exam but does not encode collusive behavior.

A second requirement is realism: the synthetic controls must resemble the observed exam sufficiently well to serve as a meaningful baseline. Across multiple diagnostics, the synthetic controls closely match the statistical structure of the corresponding real exams. At the question level, they reproduce score-frequency profiles across questions (Figure 6) and yield similar distributions of per-question means, variances, and coefficients of variation (Figure 7). At the dependence level, they preserve the overall question-question correlation structure observed in real exams, which is often positive because students tend to perform consistently well or consistently poorly across questions (Figure 7). In addition, low-dimensional embeddings of real and synthetic students show substantial mixing, quantified by the median integration local inverse Simpson’s index (miLISI).³⁴ The miLISI ranges from 1 to 2, with larger values indicating stronger mixing; the values are 1.938 for the benchmark exam 1 and

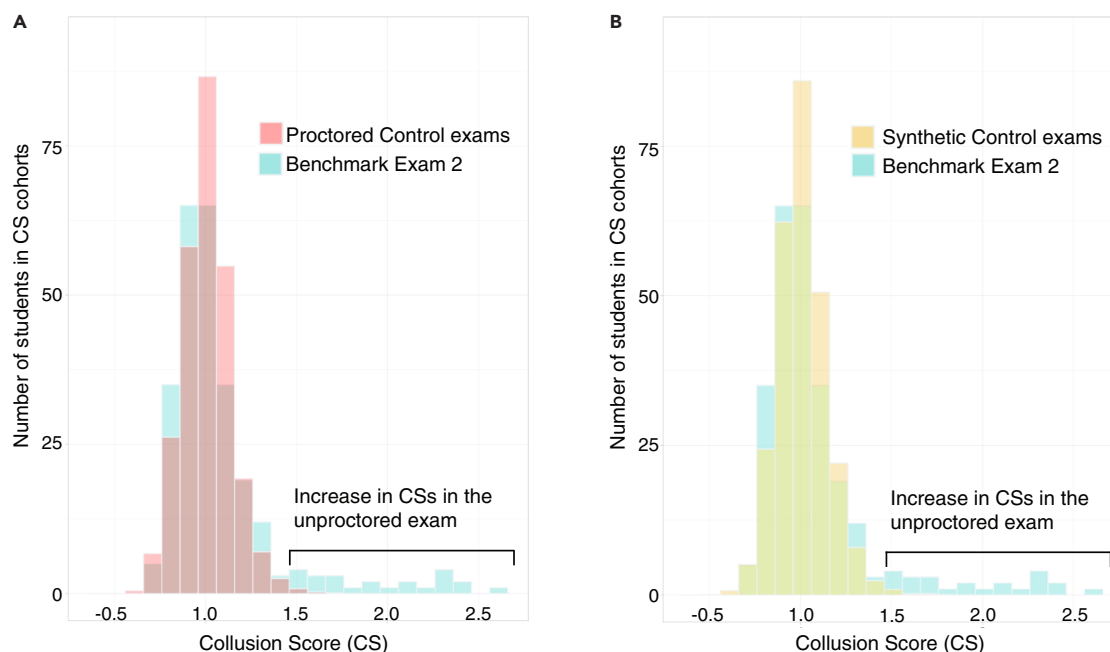


Figure 5. The distributions of CS values from the 19 proctored control exams (red), the unproctored benchmark exam 2 (blue), and the synthetic control exams (yellow)

The distribution of CSs from the benchmark exam 2 has a heavier right tail than the corresponding distributions for both the proctored control exams (A) and the synthetic control exams (B).

1.9544 for a representative proctored control exam (Figure 7). Taken together, these results support the synthetic controls as an exam-specific null for downstream FPR estimation.

Given these synthetic controls, Q-SID estimates an exam-specific *synthetic FPR* by computing the proportions of synthetic students assigned to the calibrated error-risk tiers (see “synthetic FPR estimation from synthetic control data” section under “Methods”). In contrast to the empirical FPR tiers, which are fixed by construction from proctored controls, synthetic FPRs vary across exams because they reflect the score distribution and dependence structure of the submitted exam. The synthetic FPR estimates for the two benchmark exams are summarized in Table 2.

Empirical and synthetic nulls provide complementary FPRs for Q-SID

A central objective of Q-SID is to provide error estimates that are both interpretable and credible across exam formats. To this end, we compared the two null-based FPRs introduced above: the empirical null derived from proctored exams and the exam-spe-

cific synthetic null derived from the submitted query exam. The CS distribution from synthetic control exams for benchmark exam 2 closely matches the CS distribution from the 19 proctored control exams (Figure 5, compare panels A and B), indicating that the synthetic controls reproduce the null tail behavior relevant to Q-SID calibration. This agreement is important because the two nulls are constructed from different assumptions and data sources; concordance, therefore, supports Q-SID’s FPR calibration.

We next asked whether these calibrated FPRs are informative when Q-SID is applied to unproctored exams. Across the unproctored exams (two benchmark exams plus 32 additional unproctored exams), the proportions of students placed into collusion-group bins substantially exceeded both the empirical FPRs and the synthetic FPRs (Figure 8). This pattern is consistent with true collusion rather than null variation. In addition, for many query exams, the synthetic FPRs were close to the empirical FPR benchmarks and were both far below the observed proportions of flagged students. Approximately half of the flagged students in unproctored exams fell into the most stringent bin (the 0.04% empirical FPR bin), where synthetic FPRs

Table 2. 95% CIs of true FPR for the three empirical FPR bins

Empirical FPR bin	Empirical FPR 95% CI	Synthetic FPR 95% CI (Benchmark exam 1)	Synthetic FPR 95% CI (Benchmark exam 2)
0.04%	0.037% ($\pm 0.036\%$)	0.058% ($\pm 0.001\%$)	0.138% ($\pm 0.002\%$)
0.20%	0.203% ($\pm 0.085\%$)	0.195% ($\pm 0.002\%$)	0.315% ($\pm 0.003\%$)
0.56%	0.555% ($\pm 0.140\%$)	0.623% ($\pm 0.003\%$)	0.697% ($\pm 0.003\%$)

CI, confidence interval.

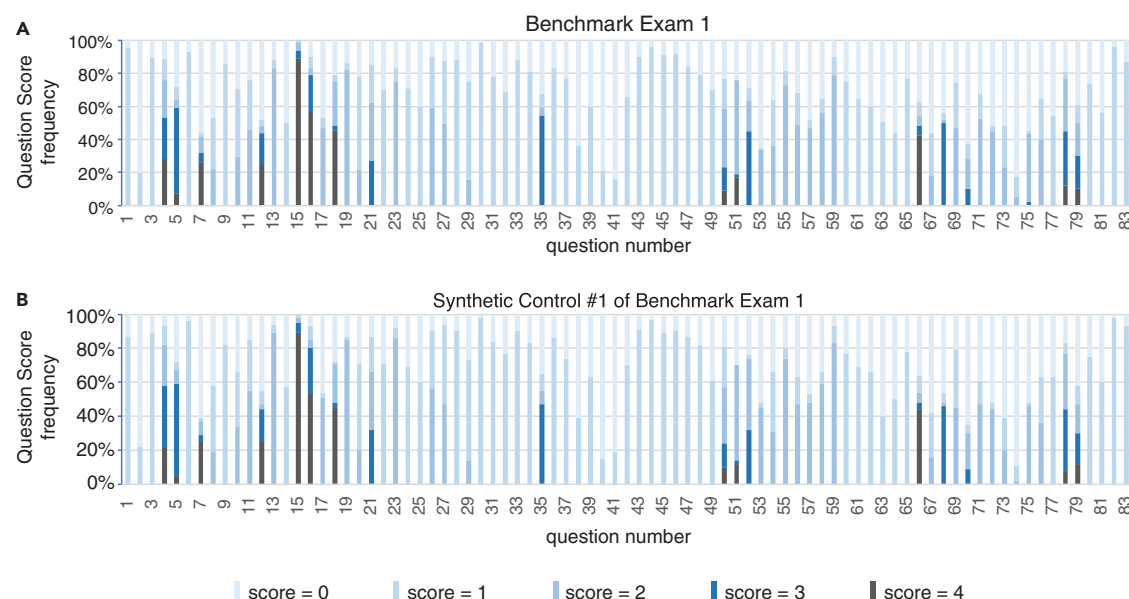


Figure 6. Comparison of question score frequencies obtained by students in real exams to those generated *in silico* for a sample synthetic exam

The frequencies of scores awarded to students for each of the 83 questions in (A) benchmark exam 1 and (B) a synthetic control exam for benchmark exam 1. The two distributions are similar, suggesting that the synthetic control exam preserves the question score distributions in benchmark exam 1.

were similarly low; the excess over null expectations was also present in the 0.20% and 0.56% bins, although the separation from the null was less pronounced than in the most stringent

bin (Figure 8). Instructor follow-up on several exams further supported that many Q-SID-identified students had in fact colluded.

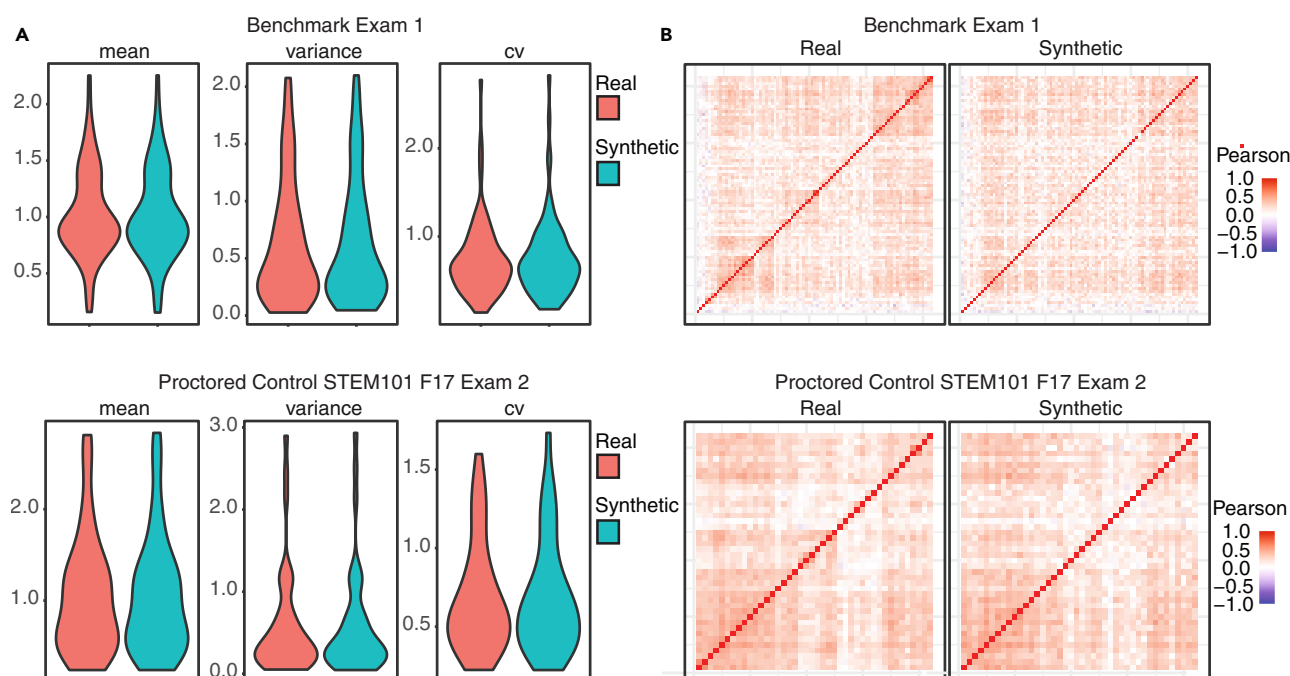


Figure 7. Synthetic control exams reproduce key distributional and dependence patterns of the real exam data

Two representative exams are shown: benchmark exam 1 (top) and proctored control STEM101 F17 exam 2 (bottom).

(A) Violin plots compare the distributions of question-wise mean, variance, and coefficient of variation (CV), where each statistic is computed across students for a given question and then summarized over all questions, between the real and synthetic exams.

(B) Heatmaps show the question-question Pearson correlation matrices for the corresponding real and synthetic question-score data.

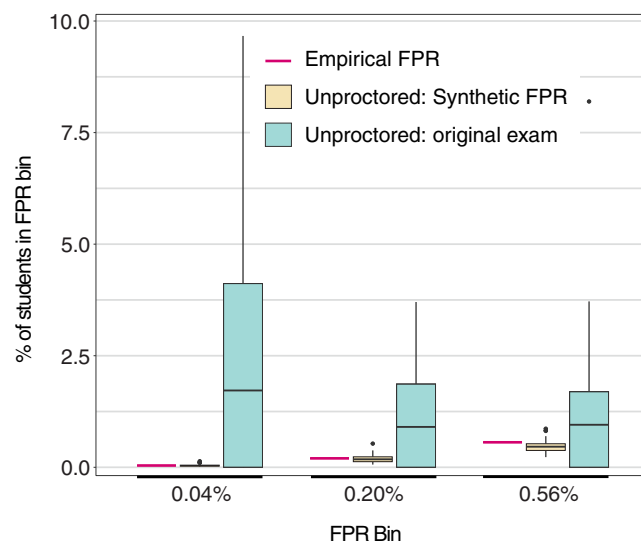


Figure 8. The percents of students placed into each empirical FPR bin for the 34 unproctored exams (green) are shown together with the synthetic FPRs calculated for these same exams (brown)

The empirical FPRs are shown by horizontal red lines. In each boxplot, the center line indicates the median, the box spans the interquartile range, the whiskers extend to $1.5 \times$ the interquartile range, and points beyond the whiskers indicate outliers.

To more systematically assess the agreement between Q-SID's empirical and synthetic FPRs, we generated randomly downsampled exams from the 19 proctored control exams that contained either different numbers of students or different numbers of questions and then calculated synthetic FPRs. We found no systematic differences between synthetic FPRs for downsampled exams that had different numbers of questions (Figure 9A). However, there is a consistent difference between synthetic FPRs for downsampled exams with different numbers of students (Figure 9B). Exams with fewer than 100 students tend to have synthetic FPRs higher than the empirical FPRs (data above the horizontal red bars), whereas exams with more than 100 students tend to have synthetic FPRs lower than the empirical FPR (data below the horizontal red bars).

The empirical and synthetic FPRs are equally valid. There is no evidence that shows one is more accurate than the other. Feedback from instructors, as well as our own inclination, strongly favor taking a conservative approach when the two FPRs diverge. As a precaution, therefore, Q-SID excludes any collusion groups whose cumulative synthetic FPR is $>0.8\%$, i.e., any cases shown above the black dashed line in Figure 9. As a result, for exams with fewer than 100 students, Q-SID will often exclude collusion groups that would normally be placed in the 0.56% empirical FPR bin, and in rare cases, in the 0.20% empirical FPR bin. As a result, fewer of the students who collude in small classes will be identified. However, in the 34 unproctored exams, because three-quarters of students who colluded were found in the 0.04% and 0.20% empirical FPR bins (Figure 8), our results

showed that Q-SID is still effective for the smaller exams within the 34.

Q-SID achieves predictable detection power across exam formats

For a collusion detection method to be informative in practice, it must retain detection power across heterogeneous exam formats. We therefore evaluated detection power using a student-level TPR, defined as the proportion of independently verified colluders placed into Q-SID collusion groups. This differs from the pair-level TPR used in the ROC and precision-recall analyses (Figure 2), which measures the proportion of verified colluding pairs recovered. Of necessity, our study uses the two benchmark exams because only for these non-multiple-choice exams has a set of students been rigorously shown to have colluded by analysis of written answers and confessions. It is likely that this true-positive set represents nearly all of the students who colluded on these exams. This conclusion is supported by the fact that an analysis of the written answers of numerous other students, whose question scores were also highly correlated, failed to detect any instances of collusion. Thus, it is reasonable to use these data to determine the TPR. We then used subsampling and exam combinations to systematically vary exam characteristics and assess how Q-SID's TPR changes with question number, question difficulty, and class size (Figure 10).

Detection power, as measured by the TPR, increases strongly with the number of questions. When virtual exams were created by subsampling questions from benchmark exam 1, the mean TPR was approximately 10% for exams with 20 questions, increased to about 55% with 60 questions, and reached about 80% with 83 questions (Figure 10A). In addition, because the CS distribution becomes unstable for exams with fewer than 20 questions, Q-SID does not analyze exams below this threshold.

Beyond question count, detection power (TPR) depends on the extent to which questions discriminate among students. Exams in which most students receive similar scores contain less information for distinguishing atypical similarity patterns from ordinary performance variation. Consistent with this pattern, restricting analysis to the 50 questions with the smallest score variability identifies only about 7% of true-positives, whereas using the 50 questions with the largest score variability identifies about 56% of true-positives. To summarize exam informativeness in a single quantity, we use an exam complexity measure (defined in the "Methods" section), which combines the number of questions and the dispersion of scores within questions. TPR increases with complexity, with the steepest gain occurring up to complexity values of approximately 20 (Figure 10B). This result suggests a practical strategy: analyze exams in sets whose total complexity is approximately 15–25, and, when individual exams have complexity below about 15, combine their question-score data before analysis. In our collection of 70 exams, the mean complexity per question is 0.33 (range 0.16–0.65), the median exam complexity is 15.46, and the median expected TPR is 46% (supplementary methods section "distribution of complexity score and predicted TPR for 70 exams").

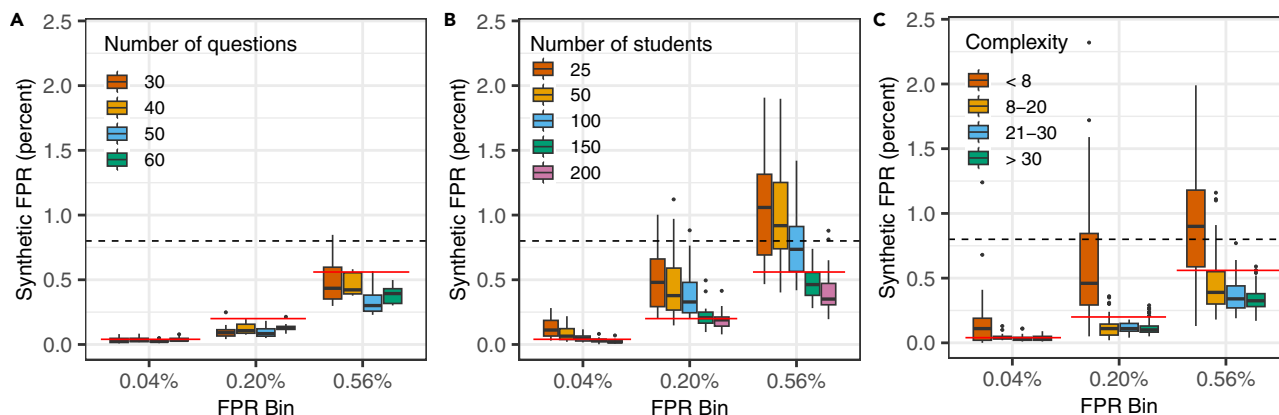


Figure 9. The relationship of synthetic FPRs with question number, student number, and complexity

Each of the 19 proctored control exams were randomly downsampled to produce 50 downsampled exams for each of (A) four question number cohorts, (B) five student number cohorts, or (C) four complexity cohorts. Synthetic control exams were then generated for sets of at least 100,000 students for each cohort from each of the 19. For each of the 19 in each cohort, the synthetic FPRs were determined and plotted as bar and whisker plots. In each boxplot, the center line indicates the median, the box spans the interquartile range, the whiskers extend to 1.5× the interquartile range, and points beyond the whiskers indicate outliers. The empirical FPRs are shown as horizontal red lines. The black dashed line indicates the synthetic FPR cutoff of 0.8%, above which detected collusion groups are conservatively excluded by Q-SID.

Finally, detection power (TPR) is relatively stable for large classes but degrades in small classes. When class size was varied by subsampling students, TPR showed little change for classes with at least 100 students, whereas for classes below 100 students, the mean TPR decreased and variability increased (Figure 10C). This reduction is driven primarily by the conservative exclusion of groups with large synthetic FPR estimates. We, therefore, set a minimum class size of 25 for Q-SID analyses, as performance variability becomes unacceptably high for smaller classes.

Because exam complexity is partly correlated with question number, we also examined how synthetic FPRs behave as a function of complexity. Synthetic FPRs are stable for exams with complexity greater than 8 but are systematically higher for complexity values below 8 (Figure 9C). Accordingly, Q-SID excludes the rare exams or exam combinations with complexity below 8 from analysis.

Informing students about Q-SID was associated with fewer detected cases

Although Q-SID was initially developed to identify students who had colluded, it soon became evident that preventing cheating before exams was more effective. Reminding students of the honor code and having them sign a statement pledging not to cheat failed to prevent collusion in the benchmark and other unproctored exams (Figure 8). Consequently, we aimed to inform students taking online exams that we were using an algorithm highly effective at detecting collusion. Students were either given a link to a website with detailed information about Q-SID or informed that others in prior classes had been caught cheating using Q-SID.

Figure 11 illustrates the percentage of students assigned to collusion groups in exams from the same UC Berkeley course. For 22 exams administered in person with strict proctoring in the 3 years before the COVID-19 pandemic, few students were

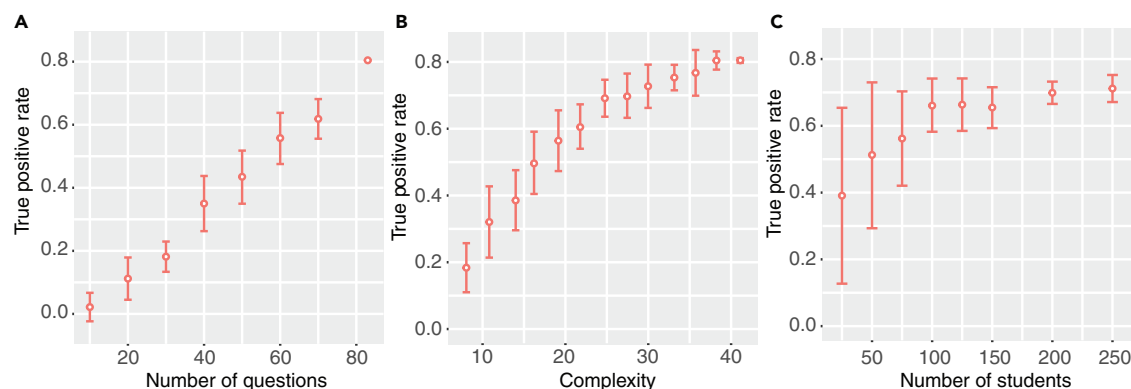


Figure 10. Q-SID TPRs across varying exam characteristics

Q-SID TPRs are shown for varying (A) numbers of questions per exam, (B) exam complexities, and (C) numbers of students per exam. At each value, points indicate the mean TPR across 20 subsampling replicates, lines connect the means, and error bars indicate ± 1 standard deviation.

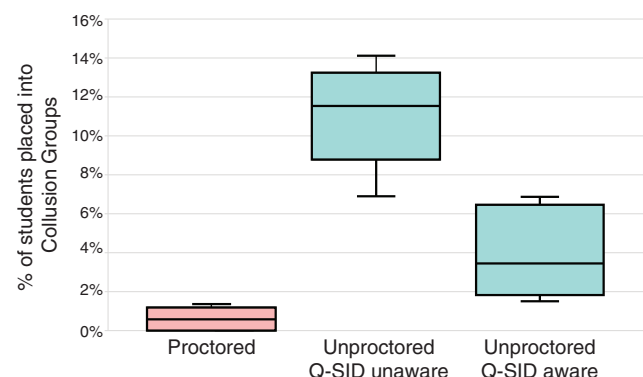


Figure 11. The percent of students placed into collusion groups for exams taken under the circumstances shown

The exam-level percentages of students placed into collusion groups are shown for proctored exams, unproctored exams for which students were not informed that Q-SID would be used, and unproctored exams for which students were informed that Q-SID would be used. The three boxplots summarize 22, 5, and 10 exams, respectively. The center line indicates the median, the box spans the interquartile range, the whiskers extend to the most extreme observed values within $1.5 \times$ the interquartile range, and points beyond the whiskers indicate outliers.

placed into collusion groups. For the first five online exams during the pandemic, students were unaware that Q-SID would be used to detect collusion, resulting in 7%–14% of the class colluding. In the subsequent 10 unproctored exams, students were either shown a website describing Q-SID or informed that students in previous classes had been caught cheating using Q-SID; in these exams, only 2%–7% of students were assigned to collusion groups. This substantial reduction is consistent with a deterrent effect and suggests that informing students about Q-SID may reduce extensive answer sharing, though we cannot exclude the possibility that some students instead adopted more limited or less-detectable forms of collusion, such as copying fewer questions.

Additionally, many students who colluded despite being informed about Q-SID confessed when confronted and indicated they would not cheat in the future. While these cases are not ideal, deterrence through sanctions is another way Q-SID discourages students from colluding in future classes.

DISCUSSION

Conclusion

We have presented Q-SID, a score-based framework for detecting likely collusion groups from exam question scores with uncertainty quantification. Q-SID applies to both multiple-choice and non-multiple-choice exams, provides complementary empirical and synthetic false-positive assessments and is computationally efficient for practical use. Together, these features make Q-SID a general and operationally useful approach for collusion screening.

Our benchmark and real-exam analyses support the effectiveness of this framework. In an analysis of 34 unproctored online exams taken during the COVID-19 pandemic by 10,526 students, Q-SID identified 4.5% of examinees as likely colluders,

with 2.4% detected at an empirical FPR of 0.04% and similarly low synthetic FPRs (Figure 8). For the four unproctored exams with full manual analysis of written answers, Q-SID placed 107 students into 39 collusion groups, of whom 101 were determined to have colluded based on textual analysis and 92 subsequently confessed (supplementary methods section “[elimination of occasional error in calculating CSs](#)”). These results show that Q-SID can recover collusion structure accurately and serve as an effective screening tool in realistic exam settings.

Finally, although we use the identity score as the default similarity measure for interpretability and computational simplicity, the Q-SID workflow is flexible: alternative similarity measures computed from item-level scores can be incorporated without changing the overall normalization, calibration, and grouping framework. This flexibility makes Q-SID adaptable to future extensions and to exam settings with richer item-level information.

Practical implications

For an exam with 200 students, there are 19,800 possible pairwise combinations of exams whose written answers would need to be compared to detect all cases of collusion. This task is impossible to perform manually, and current text analysis algorithms cannot distinguish copied answers from those that are similar because they are correct. Q-SID addresses this problem by highlighting small groups of students who are likely, but not certain, to have colluded. Instructors must then compare the written answers of these students to determine if they likely colluded. In our view, students should not be found responsible for collusion without evidence from such textual analysis.

In addition to identifying specific students for further investigation, Q-SID can rapidly measure collusion across many classes without charging individual students with cheating. Because Q-SID’s FPRs are calibrated and consistent between empirical and synthetic FPRs, the algorithm can reliably estimate the degree of collusion in each exam. We have found that some courses consistently exhibit higher levels of collusion, often reflecting students’ perceptions of the course’s importance (Table S1). For example, in a course required for entry to many medical schools, a median of 6.5% of students were identified as likely colluders across 15 unproctored online exams. Other exams showed much lower levels of collusion or none. Identifying which courses have higher levels of cheating can help instructors and administrators focus their efforts to address this problem.

Informing students about Q-SID before an exam was associated with a lower proportion of students detected as belonging to collusion groups (Figure 11). This pattern is consistent with a deterrent effect and suggests that informing students about Q-SID may help reduce extensive answer sharing.

While Q-SID can analyze multiple-choice exams, it should only be used to gauge whether collusion has occurred in such exams, not to identify and challenge specific students for cheating. The lack of written answers prevents the additional analysis needed to establish compelling evidence of individual cheating, such as required by a disciplinary hearing. When collusion is detected in multiple-choice exams, instructors should consider alternative exam formats or improved proctoring in the future.

Cheating on exams has a corrosive effect. Students who cheat often receive higher grades than they deserve, while those who

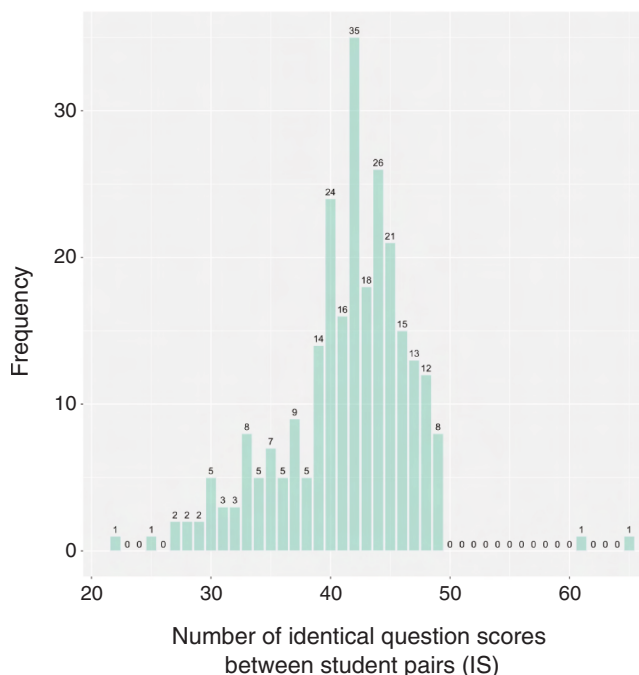


Figure 12. Histogram of the ISs of one student with the other 262 students who took benchmark exam 2

There are two large outliers of 61 and 65.

do not cheat may receive lower grades. The opportunity to cheat is greater in online exams than in traditional in-person exams due to the challenges of online proctoring. The increased use of online exams creates a profound need for methods to detect, measure, and deter collusion. Q-SID is a well-justified approach that addresses this need.

Threats to validity

There are several conditions that may affect Q-SID's validity.

First, Q-SID is less able to detect collusion when exams contain too few questions or when question scores show limited dispersion across students (e.g., most students earn very similar scores on many questions). To formalize this notion, Q-SID defines an exam complexity metric that increases with both the number of questions and the dispersion of question scores, and we observe that detection power (TPR) increases with

exam complexity (Figures 10A and 10B). Consequently, Q-SID does not analyze exams with fewer than 20 questions or with complexity below a defined threshold. We recommend combining multiple low-complexity exams taken by the same students in the same course (e.g., midterms and a final) to raise the overall complexity when appropriate. In practice, only a small minority of exams are excluded from analysis for these reasons.

Second, the TPR estimates become too variable to be reliable for small classes (Figure 10C). As a result, Q-SID is limited to class sizes of at least 25 students.

Third, for randomized exams, the question score data must be adjusted before being input into Q-SID. To deter collusion, some instructors set exams in which students receive different questions and/or varied orders of questions.³⁵ Q-SID, though, assumes students receive the same questions and computes student-to-student similarities from aligned question scores. When students receive the same questions in different orders, the question scores must be aligned across students to allow effective analysis by Q-SID. When students receive different questions, by contrast, Q-SID's strategy of computing student-to-student similarities is invalid. If only a minority of questions are set differently, however, the scores for these questions can be removed prior to Q-SID analysis with only a small loss in data complexity. Only in the extreme case where most questions differ would Q-SID analysis be ineffective.

Fourth, open-ended questions introduce the possibility of some grading variability, where students who collude may receive similar but not identical scores, which would weaken exact score agreement and reduce Q-SID's sensitivity. Nonetheless, Q-SID is effective on non-multiple-choice exams for which collusion was independently established from written answers (Figure 8), implying that grading variability is not a major concern.

Fifth, our data end in 2021 and, therefore, predate the widespread use of generative AI in exams. Generative AI has changed the academic-integrity landscape and may alter how students cheat, but current evidence shows that student-to-student collusion still occurs.^{31,32} Q-SID, therefore, remains relevant for detecting unusual similarity in student score patterns when answer sharing occurs, and it may identify groups of students who rely on the same AI tool or other external source and as a result produce unusually similar question scores.

Sixth, although Q-SID uses several strategies to reduce biases, such as the high similarity among highly scored students, and similarly for lowly scored students (Figures 12, 13, 14, and

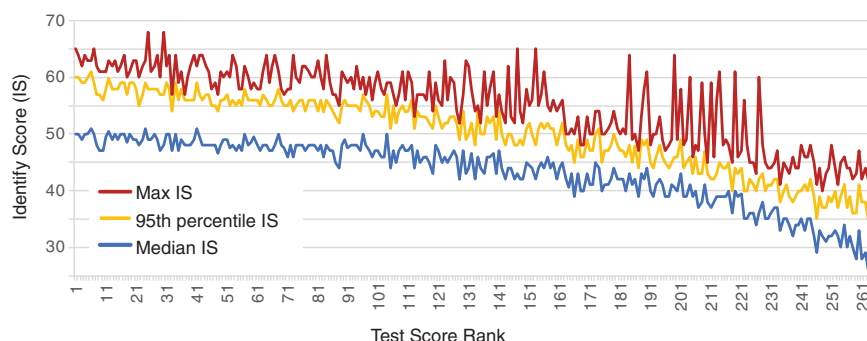


Figure 13. Max IS, 95th percentile IS, and median IS for each student who took benchmark exam 2, ranked by their test score

The student who received the highest test score was ranked number 1.

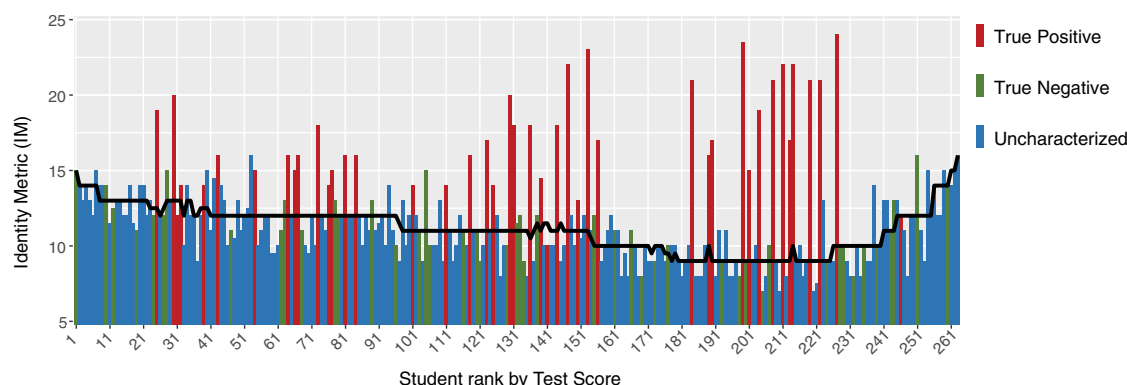


Figure 14. IMs for benchmark exam 2, ranked by student test score

15), it remains possible that some groups of students may be more likely to be detected as colluding based on which particular questions they copied. In addition, our estimate of the TPR does not account for cases in which students may have copied too few questions to be detected in the two benchmark exams, even though extensive screens for collusion were conducted on these exams. Despite these limitations, Q-SID substantially reduces the amount of collusion that can go undetected.

METHODS

The data

We used the data of 70 exams taken at UC Berkeley and UCLA to develop Q-SID (Table S1; Data S2), including two benchmark datasets with true-positives (students who colluded) and true-negatives (students who did not collude) verified by textual analysis, 36 proctored exams in which collusion is not expected to have occurred, and 32 unproctored exams collected during the COVID-19 pandemic when exams had to be given online without effective proctoring.

Benchmark datasets with true-positives and true-negatives

A challenge faced by the prior multiple-choice-only collusion detection methods is that true-positives cannot be reliably estab-

lished using the exam data. In contrast, collusion on non-multiple-choice exams has long been judged by detailed manual analysis of written answers. Students who colluded have more similar answers than the remainder of the class, often because their answers share unusual errors (see section “[manual comparison of written answers](#)” for more discussion). Hence, for non-multiple-choice exams, manual textual analysis, though time consuming, can be used to define students who colluded (i.e., true-positives) and students who did not (i.e., true-negatives).

We collected the two benchmark exam datasets from two unproctored, online exams in a UC Berkeley class during the pandemic, in which extensive collusion had occurred. Among the 263 students who took both exams, we first identified pairs of students whose question scores were highly correlated; then, from this set we defined 50 true-positives and 52 true-negatives via manual textual analysis, and we considered the remaining students as uncategorized. See section “[manual comparison of written answers](#)” for our detailed process of defining true-positives and true-negatives.

Based on the two benchmark datasets, we constructed a summary statistic for each student, called the *collusion score*, to inform the chance that the student colluded. We also used the true-positives in the two benchmark datasets to decide two CS thresholds to find the students who colluded and the collusion groups.

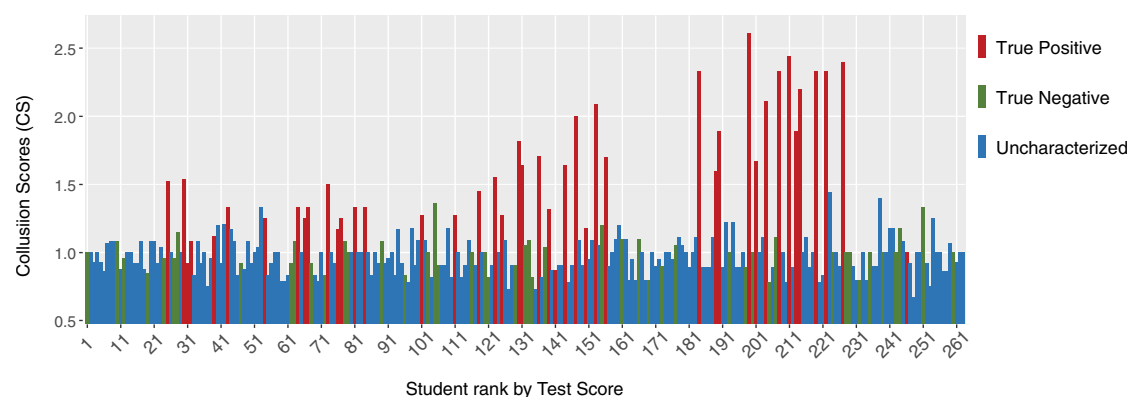


Figure 15. CSs for benchmark exam 2, ranked by student test score

Proctored exam datasets

We collected 36 exams taken between 2017 and 2021 that had been strictly proctored in person, which would make it very unlikely that the students taking these tests had been able to copy detailed written answers. Nineteen of these proctored exams are from a single STEM course taught three times per year at UC Berkeley, the same course for which the two benchmark exams were set. These 19 exams were each taken by over 300 students and thus suitable for subsampling students to create exam datasets with varying sizes for robustness analysis of Q-SID. These 19 exams are referred to as the *proctored control* and contain 6,847 students in total. The remaining 17 proctored exams were more heterogeneous, being taken by between 53 and 575 students in each exam, with a total of 3,969 students. These 17 were from different classes in several departments, representing broad subject areas, including molecular biology, chemistry, computer sciences, and statistics. Together, the 36 proctored exams contain a total of 10,816 students, which we refer to as the *proctored dataset*. Data from the proctored control and proctored dataset are used to find groups of students who did not collude but who are nonetheless detected by Q-SID and to calculate empirical FPRs.

Unproctored exam datasets for Q-SID application

In addition to the two benchmark exams that were unproctored, there are 32 additional unproctored online exams taken in 2020 and 2021 that were taken by between 54 and 639 students. This set included 13 additional exams from the same course as the benchmark exams, with the remainder drawn from a wide selection of STEM courses. In total, the 34 unproctored exams have 10,526 students. These unproctored exams will be used as application examples to demonstrate the efficacy of Q-SID in detecting collusion.

Q-SID methodology

Overview of Q-SID

Q-SID detects collusion groups in four main steps. First, for each student, it computes a similarity-based statistic from the question score matrix that summarizes the student's question score similarity with all other students and then converts this statistic into a CS by normalizing it with respect to the student's overall test score and rank (e.g., highly scoring students are naturally more similar to one another). Second, it identifies provisional collusion groups by applying thresholds to these CSs and linking students to their most similar partners. Third, it merges related provisional groups to form the final collusion groups. Fourth, it quantifies uncertainty for the detected groups using both empirical and synthetic null data, which are used to report FPRs. We next describe each component of this procedure in detail.

Metrics for collusion detection

We denote the *question score matrix* by $\mathbf{X} \in \mathbb{R}_{\geq 0}^{n \times p}$, with n students as rows and p questions as columns. For each question s , the score received by student i , denoted by $X_{is} \in \mathbb{R}_{\geq 0}$, is a non-negative value in a finite set determined by the grading rubric for that question. Using benchmark exam 2 as an example, there are $n = 263$ students who answered $p = 69$ questions. We define the *test score* of student i as the sum of student i 's question scores $\sum_{s=1}^p X_{is}$.

In the first step of Q-SID's collusion detection ("Methods" section "Q-SID algorithm"; algorithm step 1), every student is assigned a CS by the following procedure.

To begin, we define an *identity score* (IS) for each pair of students based on the students' question scores. The IS of students i and $j \neq i$ is defined as the number of identical question scores shared by the two students:

$$IS_{ij} = \sum_{s=1}^p 1\{X_{is} = X_{js}\}.$$

Hence, a student has $(n-1)$ ISs; for example, student 1 has $IS_{12}, \dots, IS_{1n}$. Figure 12 shows the ISs between one student and each of the other 262 students who took benchmark exam 2. There are two large outlier ISs of 61 and 65, which are notable since the exam only had 69 questions. Indeed, the three students corresponding to the two ISs were shown to have colluded by textual analysis of their exam answers, and all three confessed to cheating.

Q-SID is based on the observation that large IS outliers are suggestive of collusion. For this reason, each student i is assigned a "max IS" defined as $\max_{j \neq i} IS_{ij}$, the maximum of the $(n-1)$ ISs student i have with the remaining $(n-1)$ students. Accordingly, we define the student who has the max IS with student i as student i 's *first partner*: $i_1 = \arg \max_{j \neq i} IS_{ij}$. Hypothetical-

ly, if a student has a large max IS, then the student might have colluded with his or her first partner. For the purpose of detecting collusion groups, we also define student i 's *second partner* as $i_2 = \arg \max_{j \neq i, i_1} IS_{ij}$.

While the max IS is suggestive of collusion, these values alone do not indicate collusion because they suffer a systematic bias due to the dependence between a student's max IS and test score (i.e., sum of the student's question scores). The underlying reason for this dependence is that students who received higher test scores naturally have higher similarities in their question scores because they share many correct answers compared with students with lower test scores. This, in turn, leads to higher max IS values for students at the top of the class. Figure 13 (red line) illustrates this for benchmark exam 2.

To correct this bias, we sought a baseline trend of IS values along test score ranks. The ideal baseline should only reflect the bias that students who received top test scores tend to have high IS values, but should not indicate collusion. If a student receives high test scores, then the student's $(n-1)$ IS values should contain some high values due to ISs shared with other top-scoring students. Hence, an upper quantile of the student's $(n-1)$ IS scores should be high. Motivated by this rationale, we examined two quantiles of each student's $(n-1)$ IS values: the median and the 95th percentile. Based on benchmark exam 2, Figure 13 shows the trends of these two quantiles, which are highly correlated with each other (with the Pearson correlation as 0.98) and generally decrease as the students' test scores decrease, suggesting that either of these trends could be reasonable baselines. However, since it is particularly unlikely that over half of the students in a class would collude in a single, large group, the median IS value is the less likely of the two to

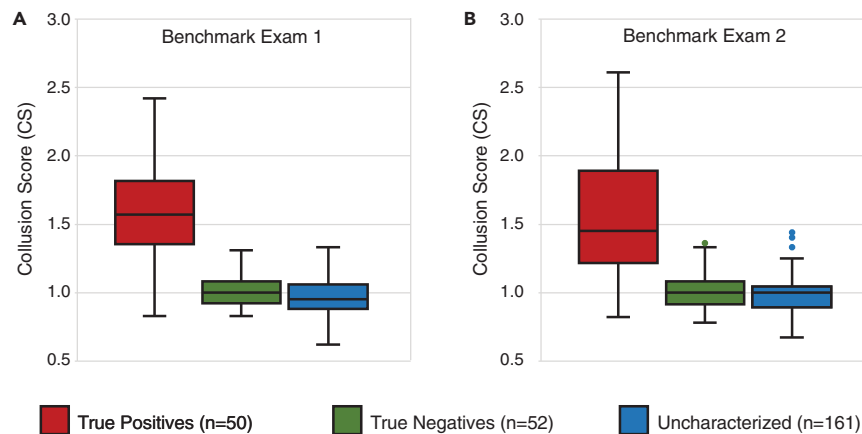


Figure 16. The distributions of CS values for true-positives, true-negatives, and uncharacterized students who took both the benchmark exams

(A) Benchmark exam 1 and (B) benchmark exam 2. Student group sizes are indicated: true positives, $n = 50$; true negatives, $n = 52$; and uncharacterized students, $n = 161$. In each boxplot, the center line indicates the median, the box spans the interquartile range, the whiskers extend to the most extreme observed values within $1.5 \times$ the interquartile range, and points beyond the whiskers indicate outliers.

indicate collusion. Hence, we adopted the median IS values to correct the bias in max IS values as follows.

We define student i 's *identity metric* (IM) as

$$IM_i = \max_{j \neq i} IS_{ij} - \text{median}_{j \neq i} IS_{ij}.$$

Pearson correlations indicate that the IM does indeed have a reduced test rank bias compared with Max IS: while the correlation between the median IS values and the max IS values is 0.87—reconfirming that there is a bias in the max IS values—the correlation between the median IS values and the IM values is only 0.06. Despite this, IM values do retain a residual test rank bias. Figure 14 shows that the local median IM values have a slight decline along the test score ranks from 1 to 220, and in addition, exhibited a rise further down the rank list. The additional bias for bottom-ranked students is due to them sharing a high percent of question scores of zeros, which causes their max IS values and median IS values to behave differently from the other students.

To remove the residual test rank biases shown in Figure 14, we developed a further metric, a CS for each student i as

$$CS_i = IM_i / (\text{localmedian}IM_i),$$

where the $\text{localmedian}IM_i$ is the median of the IMs of the immediate neighbors of student i based on the test score rank.

Specifically, the median IM (median IM) for each student is calculated as follows. Students are first ranked by test score, bringing students of similar ability into proximity. A sliding window that spans a set of adjacent students is used to produce local median values for IMs at every position in the test score rank list. For most students, the windows are centered on that student and extend an equal number of students above and below the student on the rank list. However, for the three students at each end of the rank list, the windows cannot be centered, and instead, the three students all share the same window. For classes of 31 or more students, the windows cover 31 students, except at the ends of the rank list, where they progressively shorten to 7 students. As an example, for a student test score rank 50 in a class of 100, the median of the IMs for student ranks 35–65 is calculated. A window of 31 was chosen as this provided a reasonably smoothly changing local average while not losing too much resolution, particularly for smaller classes.

Figure 15 shows the CS values of the students ordered by test score rank for benchmark exam 2. In both Figures 14 and 15, red-colored bars represent the students known to have colluded based on forensic analysis of written answers (i.e., true-positives). In contrast, green-colored bars highlight the students who have not colluded because no unusual similarities were spotted in their written answers (i.e., true-negatives). In addition, blue-colored bars show the students whose written answers have not been forensically analyzed (i.e., uncharacterized students). Inspecting the true-negatives and uncharacterized students, we found no obvious bias in the CS values for the test score ranks, confirming that CS is a better metric than IM. As a result, the true-positives became more distinct from the true-negatives and uncharacterized students in terms of CS values (Figure 15) than IM values (Figure 14).

Detection of collusion groups

Given the intuition that the students with the highest CS values are those most likely to have colluded, we developed an algorithm to define collusion groups based on students' CS values (section "Q-SID algorithm" algorithm steps 2 and 3).

First, step 2 of Q-SID identifies non-overlapping *provisional collusion groups*, each of which is a set of students. Specifically, Q-SID first identifies all {student, 1st partner} sets such that both students in a set have $CS \geq c_1$ and at least one student has $CS \geq c_2$, with the two thresholds $c_2 > c_1 > 0$. The choice of threshold values for c_1 and c_2 will be discussed in the following section "choice of CS value thresholds." Then, Q-SID combines the identified sets that overlap into K non-overlapping provisional groups $\mathcal{T}_1, \dots, \mathcal{T}_K$, which are ordered by the maximum CS of the students in each group, in a decreasing order.

Second, in step 3 of Q-SID, for each of $k = 1, \dots, K$, if at least two members of a provisional group $\mathcal{T}_k$ share a second partner that is a member of later provisional group $\mathcal{T}_s, k < s \leq K$, then $\mathcal{T}_s$ is merged into $\mathcal{T}_k$. After this merging, the remaining groups are designated as *collusion groups* $\mathcal{C}_1, \dots, \mathcal{C}_M$, with $M \leq K$.

Finally, for each collusion group $\mathcal{C}_m$, we define its maximum CS (maxCS) value maxCS_m as the maximum of the CS values of the students in it, $m = 1, \dots, M$. Due to the definition of collusion groups, all collusion groups have maxCS values at least c_2 .

Choice of CS value thresholds

We determined the CS value thresholds, c_1 and c_2 , empirically in an exam-size-dependent manner. The goal was to maintain

Table 3. The CS percentiles used to define collusion groups and FPR bins

Threshold	CSs in the benchmark exams	CDF values in the empirical null distribution of CSs	Interpretation
c_1	1.23	94.55%	Minimum CSs of students identified
c_2	1.50	99.71%	Minimum maxCSs for collusion groups in the high-risk bin (0.56% empirical FPR)
c_3	1.60	99.88%	Minimum maxCSs for collusion groups in the medium-risk bin (0.20% empirical FPR)
c_4	1.70	99.99%	Minimum maxCSs for collusion groups in the low-risk bin (0.04% empirical FPR).

Each threshold is defined as a quantile of the CS values in the proctored control exams. All collusion group members must have CS values $\geq c_1$. All student-first partner pairs must have at least one member with a CS value $\geq c_2$. Thresholds c_3 and c_4 are used to assign collusion groups to three bins with different risk levels and empirical FPR values.

sensitivity to verified collusion cases while limiting inclusion of students without evidence of collusion. To do so, we anchored these thresholds using (1) the two benchmark exams, in which collusion status was established by a manual review of the written answers (section “[benchmark datasets with true-positives and true-negatives](#)”), and (2) the 19 proctored control exams, comprising 6,847 students who are unlikely to have colluded (section “[proctored exam datasets](#)”). For each exam, we computed the CS value for every student and treated the CS values from the proctored control exams as an *empirical null distribution* of CS values.

For an input exam with more than 250 students, we used the two benchmark exams to determine the c_1 and c_2 thresholds as $c_1 = 1.23$ and $c_2 = 1.50$. Our choice of the thresholds was based on (1) the CS values of the true-positives, true-negatives, and uncharacterized students (Figure 16) and (2) the CS values of the students’ first partners (Tables S2 and S3). Using these thresholds, Q-SID identified 46 students to have likely colluded in at least one of the two benchmark exams. Among the 46 identified students, 44 students were confirmed as true-positives (given the total of 50 true-positives, the TPR was 88.00%) and 2 students were uncharacterized.

Based on preliminary evidence that the CS value distribution depends on the exam size (i.e., number of students), we pre-calculated the thresholds c_1 and c_2 for a grid of exam sizes (15, 20, ..., 45, 50, 60, ..., 240, 250) using a subsampling-and-quantile-matching procedure with three steps. First, we mapped the benchmark-exam thresholds $c_1 = 1.23$ and $c_2 = 1.50$ to empirical null cumulative distribution function (CDF) values, yielding 95.55% and 99.71%, respectively, based on the empirical null distribution of CS values from the 19 proctored control exams (Table 3; Figure 17). The CS thresholds shown in Figure 17 are also provided in a lookup table (Data S2). Q-SID uses these values to determine the exam-size-specific absolute CS thresholds for assigning students to collusion groups and FPR bins.

Second, for each exam size n in {15, 20, ..., 45, 50, 60, ..., 240, 250}, we downsampled each of the 19 proctored control exams to size n by randomly sampling n students from the class with replacement, calculated the CS value of each student within a downsampled exam, and repeated the downsampling and CS calculation 100 times. Then, we pooled the $(19 \times n \times 100)$ CS values to create an empirical null distribution for the exam size n .

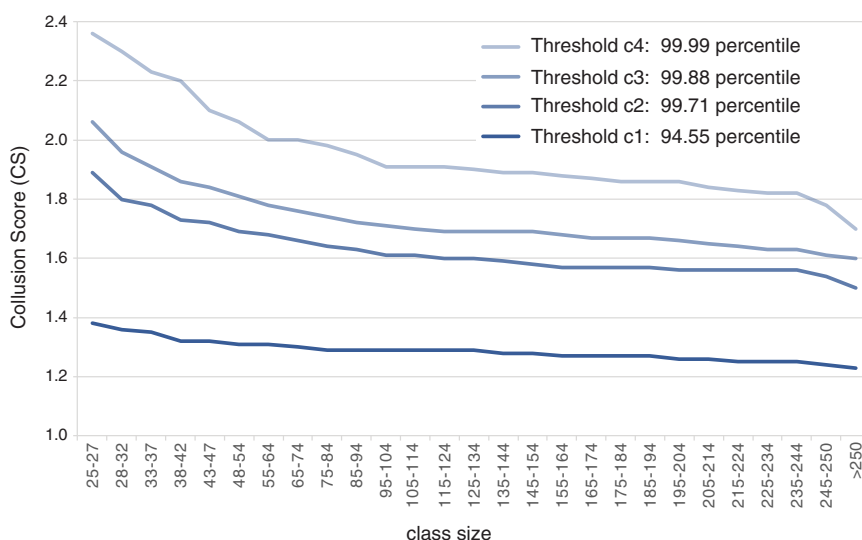


Figure 17. CS value thresholds that correspond to the four CS distribution percentiles for varying class sizes subsampled from the proctored control exams

Third, we defined the c_1 and c_2 thresholds for the exam size n as the 95.55% and 99.71% quantiles of the empirical null distribution.

For an input exam with at most 250 students, we rounded the class size to the nearest value n in $\{15, 20, \dots, 45, 50, 60, \dots, 240, 250\}$ and applied the corresponding c_1 and c_2 thresholds.

Empirical FPR estimation from proctored exams

We defined the *empirical FPR* for each collusion group based on the *empirical null data* (also referred to as the *proctored dataset*), consisting of 36 proctored exams taken between the years 2017 and 2021 and containing 10,816 students. Among the 36 exams, 19 proctored control exams, consisting of 6,847 students, were from the same course for which the two benchmark exams were set. Because the 36 exams were strictly proctored in person, it is unlikely that any of the students taking these exams colluded (section “[proctored exam datasets](#)”). Therefore, we defined false-positives as those students from the proctored exams who were placed into collusion groups by Q-SID.

To help instructors decide which of the M collusion groups, $G_1, \dots, G_M$, detected in an unproctored exam are worth examining, we categorized the collusion groups into three *bins* to reflect risk levels—*low*, *medium*, and *high*—by thresholding the maxCS values of collusion groups (section “Q-SID algorithm” algorithm step 4). Specifically, we empirically define two maxCS thresholds, c_3 and c_4 , based on the exam size, by downsampling the two benchmark exams, both containing more than 250 students.

First, based on the two benchmark exams, we set $c_3 = 1.6$ and $c_4 = 1.7$ for exams with more than 250 students. We chose the two maxCS thresholds so that the three bins, containing collusion groups with maxCS values in $[c_2, c_3)$, $[c_3, c_4)$, and $[c_4, \infty)$, have similar numbers of collusion groups.

To account for the effects of exam size on maxCS values, similar to c_1 and c_2 , we pre-calculated the thresholds c_3 and c_4 for an array of exam sizes 15, 20, ..., 45, 50, 60, ..., 240, 250 by a subsampling-and-quantile-matching approach consisting of three steps. First, we converted the benchmark exam thresholds $c_3 = 1.60$ and $c_4 = 1.70$ to empirical null CDF values as 99.88% and 99.99% based on the empirical null distribution of CS values obtained from the 19 proctored control exams (Table 3). Second, for each exam size n in $\{15, 20, \dots, 45, 50, 60, \dots, 240, 250\}$, we downsampled each of the 19 proctored control exams to size n , calculated the CS value of each student within a downsampled exam, and repeated the downsampling and CS calculation for 100 times. Then, we pooled the $(19 \times n \times 100)$ CS values to create an empirical null distribution for the exam size n . Third, we defined the c_3 and c_4 thresholds for the exam size n as the 99.88% and 99.99% quantiles of the empirical null distribution (Figure 17).

Then, for exams with size no more than 250, we rounded the exam size to the closest integer n in the set of $\{15, 20, \dots, 45, 50, 60, \dots, 240, 250\}$. Then we used the c_3 and c_4 thresholds for exam size n for this input exam to categorize the collusion groups into three bins.

With the categorization of the M collusion groups into three bins, we calculated the empirical FPR for each bin from an expanded set of 36 proctored exams taken by $n_e = 10,816$ students to increase the diversity of exams. We used the percent

of these students (whom we believed to have not colluded) placed into each bin to define the empirical FPR of that bin. Then, we obtained the empirical FPR values of the three bins as 0.04%, 0.20%, and 0.56%, corresponding to the low-, medium-, and high-risk levels. These empirical FPR values were defined to be cumulative: approximately 0.04% of students in the 36 proctored exams fell into the 0.04% empirical FPR bin, 0.16% fell into the 0.20% empirical FPR bin, and 0.36% fell into the 0.56% empirical FPR bin.

We denoted p to be the true FPR and assumed that n students are independent Bernoulli random variables with probability p to be mis-identified as a false-positive. Then, we could write out the

95% confidence interval of p as $\hat{p}_e \pm 1.96 \sqrt{\frac{\hat{p}_e(1-\hat{p}_e)}{n}}$, where $\hat{p}_e$ denoting the empirical FPR was assumed to follow an approximately normal distribution with mean p and variance $\frac{p(1-p)}{n}$. Table 2 summarizes the 95% confidence intervals of true FPR for the three empirical FPR bins.

Synthetic control data generation

Synthetic control exams are *in silico* synthetic versions of the query exam that capture the distributions of question scores and test scores, but do so in a way that does not include any correlations within the data due to collusion. To generate a realistic synthetic question score matrix, we include two steps: model fitting on the real question score matrix and generation of a synthetic question score matrix by sampling from the fitted model. Note that the marginal multinomial distribution requires integer values and many question scores are not integer, example.g., 1.5 or 0.33. Hence, before fitting the model, we first multiply all scores by 100 and then round to create integers that distinguish question scores that differ by ≥ 0.01 . Then after we fit the model and generate the synthetic score matrix; we recover the simulated score matrix back to the original scale by dividing 100.

We denote the integer question score matrix by $\mathbf{X} \in \mathbb{N}^{n \times p}$ with p questions and n students. We further assume that the n students belong to K cohorts, which are uniformly separated based on the student test scores’ quantiles (specifically, we use $K = 5$ in our analysis), and we perform the simulation independently for each student cohort. We denote the $\mathbf{X}^{(k)} \in \mathbb{N}^{n^{(k)} \times p}$ as the question score sub-matrix corresponding to the student cohort k , where $n^{(k)}$ is the number of students belonging to cohort k . To simplify the notations, we omit the superscript (k) for the k -th cohort in the following discussion of generative models.

Fitting a generative model of the question score matrix

The first step is to fit a generative model to characterize the joint distribution of question scores across students in the question score matrix. The model treats each question’s score as a random variable and the students’ question scores as independent and identically distributed observations. We model the marginal distribution of each question’s scores as multinomial, which matches the finite set of rubric-defined score categories for that question. Because scores on different questions can be dependent (e.g., a student who scores highly on one question is more likely to score highly on others), we use a Gaussian copula to fit a “null” joint distribution of the question scores. This technique couples the estimated marginal distributions while preserving rank correlations across questions and has previously been used to generate realistic synthetic data for

homogeneous cell populations in single-cell biology.^{36–38} In our setting, the fitted unimodal joint distribution serves as a no-collusion null model: synthetic students are sampled independently from a common multivariate distribution without subgroup structure to indicate collusion groups. We then generate synthetic control exams by independently sampling these synthetic students from the fitted null model.

In the following analysis, we denote the question score matrix by $\mathbf{X} \in \mathbb{R}_{\geq 0}^{n \times p}$ with n students as rows and p questions as columns, where each score $X_{is} \in \mathbb{R}_{\geq 0}$ is a non-negative value in a finite set determined by the grading rubric for that question. For each student $i = 1, \dots, n$, we further denote $X_{i\cdot} = (X_{i1}, \dots, X_{ip})^\top \in \mathbb{R}_{\geq 0}^p$ as a random vector of student i 's scores for the p questions, and $x_{i\cdot} = (x_{i1}, \dots, x_{ip})^\top$ is a realization of $X_{i\cdot}$, i.e., student i 's actual question scores in a dataset.

Note that $X_{i\cdot} = (X_{i1}, \dots, X_{ip})^\top \in \mathbb{R}^p$ is the random score vector corresponding to student $i, i = 1, \dots, n$. We assume that the random vector $X_{i\cdot}$ jointly follows a p -dimensional distribution F , independently for $i = 1, \dots, n$. Marginally for each question s , we assume that random variables X_{is} independently follow a multinomial distribution F_s with an event probability vector θ_s and a trial number 1, i.e., $X_{is} \sim \text{MultiNomial}(\theta_s, 1)$. The event probability vector θ_s could be estimated by the proportion vector of question scores' occurrence.

Given the estimated the marginal distributions $\hat{F}_1, \dots, \hat{F}_p$, we use a copula to model the joint distribution F of $X_{i\cdot}$. A copula is defined as a CDF of p random variables, for which marginally each random variable is following a uniform distribution Uniform[0,1]. Sklar's theorem states that for continuous marginal CDF distributions $F_s, s = 1, \dots, p$, there exists a unique copula function $C: [0,1]^p \rightarrow [0,1]$ such that

$$F(x_{i1}, \dots, x_{ip}) = C(F_1(x_{i1}), \dots, F_p(x_{ip})),$$

where $x_{i\cdot} = (x_{i1}, \dots, x_{ip})^\top$ is a realization of random vector $X_{i\cdot} = (X_{i1}, \dots, X_{ip})^\top$. However, in our analysis, the marginal multinomial distributions F_s are not continuous; thus, the copula function is not unique, i.e., unidentifiable. We further deploy the distributional transform technique to construct the continuous random variables $\tilde{X}_{is} = X_{is} + V_{is}$, where $V_{is}, i = 1, \dots, n, s = 1, \dots, p$ are independent random variables following Uniform[0,1]. We denote the continuous CDF of $\tilde{X}_{is}$ as $\tilde{F}_s$ and further define

$$U_{is} = V_{is}F_s(X_{is} - 1) + (1 - V_{is})F_s(X_{is}).$$

As shown in our previous paper,³⁷ U_{is} is essentially deploying the CDF $\tilde{F}_s$ onto the random variable $\tilde{X}_{is}$, i.e., $U_{is} = \tilde{F}_s(X_{is} + V_{is})$. Hence $\{U_{is}; i = 1, \dots, n\}$ are independent random variables following Uniform[0,1]. Therefore, the joint CDF $\tilde{F}$ of the continuous random variables $\tilde{X}_{i\cdot} = (\tilde{X}_{i1}, \dots, \tilde{X}_{ip})^\top$ has the unique copula expression $\tilde{F}(\tilde{x}_{i1}, \dots, \tilde{x}_{ip}) = C(u_{i1}, \dots, u_{ip})$ according to Sklar's theorem, where $\tilde{x}_{is}$ and u_{is} are realizations of $\tilde{X}_{is}$ and U_{is} . We thus use $\tilde{F}$'s unique copula expression $C(u_{i1}, \dots, u_{ip})$ to approximate the discrete random variables' joint CDF F :

$$F(x_{i1}, \dots, x_{ip}) \approx C(u_{i1}, \dots, u_{ip}).$$

Specifically, we use a Gaussian copula structure as an approximation of $C(u_{i1}, \dots, u_{ip})$ in our analysis; hence the CDF F can be expressed through the following formula:

$$F(x_{i1}, \dots, x_{ip}) \approx \Phi_p(\Phi^{-1}(u_{i1}), \dots, \Phi^{-1}(u_{ip}); \mathbf{R}),$$

where Φ_p is the CDF of a p -dimensional Gaussian distribution with a zero mean vector and a covariance matrix $\mathbf{R}$ and Φ is the CDF of a standard Gaussian distribution. The copula covariance matrix $\mathbf{R}$ is equal to the correlation matrix of $(\Phi^{-1}(u_{i1}), \dots, \Phi^{-1}(u_{ip}))^\top$. To estimate the covariance matrix $\mathbf{R}$, we (1) sample v_{is} from Uniform[0,1] independently for $i = 1, \dots, n$ and $s = 1, \dots, p$; (2) calculate quantities u_{is} as

$$u_{is} = v_{is}\hat{F}_s(x_{is} - 1) + (1 - v_{is})\hat{F}_s(x_{is}),$$

where $\hat{F}_s$ is the estimated marginal CDF function for the s -th question; and (3) use the sample covariance matrix $\hat{\mathbf{R}}$ of $(\Phi^{-1}(u_{i1}), \dots, \Phi^{-1}(u_{ip}))^\top, i = 1, \dots, n$ as the estimator of $\mathbf{R}$.

In summary, we first estimate p marginal multinomial distributions $F_1, \dots, F_p$ as $\hat{F}_1, \dots, \hat{F}_p$. Then we calculate the distribution transformed quantities u_{ij} and estimate the Gaussian copula covariance matrix $\hat{\mathbf{R}}$ to estimate the joint distribution F as:

$$\hat{F}(x_{i1}, \dots, x_{ip}) = \Phi_p(\Phi^{-1}(u_{i1}), \dots, \Phi^{-1}(u_{ip}); \hat{\mathbf{R}}),$$

with $\hat{\Theta} = \{\hat{\theta}_1, \dots, \hat{\theta}_p, \hat{\mathbf{R}}\}$ as the estimated parameters.

Generation of a synthetic question score matrix

For each student cohort k , given the fitted p -dimensional joint distribution $\hat{F}^{(k)}$ with parameters $\hat{\Theta}^{(k)} = \{\hat{\theta}_1^{(k)}, \dots, \hat{\theta}_p^{(k)}, \hat{\mathbf{R}}^{(k)}\}$, we generate the synthetic question score matrix for $n^{(k)}$ students through the following procedures: first, we draw $n^{(k)}$ vectors, denoted by $z_i^{(k)'} \in \mathbb{R}^p, i = 1, \dots, n^{(k)}$, independently from the p -dimensional Gaussian distribution $\Phi_p(\cdot; \hat{\mathbf{R}}^{(k)})$; then, for $s = 1, \dots, p$, we convert $z_{is}^{(k)'}$ to $x_{is}^{(k)'}$ by setting $x_{is}^{(k)'}$ as the $\Phi(z_{is}^{(k)'})$ -th quantile of distribution $\hat{F}_s^{(k)}$, i.e., $\text{MultiNomial}(\hat{\theta}_s^{(k)}, 1)$.

Therefore, matrix $\mathbf{X}^{(k)'} = (x_{is}^{(k)'}) \in \mathbb{N}^{n^{(k)} \times p}$ is the synthetic question score submatrix corresponding to the student cohort k . By horizontally concatenating the K synthetic question score matrices $\mathbf{X}^{(k)'}$, we output $\mathbf{X}' = \left[(\mathbf{X}^{(1)'})^\top \dots (\mathbf{X}^{(K)'})^\top \right]^\top \in \mathbb{N}^{n \times p}$ as the final synthetic question score matrix.

Evaluating similarity between synthetic data and real data

To assess whether synthetic control exams preserve the statistical structure of real question-score data, we evaluated the similarity between synthetic and real question-by-student score matrices using three complementary diagnostics. As representative examples, Figure 7 displays these diagnostics for benchmark exam 1 and proctored control STEM101 F17 Exam 2.

First, we compared marginal question-level summaries by computing, for each question, the sample mean, sample variance, and coefficient of variation across students in the real and synthetic matrices and then visualizing the distributions of these per-question summaries using violin plots.

Second, we compared dependence structure across questions by computing the question-question Pearson correlation matrix separately for the real and synthetic matrices and displaying the resulting correlation heatmaps. Third, we quantified mixing between real and synthetic students in a shared low-dimensional representation using the miLISI.³⁴

For the embedding-based comparison, we vertically concatenated the real and synthetic question score matrices (with students as rows and questions as columns), and then performed dimension reduction across the question dimension. Specifically, we applied principal-component analysis (PCA) and retained the first 30 principal components, followed by uniform manifold approximation and projection (UMAP)^{39,40} to obtain a two-dimensional embedding of students. We then computed miLISI on the combined embedding using the binary label indicating whether a student originated from the real or synthetic matrix. The miLISI ranges from 1 to 2, where larger values indicate greater local mixing between real and synthetic students and therefore, stronger similarity between the two datasets in the embedded space.

These diagnostics were used jointly to evaluate similarity at multiple levels: marginal question-score distributions (question-level summaries), inter-question dependence structure (correlation matrices), and global/local alignment of student-level score profiles (miLISI in a shared embedding). Numerical results for representative exams are reported in “results” section and supplementary figures.

Synthetic FPR estimation from synthetic control data

For a given query exam, Q-SID estimates an exam-specific FPR by applying the full Q-SID pipeline to synthetic control exams generated from that query exam (section “[synthetic control data generation](#)”). Synthetic control data are generated until the total number of synthetic students exceeds $n_s > 100,000$. For each synthetic control exam, we compute CS values, form collusion groups using the class-size-specific thresholds, and assign detected groups to the same three tiers defined by the cutoffs $\{c_2, c_3, c_4\}$.

Let $\hat{p}_s(c)$ denote the cumulative synthetic FPR at cutoff $c \in \{c_2, c_3, c_4\}$, defined as the proportion of synthetic students placed into collusion groups satisfying $\max CS \geq c$. We estimate $\hat{p}_s(c)$ by the corresponding empirical proportion across the pooled synthetic students. To quantify uncertainty, we construct a 95% confidence interval using the normal approximation to the binomial,

$$\hat{p}_s(c) \pm 1.96 \sqrt{\frac{\hat{p}_s(c)(1 - \hat{p}_s(c))}{n_s}},$$

and we report these intervals for the benchmark exams in [Table 2](#).

Synthetic and empirical FPR estimates may differ, particularly for small class sizes (“[results](#)” section; [Figure 9B](#)). When the cumulative synthetic FPR exceeds 0.8%, Q-SID excludes collusion groups assigned to that tier from reporting, a rule that most frequently affects groups in the least stringent tier for classes with fewer than 100 students.

Evaluation of synthetic FPR across class size, question number, and exam complexity

To evaluate exam’s informativeness, we define an exam *complexity* measure based on the dispersion of observed numeric scores within each question and aggregate it across questions. Specifically, for each question $s = 1, \dots, p$, let $\hat{p}_s(x)$ denote the empirical relative frequency of score value x among students and let $\mathcal{X}$ be the finite set of observed scores for that question. The question-level complexity is defined as

$$O_s = -\log_{10} \left(\sum_{x \in \mathcal{X}} \hat{p}_s(x)^2 \right), \text{ where } \hat{p}_s(x) = \frac{1}{n} \sum_{i=1}^n 1\{X_{is} = x\}.$$

The exam-level complexity is then

$$O = \sum_{s=1}^p O_s.$$

To examine whether synthetic FPR estimates depend on key exam design characteristics, we performed a resampling-based sensitivity analysis using the proctored control exams. Specifically, we evaluated the behavior of synthetic FPR across class size, question number, and exam complexity.

For the question-number analysis ([Figure 9A](#)), we restricted attention to the eight proctored control exams with more than 60 questions, so that multiple downsampled question sets could be generated from each exam. We generated virtual exams with 30, 40, 50, or 60 questions. For each source exam and each target question count, we created 50 virtual exams by sampling questions without replacement. For each virtual exam, we generated synthetic control exams with a combined total of at least 100,000 synthetic students and computed the synthetic FPRs. Virtual exams with 20 questions were not included because, for these eight source exams, such subsets typically had exam complexity values below 8.

For the class-size analysis ([Figure 9B](#)), we used all 19 proctored control exams. Each exam was randomly downsampled 50 times at each of the five class sizes 25, 50, 100, 150, and 200. For each downsampled exam, we generated synthetic control exams with a combined total of at least 100,000 synthetic students and then computed the synthetic FPRs using the standard Q-SID procedure.

For the exam-complexity analysis ([Figure 9C](#)), we again used the eight proctored control exams with more than 60 questions. We generated virtual exams with question counts in $\{10, 30, \dots, 150\}$. For each source exam and each question count, we created 20 virtual exams by sampling questions (not students) with replacement so that the virtual exams could span a broader range of question numbers and exam complexity values than the source exams alone. For each virtual exam, we generated synthetic control exams with a combined total of at least 100,000 synthetic students and computed the synthetic FPRs. We then grouped the resulting virtual exams into four categories according to their exam complexity values: <8 , 8–20, 20–30, and >30 . The corresponding synthetic FPR distributions are summarized by boxplots in [Figure 9C](#).

Interpreting Q-SID outputs: FPR bins and optional p values

Q-SID is calibrated using two sources of null data: an empirical null and a synthetic null. It can optionally report p values as null

tail probabilities for any Q-SID statistic T , such as an identity score or a group-level summary. Specifically, given B null replicates $\{T^{(b)}\}_{b=1}^B$ generated from either the empirical null (proc-tored exams) or the synthetic null (*in silico* examinees), a one-sided empirical p value is computed as

$$p = \frac{1 + \sum_{b=1}^B 1\{T^{(b)} \geq T_{\text{obs}}\}}{1+B},$$

where $1\{\cdot\}$ denotes the indicator function, and the tail direction is chosen so that larger values correspond to stronger evidence of collusion. Thus, when desired, Q-SID can provide pair-level or group-level p values under either null.

In the main text, however, we report calibrated FPR bins rather than p values. This choice is motivated by interpretability and robustness in operational settings. First, an FPR has a direct practical meaning: it quantifies the expected frequency of false-positives among flagged outputs for a given exam and, therefore, translates naturally into an expected number of erroneous flags. Second, collusion detection is inherently multiple comparison in nature—many pairs and many candidate groups may be implicitly considered—so raw p values require multiplicity adjustments and can be difficult to interpret consistently across exams with different sizes, score distributions, and dependence structures. By contrast, our FPR bins provide a stable, exam-specific error-control scale that is easy to communicate and compare across exams, while still allowing p values to be reported as a supplementary output when needed.

Comparison with existing methods

We compared Q-SID with the pairwise methods GBT, M4, unweighted ω , and weighted ω , and with the groupwise methods CD and ASlg. Applying GBT, M4, and the ω methods requires first fitting an item-response model, where an item corresponds to an individual exam question, to estimate item parameters and student abilities. These estimates determine the expected probabilities that two independently responding students receive the same scores on each item. Some method-specific similarity statistics are not one-dimensional, so they cannot all be used directly to rank student pairs. Then these statistics are conveyed into upper-tail probabilities, which we used as scalar ranking scores; smaller values indicate greater evidence of similarity.

The benchmark exams contained heterogeneous partial-credit item scales with up to \$12 or \$13 observed score values, which did not support stable estimation of a fully parameterized polytomous item-response model without aligning, collapsing, or otherwise recoding item-specific score categories. To retain the original partial-credit scores while obtaining a scalar ranking for each existing method, we specified a common graded-response reference model rather than dichotomizing the responses or fitting a high-dimensional polytomous model. This is a compromise in applying the methods developed for multiple-choice exams to our setting, in the absence of reliably estimated item parameters. Item discriminations were set to \$1 so that all items were assigned the same discrimination, category thresholds were placed at equally spaced values from \$−1 to \$1 to define a common ordered scale, and standardized total score was used to account for differences in overall student performance. The model-dependent methods used

this reference model to convert their method-specific statistics into upper-tail probabilities with which student pairs could be ranked. Because the reference model was specified rather than fitted, these probabilities are interpreted as ranking scores rather than calibrated inferential p values. We acknowledge that this implementation may not represent the optimal adaptation of these existing methods, which were designed for multiple-choice exams, to our setting. The results should therefore be interpreted only as evidence that applying existing methods to non-multiple-choice exams is not straightforward.

Q-SID itself detects students and collusion groups rather than classifying individual pairs as positive or negative. For the pair-level comparison only, we used the identity score $\text{I-S}_{\{ij\}}$ as Q-SID's score for each student pair. Pairs in which both students belonged to the same manually verified collusion group were defined as positive. All other pairs were assigned to a reference-negative class, including pairs involving uncharacterized students; thus, a false-positive pair may include a colluding student and a non-colluding student.

For GBT, M4, and the Omega methods, ROC and precision-recall curves were calculated by varying the reference-model upper-tail probability threshold, with smaller values indicating greater similarity. For Q-SID, the curves were calculated by varying the identity-score threshold, with larger values indicating greater similarity. At each threshold, we calculated $\text{TPR} = \text{TP} / (\text{TP} + \text{FN})$, $\text{FPR} = \text{FP} / (\text{FP} + \text{TN})$, and $\text{precision} = \text{TP} / (\text{TP} + \text{FP})$, where TP, FP, FN, and TN denote the numbers of true-positive, false-positive, false-negative, and true-negative pairs, respectively. Accordingly, TPR (equivalently, recall) in these ROC and precision-recall analyses is defined at the student-pair level. This differs from TPR in the Q-SID power analyses below, which is defined at the student level as the proportion of individually verified colluders assigned to final Q-SID collusion groups.

Detection power assessment

We quantify detection power using the TPR, the proportion of benchmark true-positives placed into Q-SID collusion groups. The benchmark exams are used because they provide independently verified colluders based on manual analysis of written work and confessions. For each power analysis, we generate virtual exams by subsampling either questions or students, run the full Q-SID pipeline using the class-size-specific thresholds established earlier, and compute TPR as the fraction of benchmark true-positives assigned to any reported collusion group.

To evaluate dependence on question number, we construct virtual exams by randomly subsampling questions from benchmark exam 1 and retaining the corresponding student-by-question score matrix. We then compute Q-SID outputs on each virtual exam and record the resulting TPR (Figure 10A). Exams with fewer than 20 questions are excluded because the resulting CS distribution is empirically unstable.

To study power as a function of complexity, we first pool question-score data across the two benchmark exams and then subsample varying numbers of questions to generate virtual exams spanning a range of complexity values. Q-SID is applied to each virtual exam, and TPR is computed as above (Figure 10B). When multiple exams are combined for analysis, the complexity of the

combined exam is taken to be the sum of the component exam complexities.

To evaluate dependence on class size, we generate virtual exams by randomly subsampling students from the benchmark exams across a range of class sizes (Figure 10C). For each virtual exam, Q-SID is applied using the class-size-specific thresholds and the same decision rules used elsewhere, including the exclusion of groups whose cumulative synthetic-FPR estimate exceeds 0.8%. TPR is then computed as the proportion of benchmark true-positives assigned to any reported collusion group.

Manual comparison of written answers

To determine if two or more students have colluded with sufficient certainty to charge them, the classic approach of manually comparing their written answers must be used. Collusion can be quickly ruled out for students who have not cheated. In our experience, a lack of similarity in detail in their answers is readily apparent after only a few minutes of comparing the group's exams. For collusion groups where answers appear overly similar on a first pass through the exam, a more time-consuming process is then required, which can take several hours. This involves first finding unlikely or unusual aspects of the suspect's answers to specific questions and then looking at all other students' answers. Where students have colluded, at least several examples can be found for which the only students giving a particular answer are the members of that group.

Certain questions are more dispositive of collusion than other. Within an instructors' style of question, those that elicit the greatest range of answers among students are typically those for which collusion can most clearly be demonstrated. Short-answer questions for which the correct answer closely follows the lecture notes will likely be very similar for many members of the class, most of whom did not collude.

Among the 263 students who took both benchmark exam datasets from two unproctored, online exams in a UC Berkeley class during the pandemic, initially, many pairs of students were identified whose scores for each question had high Pearson correlation coefficients. Fifty of these students were subsequently found to have colluded by a time-consuming, manual analysis of written answers. The 50 colluded in 19 groups, composed of between two to five members. Forty-six of them were identified within the time allowed for formal charges to be made, and 37 of these (80%) subsequently confessed to collusion. The written answers of the nine students who did not confess contained as many compelling instances of collusion as the 37 who did confess and the four who went uncharged. All 50 students identified by textual analysis of answers are considered to be true-positives.

As part of the month-long process to identify the true-positives, the written answers of an additional 26 pairs of students whose question scores correlated highly were also compared and found to show no evidence of collusion. This set of 52 students are designated as true-negatives. There are >30,000 additional pairwise combinations of student exams for this class whose written answers could not be compared manually. The Pearson correlation coefficients between these other pairs' question scores were generally lower than those of the true-positives or true-negatives, and thus these other pairs are unlikely to have colluded. These other students are not included in the true-negative set and instead are referred to as uncharacterized students.

Manual analysis of written answers was also used to assess students who were placed into collusion groups by Q-SID for many of the other 32 unproctored exams. For two of these exams (STEM 101 Fall 2020, exam 1, and STEM 101 Summer 2021, exam 3), full information on the results of this analysis have been provided by their instructors. A total of 611 students took these two exams, with Q-SID placing 61 of them into 23 collusion groups. Textual analysis showed that 21 of these groups had colluded (57 students), all of whom confessed and all of whom were found responsible for collusion by a university disciplinary process. The number of false-positives observed (4) is consistent with that expected from the empirical FPR (3.4), further supporting the accuracy of Q-SID's FPRs. Feedback from the instructors of other 30 unproctored exams was more fragmentary, but was consistent with Q-SID being effective in those cases where manual analysis was conducted.

Evaluation metrics

The Jaccard index⁴¹ measures overlap between a detected collusion group $\hat{G}$ and a verified collusion group G ,

$$J(G, \hat{G}) = \frac{|G \cap \hat{G}|}{|G \cup \hat{G}|}.$$

It takes values in $[0, 1]$, where larger values indicate closer agreement in group membership, 1 indicates identical membership, and 0 indicates no shared students. We used the Jaccard index because it directly quantifies how well a detected group recovers the members of a verified collusion group.

To assess agreement at the level of the full class partition, we used the ARI.⁴² Let $\{G_1, \dots, G_r\}$ and $\{\hat{G}_1, \dots, \hat{G}_s\}$ denote the verified and detected partitions of all students, and let $n_{uv} = |G_u \cap \hat{G}_v|$, $a_u = \sum_v n_{uv}$, and $b_v = \sum_u n_{uv}$. Then

$$\text{ARI} = \frac{\sum_{u,v} \binom{n_{uv}}{2} - \frac{\sum_u \binom{a_u}{2} \sum_v \binom{b_v}{2}}{\binom{n}{2}}}{\frac{1}{2} \left[\sum_u \binom{a_u}{2} + \sum_v \binom{b_v}{2} \right] - \frac{\sum_u \binom{a_u}{2} \sum_v \binom{b_v}{2}}{\binom{n}{2}}}.$$

The ARI equals 1 for identical partitions and is approximately 0 under chance-level agreement, with negative values possible in unfavorable cases. We used the ARI because it evaluates recovery of the overall collusion structure across all students and penalizes both over-merging and over-splitting of groups.

Q-SID algorithm

Algorithm: Q-SID

Input: a student-by-question score matrix $\mathbf{X} \in \mathbb{R}_{\geq 0}^{n \times p}$, with n students and p questions

Internal parameters:

Collusion score (CS) thresholds $0 < c_1 < c_2 < c_3 < c_4$ given the class size n ;

Empirical FPR levels $f_1 = 0.04\%$, $f_2 = 0.20\%$, and $f_3 = 0.56\%$.

Step 1 (collusion scores and partners). For each student $i = 1, \dots, n$,

1. Calculate IS_{ij} for every other student $j \neq i$.

2. Calculate $IM_i = \max_{j \neq i} IS_{ij} - \text{median}_{j \neq i} IS_{ij}$.

3. Calculate $CS_i = IM_i / (\text{local median } IM_i)$, where the local median IM_i is the median IM of student i 's neighbors (defined based on the test score rank).

4. Identify student i 's first partner $i_1 = \arg \max_{j \neq i} IS_{ij}$ and 2nd second partner $i_2 = \arg \max_{j \neq i, i_1} IS_{ij}$.

Step 2 (provisional collusion groups). Identify the {student, 1st partner} sets $\{i, i_1\}$

such that $CS_i \geq c_1$, $CS_{i_1} \geq c_1$, and at least one of CS_i and CS_{i_1} is no less than c_2 , with $i \neq i_1 \in \{1, \dots, n\}$. Combine the overlapped sets into K non-overlapping provisional collusion groups $\mathcal{T}_1, \dots, \mathcal{T}_K$, each of which is a set of students.

The K groups are ordered by the maximum CS of the students in each group, in a decreasing order.

Step 3 (collusion groups). Detect collusion groups $\mathcal{G}_1, \dots, \mathcal{G}_M$, where $M \leq K$. Initialize $m = 1$.

For $k = 1, \dots, K$,

- If $\mathcal{T}_k \neq \emptyset$:

1) Initialize $\mathcal{G}_m = \mathcal{T}_k$.

2) Identify the duplicate second partner set.

$\mathcal{P}_k = \{l \in \{1, \dots, n\} \setminus \mathcal{T}_k : \exists i \neq j \in \mathcal{T}_k \text{ s.t. } i_2 = j_2 = l\}$,

where i_2 and j_2 are students i and j 's second partners, respectively.

3) For each student l in $\mathcal{P}_k$, if $\exists \mathcal{T}_s, s > k$, s.t. $l \in \mathcal{T}_s$:

- update $\mathcal{G}_m = \mathcal{G}_m \cup \mathcal{T}_s$,

- update $\mathcal{T}_s = \emptyset$.

4) Update $m = m + 1$.

- Else if $\mathcal{T}_k = \emptyset$, skip.

Step 4 (empirical FPRs). For each collusion group $\mathcal{G}_m, m = 1, \dots, M$, denote its maximum group member CS value as $CS_m^* = \max_{i \in \mathcal{G}_m} CS_i$, which is at least c_2 by the definition of $\mathcal{G}_m$.

1) If $CS_m^* \geq c_4$, assign $\mathcal{G}_m$ with the empirical FPR empirical $FPR_m = f_1$.

2) If $c_3 \leq CS_m^* < c_4$, assign $\mathcal{G}_m$ with the empirical FPR empirical $FPR_m = f_2$.

3) If $c_2 \leq CS_m^* < c_3$, assign $\mathcal{G}_m$ with the empirical FPR empirical $FPR_m = f_3$.

Step 5 (synthetic FPRs: calibration of empirical FPRs).

1) Simulate a synthetic null student-by-question score matrix $\tilde{\mathbf{X}} \in \mathbb{R}_{\geq 0}^{n \times p}$ by fitting a null model on $\mathbf{X}$. In the null model, no students colluded.

2) Determine the synthetic null collusion group $\tilde{\mathcal{G}}_m, m = 1, \dots, M$, and their corresponding empirical FPRs $\in \{f_1, f_2, f_3\}$ using the synthetic null data $\tilde{\mathbf{X}}$ by following step 1–4. Then each synthetic null student in any of $\tilde{\mathcal{G}}_1, \dots, \tilde{\mathcal{G}}_M$ is assigned with an empirical FPR of f_1, f_2 , or f_3 .

3) For each of $l = 1, 2, 3$, calculate the proportion of synthetic null students that are assigned with the empirical FPR level f_l ; denote the proportion as $\tilde{f}_l$.

4) For each real collusion group $\mathcal{G}_m, m = 1, \dots, M$,

(1) if $CS_m^* \geq c_4$, assign $\mathcal{G}_m$ with the synthetic FPR synthetic $FPR_m = \tilde{f}_1$;

(2) if $c_3 \leq CS_m^* < c_4$, assign $\mathcal{G}_m$ with the synthetic FPR synthetic $FPR_m = \tilde{f}_2$;

(3) if $c_2 \leq CS_m^* < c_3$, assign $\mathcal{G}_m$ with the synthetic FPR synthetic $FPR_m = \tilde{f}_3$.

Output: Student collusion groups $\mathcal{G}_1, \dots, \mathcal{G}_M$ with their corresponding empirical FPRs (empirical $FPR_1, \dots$, empirical FPR_M) and synthetic FPRs synthetic ($FPR_1, \dots$, synthetic FPR_M).

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Mark D. Biggin (markbiggin@sbcglobal.net).

Materials availability

This study did not generate new materials.

Data and code availability

The deidentified supplemental data supporting this study are available through the Open Science Framework: <https://doi.org/10.17605/OSF.IO/5YQV6> under the CC0 1.0 Universal Public Domain Dedication.⁴³ The version of the Q-SID source code used in this study is archived through the Open Science Framework: <https://doi.org/10.17605/OSF.IO/47ZGP> and distributed under the Q-SID Academic License included in the archived record.⁴⁴

ACKNOWLEDGMENTS

We are grateful to various instructors at UC Berkeley and UCLA who provided question score data from their classes as well as feedback on the effectiveness of Q-SID in identifying collusion. This work was funded in part by a Sloan Research Fellowship, a Johnson and Johnson WiSTEM2D Scholars Award, a Guggenheim Fellowship, and the NIGMS R35GM140888 to J.J.L.

AUTHOR CONTRIBUTIONS

Conceptualization, M.D.B.; methodology, G.Y., J.J.L., and M.D.B.; and writing, G.Y., J.J.L., and M.D.B.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

During the preparation of this work, the authors used ChatGPT in order to refine the language. After using this tool or service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.patter.2026.101642>.

Received: March 2, 2026

Revised: May 8, 2026

Accepted: July 27, 2026

REFERENCES

- Baird, J.S. (1980). Current trends in college cheating. *Psychol. Sch.* 17, 515–522. [https://doi.org/10.1002/1520-6807\(198010\)17:4<515::aid-pits2310170417>3.0.co;2-3](https://doi.org/10.1002/1520-6807(198010)17:4<515::aid-pits2310170417>3.0.co;2-3).
- McCabe, D.L. (2005). Cheating among college and university students: A North American perspective. *Int. J. Educ. Integr.* 1. <https://doi.org/10.21913/ijei.v1i1.14>.
- Fox, J.P., and Meijer, R.R. (2008). Using item response theory to obtain individual information from randomized response data: An application using cheating data. *Appl. Psychol. Meas.* 32, 595–610. <https://doi.org/10.1177/0146621607312277>.
- Watson, G.R., and Sottile, J. (2010). Cheating in the digital age: Do students cheat more in online courses? *Online J. Dist. Learn. Adm.* 13.
- Fask, A., Englander, F., and Wang, Z. (2014). Do online exams facilitate cheating? An experiment designed to separate possible cheating from the effect of the online test taking environment. *J. Acad. Ethics* 12, 101–112. <https://doi.org/10.1007/s10805-014-9207-1>.
- Diedenhofen, B., and Musch, J. (2017). PageFocus: Using paradata to detect and prevent cheating on online achievement tests. *Behav. Res. Methods* 49, 1444–1459. <https://doi.org/10.3758/s13428-016-0800-7>.
- Dendir, S., and Maxwell, R.S. (2020). Cheating in online courses: Evidence from online proctoring. *Comput. Hum. Behav. Rep.* 2, 100033. <https://doi.org/10.1016/j.chbr.2020.100033>.
- Bilen, E., and Matros, A. (2021). Online cheating amid COVID-19. *J. Econ. Behav. Organ.* 182, 196–211. <https://doi.org/10.1016/j.jebo.2020.12.004>.
- Angoff, W.H. (1974). The development of statistical indices for detecting cheaters. *J. Am. Stat. Assoc.* 69, 44–49. <https://doi.org/10.2307/2285498>.
- Frary, R.B., Tideman, T.N., and Watts, T.M. (1977). Indices of cheating on multiple-choice tests. *J. Educ. Stat.* 2, 235–256. <https://doi.org/10.3102/10769986002004235>.
- Belleza, F.S., and Belleza, S.F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teach. Psychol.* 16, 151–155.
- Jennings, J.S., Harpp, D.N., and Hogan, J.J. (1996). Crime in the classroom: Part II. An update. *J. Chem. Educ.* 73, 349. <https://doi.org/10.1021/ed073p349>.
- Wesolowsky, G.O. (2000). Detecting excessive similarity in answers on multiple choice exams. *J. Appl. Stat.* 27, 909–921. <https://doi.org/10.1080/02664760050120588>.
- McManus, I.C., Lissauer, T., and Williams, S.E. (2005). Detecting cheating in written medical examinations by statistical analysis of similarity of answers: pilot study. *BMJ* 330, 1064–1066. <https://doi.org/10.1136/bmj.330.7499.1064>.
- Sotaridona, L.S., Van Der Linden, W.J., and Meijer, R.R. (2006). Detecting answer copying using the kappa statistic. *Appl. Psychol. Meas.* 30, 412–431. <https://doi.org/10.1177/0146621606288891>.
- Andries van der Ark, L., Emons, W.H.M., and Sijtsma, K. (2008). Detecting Answer Copying Using Alternate Test Forms and Seat Locations in Small-Scale Examinations. *J. Educ. Meas.* 45, 99–117. <https://doi.org/10.1111/j.1745-3984.2007.00055.x>.
- DiSario, R., Olinsky, A., Quinn, J., and Schumacher, P. (2009). Applying Monte Carlo simulation to determine the likelihood of cheating on a multiple-choice professional exam. *Case Studies In Business, Industry And Government Statistics* 3, 30–36.
- van der Linden, W.J., and Lewis, C. (2015). Bayesian checks on cheating on tests. *Psychometrika* 80, 689–706. <https://doi.org/10.1007/s11336-014-9409-x>.
- Richmond, P., and Roehner, B.M. (2015). The detection of cheating in multiple choice examinations. *Phys. Stat. Mech. Appl.* 436, 418–429. <https://doi.org/10.1016/j.physa.2015.05.040>.
- Romero, M., Riascos, Á., and Jara, D. (2015). On the optimality of answer-copying indices: Theory and practice. *J. Educ. Behav. Stat.* 40, 435–453. <https://doi.org/10.3102/1076998615595628>.
- Maynes, D.D. (2016). Detecting potential collusion among individual examinees using similarity analysis. In *Handbook of quantitative methods for detecting cheating on tests*, G.J. Cizek and J.A. Wollack, eds. (Routledge), pp. 47–69.
- Fendler, R.J., Yates, M., and Godbey, J. (2018). Observing and Detering Social Cheating on College Exams. *ij-sotl*. 12, 4. <https://doi.org/10.20429/ij-sotl.2018.120104>.
- Lin, M.J., and Levitt, S.D. (2020). Catching cheating students. *Economica* 87, 885–900. <https://doi.org/10.1111/ecca.12331>.
- Hanson, B., Harris, D., and Brennan, R. (1987). A comparison of several statistical methods for examining allegations of copying (ACT Research Report Series No. 87-15). Iowa City, IA: (American College Testing).
- Van der Linden, W.J., and Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *J. Educ. Behav. Stat.* 31, 283–304. <https://doi.org/10.3102/10769986031003283>.
- Maynes, D. (2014). Detection of non-independent test taking by similarity analysis. In *Test Fraud*. Routledge, N. Kingston and A. Clark, eds., pp. 53–80.
- Wollack, J.A. (1997). A nominal response model approach for detecting answer copying. *Appl. Psychol. Meas.* 21, 307–320. <https://doi.org/10.1177/01466216970214002>.
- Trout, N., and Gorney, K. (2025). Weighted Answer Similarity Analysis. *Appl. Psychol. Meas.* <https://doi.org/10.1177/01466216251322353>.
- Belov, D.I., and Wollack, J.A. (2021). Graph theory approach to detect examinees involved in test collusion. *Appl. Psychol. Meas.* 45, 253–267. <https://doi.org/10.1177/01466216211013902>.
- Eckerly, C. (2021). Answer similarity analysis at the group level. *Appl. Psychol. Meas.* 45, 299–314. <https://doi.org/10.1177/01466216211013109>.
- Lee, V.R., Pope, D., Miles, S., and Zárate, R.C. (2024). Cheating in the age of generative AI: A high school survey study of cheating behaviors before and after the release of ChatGPT. *Comput. Educ. Artif. Intell.* 7, 100253. <https://doi.org/10.1016/j.caeai.2024.100253>.
- Chen, B., Lewis, C.M., West, M., and Zilles, C. (2024). Plagiarism in the age of generative ai: cheating method change and learning loss in an intro to CS course. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*. pp. 75–85.
- Wollack, J.A., and Maynes, D.D. (2016). Detection of Test Collusion Using Cluster Analysis. In *Handbook of Quantitative Methods for Detecting Cheating on Tests* (Routledge), pp. 124–150.
- Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* 16, 1289–1296. <https://doi.org/10.1038/s41592-019-0619-0>.
- Manoharan, S. (2019). Cheat-resistant multiple-choice examinations using personalization. *Comput. Educ.* 130, 139–151. <https://doi.org/10.1016/j.compedu.2018.11.007>.
- Nelsen, R. (2006). *An Introduction to Copulas* (Springer Series in Statistics).
- Sun, T., Song, D., Li, W.V., and Li, J.J. (2021). scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome Biol.* 22, 163. <https://doi.org/10.1186/s13059-021-02367-2>.
- Song, D., Wang, Q., Yan, G., Liu, T., Sun, T., and Li, J.J. (2024). scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nat. Biotechnol.* 42, 247–252. <https://doi.org/10.1038/s41587-023-01772-1>.
- McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*. 3, 861. <https://doi.org/10.21105/joss.00861>.

40. Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37, 38–44. <https://doi.org/10.1038/nbt.4314>.
41. Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat.* 37, 547–579.
42. Hubert, L., and Arabie, P. (1985). Comparing partitions. *J. Classif.* 2, 193–218. <https://doi.org/10.1007/bf01908075>.
43. Yan, G., Li, J.J., Biggin, M.D., and Supplementary, Data. (2026). Open Science Framework. <https://doi.org/10.17605/OSF.IO/5YQV6>.
44. Yan, G., Li, J.J., and Biggin, M.D. (2026). Q-SID Code and Software. Open Science Framework. <https://doi.org/10.17605/OSF.IO/47ZGP>.