A Fairness-Promoting Detection Objective with Applications in AI-Assisted Test Security

Michael Fauss and Ikkyu Choi
ETS Research Institute, Princeton, NJ
{mfauss, choi001}@ets.org

Abstract

A detection objective based on bounded groupwise false alarm rates is proposed to promote fairness in the context of test fraud detection. The paper begins by outlining key aspects and characteristics that distinguish fairness in test security from fairness in other domains and machine learning in general. The proposed detection objective is then introduced, the corresponding optimal detection policy is derived, and the implications of the results are examined in light of the earlier discussion. A numerical example using synthetic data illustrates the proposed detector and compares its properties to those of a standard likelihood ratio test.

1 Introduction

Test security refers to the policies, procedures, and technologies used to protect the integrity and fairness of tests. A key component of test security is *test fraud detection*, that is, detection of unauthorized access to content, tools, or third-party assistance. Statistical methods for test fraud detection have been researched since at least the 1920s (Bird, 1927, 1929), with significant advances happening in the late 20th and early 21st century (Sotaridona and Meijer, 2002; Wollack, 2003; van der Linden and Sotaridona, 2004)—see (Kingston and Clark, 2014; Cizek and Wollack, 2016) for comprehensive overviews. In recent years, however, several developments have significantly expanded both the scope and urgency of test fraud detection efforts:

- The COVID-19 pandemic prompted a sudden shift from testing in tightly controlled test centers to remote testing in environments chosen by the test takers. While this transition offered significant convenience (Zheng et al., 2021; St-Onge et al., 2022), it also introduced numerous new opportunities for cheating (Bilen and Matros, 2021; Janke et al., 2021; Newton and Essex, 2023).
- Generative artificial intelligence (GenAI) models are now powerful enough to solve or assist with a wide range of item types, from simple multiple-choice questions to free-form essays and coding exercises, making them highly effective tools for cheating. (Yan et al., 2023; Susnjak and McIntosh, 2024)
- There is a movement towards more socioculturally responsive (Bennett, 2023) and personalized (Bennett, 2024; Sinharay et al., 2025) assessments to promote fairness and better capture the growing diversity of knowledge and abilities in increasingly heterogeneous test taker populations. This shift has led to greater item variety, resulting in fewer test takers responding to the same items.

These developments have made test fraud detection increasingly challenging: impostors and proxy test takers are more difficult to identify in remote settings than in test centers; AI-generated responses are harder to detect than content copied from traditional sources; and typical response times are difficult to establish for items that have been answered by only a handful of test takers. Consequently, test security reviews tend to require more time, expertise and data than they did in the past.

One approach to addressing these challenges is to delegate tasks to various AI systems, both generative and predictive. Building on the examples above: facial recognition could help detect impostors; typing pattern anomalies could signal proxy test takers; AI-content detectors could identify non-authentic writing or speech; and trained models, rather than empirical distributions, could be used to flag abnormal response times.

However, this approach typically and rightly raises questions regarding the reliability, accuracy and fairness of decision made by AI systems, especially in the context of high-stakes tests. (Weber-Wulff et al., 2023; Perkins et al., 2024) While con-

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 3: Coordinated Session Papers, pages 107–114
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

siderable research is being devoted to making AI fairer, more transparent and more reliable, biases and differential treatment continue to be observed in practice. (Stureborg et al., 2024; Bai et al., 2025; Maslej et al., 2025)

In contrast, many methods traditionally used in psychometrics and, more broadly, decision-making under uncertainty, have transparent objectives and strong accuracy and/or fairness properties. (Dorans and Cook, 2016; Johnson et al., 2022) In this paper, we propose addressing the uncertainty and potential biases of AI outputs not by using them directly, but by feeding them into a system that fuses and processes them. In essence, the idea is to delegate complex subtasks to advanced AIs while anchoring the final decision-making procedure in traditional statistical methods, thereby enabling the use of well-established techniques to define, measure, and promote fairness.

To clarify, this paper does not address the design or architecture of the complete system described above, which remains an ongoing research effort. (Fauss et al., 2025) Instead, it focuses on a specific subtask: designing a detector that flags test takers for potential fraud in a way that balances test integrity and group fairness. This task is formulated and analyzed as a standalone problem, meaning the proposed detector is largely agnostic to the specific detection context. As such, it may be of theoretical or practical interest beyond the use case discussed here. However, as will become clear throughout the paper, its design is explicitly guided by assumptions tailored to the intended application of AI-assisted test fraud detection.

2 Fairness in Test Security

In this section, we discuss some aspects and characteristics that set fairness in test security apart from fairness as a general concept in statistics and machine learning. Specifically, we will make and justify five *claims*. These claims are not intended to be "truths"; rather, we see them as important, sometimes overlooked aspects that can contribute to a more informed discussion of what constitutes fairness in test security applications.

Claim 1: *Fairness and performance are not in conflict.*

A concept commonly encountered in the literature on statistical fairness is the so-called *performance–fairness tradeoff* (Prost et al., 2019), which implies that a procedure's performance and fairness are often in tension with one another. The underlying idea is that in order to make a procedure fairer, additional *constraints* have to be introduced that shrink the space of feasible solutions, and, in turn, reduce the performance. While this is true from a purely mathematical perspective, we would argue that the idea of a performance-fairness tradeoff can be misleading in a test security context. This is the case because detecting test fraud is in itself an objective that, in principle, *promotes fairness*. Among other consequences, widespread, undetected cheating devalues the scores of honest test takers, potentially harming their future opportunities. In general, we consider the idea that a procedure can be "bad" at its dedicated task, yet still perfectly fair problematic. One can even argue that fairness issues are a *consequence* of performance issues. A fraud detector achieving perfect accuracy is not only highly performant, it is also fair by all common criteria. Fairness issues arise once a procedure starts making mistakes, and certain groups are more frequently or more severely affected by these mistakes. Therefore, we argue that in the context of test security fairness and performance should be considered two sides of the same coin—often, a *better* detector will also be a *fairer* detector.

Claim 2: *Equality $\neq$ fairness.*

This claim is closely related to Claim 1. We single it out to highlight the critical role that *equality* plays in virtually all fairness criteria in the literature. For example, *separation fairness* (Barocas et al., 2023) is defined in terms of equal true and false positive rates among all groups. Analogously, *sufficiency fairness* (Barocas et al., 2023) implies that the probability of predicted labels being correct is equal for all groups. Again, we would argue that this idea can be misleading in a test security context. For example, a fraud detector that randomly declare test takers cheaters is perfectly fair by many criteria, yet clearly dysfunctional and unfair in practice. Similarity, by most fairness criteria, a detector with groupwise false alarm rates of, say, 30 % and 35 % is fairer than a detector with groupwise false alarm rates of, say, 5 % and 15 %. In reality, it is far from clear that test takers would view the higher false alarm rate of the first detector as fairer than the larger disparity in groupwise false alarm rates produced by the second.

Claim 3: *Fairness needs a concrete target.*

We argue that any nontrivial measure or intervention aimed at promoting fairness in test security must clearly specify the type of discrimination it seeks to address and provide strong evidence that it effectively mitigates or eliminates it. While this may seem obvious, our experience suggests it is not consistently implemented in practice. Frequently, existing detectors or classifiers are made fair by picking an arbitrary or convenient fairness criterion, adding a corresponding penalty term to the training objective, and adjusting its weight until a "good performance-fairness tradeoff" is reached. We believe that promoting fairness in this manner can be superficial and ineffective. It will typically lead to a slightly more uniform distribution of the groupwise metric the fairness criterion considers important. However, showing that the combined effects on all groups and on the overall performance really address unfair treatment is usually difficult. In fact, the case for this kind of fairness measure is often made in a circular manner: it promotes fairness *because* it improves the fairness criterion underpinning its design.

Claim 4: *Fairness should not be a black box.*

While Claim 3 argues that it should be clear *what* a fairness-promoting procedure tries to accomplish, here we argue that it should also be clear *how* the procedure promotes fairness. This claim is based on the observation that, in particular in test security, fairness is closely connected to trust and transparency. To clarify, we do not claim that one should be able to explain every technical detail of a fairness-promoting procedure to a non-technical audience. However, we do believe that a sincere attempt at making a procedure fairer should be implemented in way that, at least conceptually, can be communicated to those affected by it. This also opens the door for broader discussions of what constitutes fairness and how it can be improved.

Claim 5: *Fairness should be measurable.*

Naturally, the vast majority of statistical fairness criteria are defined in terms of *probabilities*. However, these probabilities are typically unknown and must be *estimated* from data. This can lead to problems when certain events occur so infrequently that reliably assigning them an empirical probability becomes infeasible. This problem is more prominent the smaller the population and the more groups are considered. For example, in the context of test fraud detection, a fairness criterion that incorporates groupwise cheating rates might run into the problem that for some groups no cheaters have been observed yet. Does this mean that the respective cheating rates are low? Or that the detection rates are low? Can, often self-declared, group variables of cheaters be trusted in the first place? In a nutshell, we argue that fairness should be based on quantities that can accurately and reliably be inferred from the data.

In the next section, we present a fairness promoting detection objective that is informed by and largely aligned with the above claims.

3 A Fairness-Promoting Detection Objective

In this section, we propose a fairness-promoting detection objective, derive the corresponding optimal detector, and discuss its properties in light of the claims in Section 2. While the intended use case of the proposed detector is test fraud detection, it is not limited to this context and likely has applications in other areas.

A quick note on notation: In what follows, uppercase letters, X, denote random variables, lowercase letters, x, denote their realizations, and boldface, $\boldsymbol{x}$, indicates vectors. Probability distributions are denoted by P, and probability density functions (PDFs) by p.

3.1 Problem Formulation

Let $N \in \mathbb{N}_{\geq 1}$ be the number of test takers. For every test taker we observe a random vector $\boldsymbol{X}_n \in \mathbb{R}^M$, $M \in \mathbb{N}_{\geq 1}$, which is a collection of relevant observations and features. In this paper, we do not make further assumptions about the nature or meaning of $\boldsymbol{X}$ or its elements. However, as discussed above, in the intended application of (AI-assisted) test fraud detection, $\boldsymbol{X}$ is assumed to consist of high-level features that themselves are outputs of AI systems (likelihood of AI-generated content, likelihood of copy-typing, likelihood of impostor, etc.).

In addition to the feature vector, we assume that a discrete random variable, $G_n \in \{1, \dots, N_G\}$, $N_G \in \mathbb{N}_{\geq 1}$ is observed for every test taker indicating membership in one of N_G groups. Every test taker is assumed to belong to exactly one group. These groups are typically defined by demographic attributes such as gender, race, age, or first language. However, depending on the application, one might also consider externally defined groups, such as test takers receiving a certain form or taking the test remotely versus in a test center.

Finally, we assume that every test taker is either fraudulent ("cheater") or honest ("non-cheater"). This is indicated by a binary random variable $C_n \in \{0, 1\}$, with $C_n = 1$ indicating a cheater and $C_n = 0$ indicating a non-cheater. Naturally, C_n is assumed to be a latent variable.

Finally, we assume that the feature vectors of all test takers are independent, conditioned on their group membership and honesty. That is, there are random variables $\boldsymbol{X}$, G and C such that

$$\boldsymbol{X}_n \,|\, (G_n = g, C_n = c) \overset{d}{=} \boldsymbol{X} \,|\, (G = g, C = c)$$

for all $n \leq N$, where $\overset{d}{=}$ denotes equality in distribution. Therefore, the index n is omitted in what follows. The assumption may not always hold in practice, but it offers a useful approximation that suffices for the discussion at hand.

The detector we seek to design is assumed to generate a random variable $\hat{C} \in \{0, 1\}$ that indicates whether the respective test taker is classified as cheater ($\hat{C} = 1$) or non-cheater ($\hat{C} = 0$). It is defined by a function $f \colon \mathbb{R}^M \to [0, 1]$ that maps a feature vector to a probability of classifying the corresponding test taker as a cheater, that is:

$$P(\hat{C} = 1 \,|\, \boldsymbol{X} = \boldsymbol{x}, G = g, C = c) = f(\boldsymbol{x}). \quad (1)$$

for all $\boldsymbol{x}$, g and c. Note that f is a function only of the feature vector, $\boldsymbol{x}$, but not of the group variable, g, even though g is known. This is intentional, as incorporating group information into a detector is generally considered problematic. Most importantly, it can lead to cases in which two test takers with identical feature vectors are classified differently depending on which group they belong to.

We next present the proposed detection objective:

$$\max_f \; P(\hat{C} = 1 \,|\, C = 1) \quad \text{s.t.} \quad (2)$$

$$P(\hat{C} = 1 \,|\, G = g, C = 0) \leq \alpha \quad \forall g \leq N_G, \quad (3)$$

where $\alpha \in (0, 1)$ is a free parameter. The constraints in (3) enforce an upper bound on the false alarm rate (FAR) of each group. We refer to a detector that satisfies these constraints as fair in the sense of bounded FARs, or *BFAR-fair* for short. For a given α, the objective in (2) picks the BFAR-fair detector with the highest detection rate. This problem formulation will be discussed and justified in more detail shortly.

3.2 Optimal Detector

The main result of this paper, a detector that is optimal in the sense of BFAR fairness, is stated in the following theorem:

Theorem 1. *The detector that solves the problem in* (2) *and* (3) *is given by*

$$\hat{C}^* = \begin{cases} 0, & g_{\boldsymbol{\lambda}^*}(\boldsymbol{x}) \leq 0 \\ 1, & g_{\boldsymbol{\lambda}^*}(\boldsymbol{x}) > 0 \end{cases}, \quad (4)$$

where

$$g_{\boldsymbol{\lambda}}(\boldsymbol{x}) = p(\boldsymbol{x} \,|\, C = 1) - \sum_{g=1}^{N_G} \lambda_g p(\boldsymbol{x} \,|\, G = g, C = 0) \quad (5)$$

and $\boldsymbol{\lambda}^*$ *is such that*

$$P[\hat{C}^* = 1 \,|\, G = g, C = 0] = \alpha \quad (6)$$

if $\lambda_g^* > 0$ *and*

$$P[\hat{C}^* = 1 \,|\, G = g, C = 0] < \alpha \quad (7)$$

if $\lambda_g^* = 0$.

Proof. The statement in the theorem can be proven using standard arguments in constrained optimization. The Lagrange dual (Boyd and Vandenberghe, 2004, Ch. 5.2) of the problem in (2) is given by

$$\min_{\boldsymbol{\lambda} \geq 0} \max_f \; L_\alpha(f, \boldsymbol{\lambda}), \quad (8)$$

where

$$L_\alpha(f, \boldsymbol{\lambda}) = P[\hat{C} = 1 \,|\, C = 1] - \sum_{g=1}^{N_G} \lambda_g P[\hat{C} = 1 \,|\, G = g, C = 0] + \sum_{g=1}^{N_G} \lambda_g \alpha.$$

By conditioning and marginalizing over $\boldsymbol{X}$ we can write L_α as

$$L_\alpha(f, \boldsymbol{\lambda}) = \int f(\boldsymbol{x}) g(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \alpha \sum_{g=1}^{N_G} \lambda_g, \quad (9)$$

where $g_{\boldsymbol{\lambda}}$ is defined in (5) and we used (1) to write the relevant probabilities in terms of f. Since L_α in (9) is linear in f, the maximizer of the inner problem in (8) is given by

$$f^*(\boldsymbol{x}) = \begin{cases} 0, & g_{\boldsymbol{\lambda}}(\boldsymbol{x}) \leq 0 \\ 1, & g_{\boldsymbol{\lambda}}(\boldsymbol{x}) > 0 \end{cases}. \quad (10)$$

It remains to show that the optimal Lagrange multiplier satisfy (6) and (7). However, this property follows immediately from the complementary slackness condition of the KKT conditions (Boyd and Vandenberghe, 2004, Ch. 5.5). Finally, note that for $f = f^*$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ the constraints in (3) are satisfied by construction, which in turn implies that the solution of the dual problem also solves the primal problem. (Boyd and Vandenberghe, 2004, Ch. 5.5) This completes the proof. □

3.3 Discussion

In this section, we discuss the problem formulation in (2) and (3) in more detail and explain why we consider BFAR fairness an appropriate and practical approach to promoting fairness in the context of (AI-assisted) test fraud detection.

1. BFAR fairness requires the detector to operate at a false alarm rate (type II error probability) below α for all groups. This means that, in the spirit of Claim 1, there is a *minimum performance level* that the detector needs to meet in order to be considered fair.

2. In the spirit of Claim 2, BFAR fairness *promotes* equality, but does not *enforce* it. As long as the error probabilities are acceptable for all groups, it does not detract from the detector's fairness if it performs better for some groups.

3. BFAR fairness deliberately constraints false alarm rates instead of alternative metrics, such as false discover rates or detection rates. This is in the spirit of Claims 3 and 4. BFAR fairness targets unfairness from the perspective of *honest test takers* and, consequently, can be communicated in a straightforward manner: For an honest taker, the probability of being falsely flagged by a BFAR-fair detector is at most α, irrespective of their race/age/first language etc. Appropriate values of α might be subject to debate, but we believe that both the target group and the concept of BFAR fairness are clear and transparent.

4. BFAR fairness does not require groupwise detection rates. This is in the spirit of Claim 5. For any reputable test, cheaters are a small minority of the test taker population. Therefore, as explained in the discussion of Claim 5, estimating groupwise detection rates is notoriously difficult for smaller groups. Moreover, groups can sometimes lose their meaning if the corresponding test taker committed fraud. For example, a native French speaker might copy an essay written by a native Mandarin speaker. Therefore, BFAR fairness avoids grouping cheaters in the first place.

5. By inspection of (4) and (5), the BFAR-fair detector is implemented via a modified likelihood ratio test. More specifically, it compares the likelihood of the observed feature vector under the cheater versus the honest hypothesis. However, while a standard likelihood ratio test marginalizes over the group variables using their true probabilities, the marginalization in the BFAR-fair test statistic in (5) is performed with custom weights, $\boldsymbol{\lambda}^*$, that do not necessarily reflect the actual group sizes. That is, the BFAR-fair detector is implemented by *re-weighting* or *oversampling* groups that would otherwise violate the false alarm rate constraints. While details on how to obtain these weights and how they enter the test statistic may be more intricate, the underlying idea of re-weighting or oversampling is well-established, conceptually simple, and easy to communicate—which aligns well with the spirit of Claim 4.

However, BFAR fairness also has its shortcomings. For example, two arguments against its use in operation are the following:

1. The detection rate and groupwise false alarm rates can not be observed directly, but have to be inferred based on some statistical model of the test taker population. This aspect can be argued to be in conflict with Claim 5. However, as discussed above, the quantities of interest were deliberately chosen to avoid problematic corner cases or small-sample scenarios, and we expect that they can typically be estimated with reasonable accuracy.

2. The focus on honest test takers can conflict with the goal of test integrity. Traditionally, fraud detectors are tuned to meet specific detection rate requirements, accepting potentially high false alarm rates as a necessary cost. From the BFAR perspective, one first determines a justifiable burden on honest test takers and then accepts the corresponding detection rates. On the one hand, this approach can be difficult to defend in practice. On the other hand, in the spirit of Claim 1, any detector that can only satisfy integrity standards by imposing an unacceptable burden on honest test takers may not be ready for operational use.

In summary, while BFAR fairness may not be suitable or implementable in every setting and application, we believe that it is a useful, transparent, practical and well-justified approach to promoting fairness in test fraud detection.

Table 1: Groupwise false alarm rates of likelihood ratio and BFAR-fair detector with detection rate of $\approx 87\,\%$.

	False Alarm Rate		
Detector	$G=1$	$G=2$	$G=3$
Likelihood Ratio	0.0416	0.1298	0.0825
BFAR-fair	0.0835	0.1011	0.0997

4 Numerical Example

In this section, we demonstrate the BFAR fair detector proposed in the previous section with a numerical example. Since it is merely supposed to provide a proof of concept, we deliberately keep this example simple. Specifically, we assume that the test taker population consists of three equally likely groups of interest ($N_G = 3$) and that two features ($M = 2$) are observed for each test taker. In line with the assumption that these features are themselves probabilities of a test taker having committed fraud, likely generated by large, high-level AI models, we assume that the feature vectors are distributed on the unit square. We model these features via a multivariate beta distribution in (Fauss, 2024). The exact parameters for each group are given in Appendix A.

In order to establish a baseline performance, and in light of Comment 5 in Section 3.3, we compare the proposed BFAR-fair detector to a standard likelihood ratio test, that is, a detector with decision rule

$$\hat{C} = \begin{cases} 1, & \frac{p(\boldsymbol{x}\,|\,C=1)}{p(\boldsymbol{x}\,|\,C=0)} \geq \nu \\ 0, & \frac{p(\boldsymbol{x}\,|\,C=1)}{p(\boldsymbol{x}\,|\,C=0)} < \nu \end{cases}, \tag{11}$$

where $\nu \in (0,1)$ is a threshold that balances the detection and false alarm rates.

We set the parameter of the BFAR-fair test to $\alpha = 0.1$, that is, the false alarm rate must not exceed $10\,\%$ for any group. The corresponding weights, $\boldsymbol{\lambda}^*$, were determined by numerically solving the optimality conditions in Theorem 1 and are given by $\boldsymbol{\lambda}^* \approx (0, 0.7682, 0.4266)$. The probabilities on the left-hand sides of (6) and (7) were approximated by sampling from the specified distributions. The threshold ν was selected so that the detection rate of the likelihood ratio detector matches that of the BFAR-fair detector, which was evaluated to $87\,\%$ in this case. Again, we used sampling to approximate this rate. The resulting groupwise false alarm rates for both detectors are reported in Table 1.

By inspection, the false alarm rates of the likelihood ratio detector vary substantially across groups, ranging from just above $4\,\%$ for group 1 to nearly $13\,\%$ for group 2. In contrast, by design, the BFAR-fair detector keeps all false alarm rates below the $10\,\%$ threshold. Note that while the false alarm rates for groups 2 and 3 are close to this threshold, the rate for group 1 is lower by a margin that cannot be attributed to approximation errors alone. This gap is consistent with the first element of $\boldsymbol{\lambda}^*$ being zero, which indicates that the false alarm rate constraint for group 1 is non-binding. In fact, the BFAR detector uses effective group probabilities/sizes of $P(G=1) = 0$, $P(G=2) \approx 0.64$ and $P(G=3) \approx 0.26$. In words, the assumed probability of group 3 remains close to its true value of $\frac{1}{3}$, the probability of group 2 approximately doubles, increasing its influence on the test statistic, while the effective size of group 1 set to zero, effectively ignoring it in the calculation of the non-cheater likelihood. This implies that the false alarm rate constraint for group 1 is redundant given the constraints for groups 2 and 3.

The decision boundaries of the two detectors are shown in Figure 1. For illustration purposes, Figure 1 also shows samples of feature vectors drawn from the respective distributions. Both decision boundaries approximately follow the negative diagonal of the unit square, with a noticeable "bulge" in the region where the feature distribution of honest test takers in group 2 strongly overlaps with that of the cheaters. However, the bulge is much more pronounced in case of the BFAR-fair detector. This increased lenience towards test takers in group 2 is (partially) compensated by tightening the decision boundary in the upper left region, which is unlikely to contain members of group 2. This adjustment explains the observed increase in false alarm rates for test takers in groups 1 and 3.

In summary, at the same detection rate, the BFAR-fair detector admits a significantly more uniform false alarm rate profile compared to a standard likelihood ratio test and keeps the "worst case" false alarm rate across all groups below the targeted $10\,\%$. On the downside, the overall false alarm rate, which, in this case, is given by the average of the groupwise rates, increases from $8.5\,\%$ to $9.5\,\%$. Whether or not this drawback outweighs the benefits of the BFAR-fair detector has to be evaluated on a case-by-case basis. We hope that the discussions in Section 2 and 3.3 provide valuable guidelines for this evaluation.

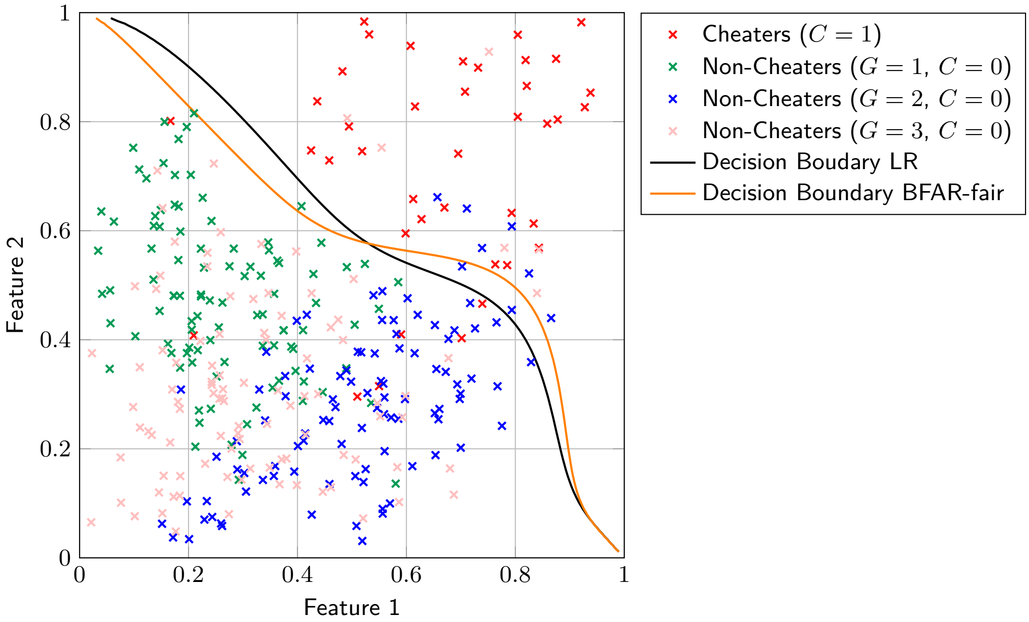

Figure 1: Feature sample and decision boundaries of BFAR-fair and likelihood ratio detector.

References

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2025. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8):e2416228122.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, Cambridge, MS, USA.

Randy E. Bennett. 2023. Toward a theory of socioculturally responsive assessment. *Educational Assessment*, 28(2):83–104.

Randy E. Bennett. 2024. Personalizing assessment: Dream or nightmare? *Educational Measurement: Issues and Practice*, 43(4):119–125.

Eren Bilen and Alexander Matros. 2021. Online cheating amid COVID-19. *Journal of Economic Behavior & Organization*, 182:196–211.

Charles Bird. 1927. The detection of cheating in objective examinations. *School & Society*, 25:261–262.

Charles Bird. 1929. An improved method of detecting cheating in objective examinations. *The Journal of Educational Research*, 19(5):341–348.

Stephen Boyd and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

Gregory J. Cizek and James A. Wollack, editors. 2016. *Handbook of Quantitative Methods for Detecting Cheating on Tests*. Routledge, New York City, NY, USA.

Neil J. Dorans and Linda L. Cook. 2016. *Fairness in Educational Assessment and Measurement*. Routledge.

Michael Fauss. 2024. tmvbeta: Truncated multivariate beta distribution on the unit hypercube.

Michael Fauss, Xiang Liu, Chen Li, Ikkyu Choi, and H. Vincent Poor. 2025. Bayesian selection policies for human-in-the-loop anomaly detectors with applications in test security. Under review for publication in Psychometrika.

Stefan Janke, Selma C. Rudert, Änne Petersen, Tanja M. Fritz, and Martin Daumiller. 2021. Cheating in the wake of COVID-19: How dangerous is ad-hoc online testing for academic integrity? *Computers and Education Open*, 2:100055.

Matthew S. Johnson, Xiang Liu, and Daniel F. McCaffrey. 2022. Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. *Journal of Educational Measurement*, 59(3):338–361.

Neil Kingston and Amy Clark. 2014. *Test Fraud: Statistical Detection and Methodology*. Routledge Research in Education. Taylor & Francis, Milton Park, Abingdon-on-Thames, UK.

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, and 4 others. 2025. Artificial intelligence index report 2025. Annual report, Stanford University Institute for Human-Centered AI.

Philip M. Newton and Keioni Essex. 2023. How common is cheating in online exams and did it increase during the COVID-19 pandemic? A systematic review. *Journal of Academic Ethics*, pages 1–21.

Mike Perkins, Jasper Roe, Binh H. Vu, Darius Postma, Don Hickerson, James McGaughran, and Huy Q. Khuat. 2024. Simple techniques to bypass genAI text detectors: implications for inclusive education. *International Journal of Educational Technology in Higher Education*, 21(1).

Flavien Prost, Hai Qian, Qiuwen Chen, Ed H. Chi, Jilin Chen, and Alex Beutel. 2019. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *ArXiv*, abs/1910.11779.

Sandip Sinharay, Randy E. Bennett, Michael Kane, and Jesse R. Sparks. 2025. Validation for personalized assessments: A threats-to-validity approach. *Journal of Educational Measurement*.

Leonardo S. Sotaridona and Rob R. Meijer. 2002. Statistical properties of the k-index for detecting answer copying. *Journal of Educational Measurement*, 39(2):115–132.

Christina St-Onge, Kathleen Ouellet, Sawsen Lakhal, Tim Dubé, and Mélanie Marceau. 2022. COVID-19 as the tipping point for integrating e-assessment in higher education practices. *British Journal of Educational Technology*, 53(2):349–366.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *Preprint*, arXiv:2405.01724.

Teo Susnjak and Timothy R. McIntosh. 2024. Chatgpt: The end of online exam integrity? *Education Sciences*, 14(6).

Wim J. van der Linden and Leonardo Sotaridona. 2004. A statistical test for detecting answer copying on multiple-choice tests. *Journal of Educational Measurement*, 41(4):361–377.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1).

James A. Wollack. 2003. Comparison of answer copying indices with real data. *Journal of Educational Measurement*, 21(4):307–320.

Duanli Yan, Michael Fauss, Jiangang Hao, and Wenju Cui. 2023. Detection of AI-generated essays in writing assessments. *Psychological Test and Assessment Modeling*, 65(1):125–144.

Meixun Zheng, Daniel Bender, and Cindy Lyon. 2021. Online learning during COVID-19 produced equivalent or better student course performance as compared with pre-pandemic: empirical evidence from a school-wide comparative study. *BMC medical education*, 21:1–11.

A Simulation Parameters

Let the parameters of the multivariate beta distribution in (Fauss, 2024) be denoted by $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{\Sigma}$.In our simulation, the feature distribution of the cheaters was assumed to be independent of the group and given by:

$C = 1$:

$$\begin{aligned} \boldsymbol{a}_1 &= \begin{bmatrix} 4 & 4 \end{bmatrix} \\ \boldsymbol{b}_1 &= \begin{bmatrix} 2 & 2 \end{bmatrix}, \\ \boldsymbol{\Sigma}_1 &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned}$$

The feature distribution of honest test takers was modeled groupwise with parameters:

$C = 0, G = 1$:

$$\begin{aligned} \boldsymbol{a}_{01} &= \begin{bmatrix} 2 & 4 \end{bmatrix} \\ \boldsymbol{b}_{01} &= \begin{bmatrix} 6 & 4 \end{bmatrix}, \\ \boldsymbol{\Sigma}_{01} &= \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} \end{aligned}$$

$C = 0, G = 2$:

$$\begin{aligned} \boldsymbol{a}_{02} &= \begin{bmatrix} 4 & 2 \end{bmatrix} \\ \boldsymbol{b}_{02} &= \begin{bmatrix} 4 & 6 \end{bmatrix}, \\ \boldsymbol{\Sigma}_{02} &= \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \end{aligned}$$

$C = 0, G = 3$:

$$\begin{aligned} \boldsymbol{a}_{03} &= \begin{bmatrix} 2 & 2 \end{bmatrix} \\ \boldsymbol{b}_{03} &= \begin{bmatrix} 4 & 4 \end{bmatrix}, \\ \boldsymbol{\Sigma}_{03} &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned}$$