Decision Thresholds and Policy
Volume 1 · Number 1 · 2026

Calibration Targets and Held-Identity Face-Threshold Tradeoffs

William Cooper[1] · Eric Palmer[2] · Evan Anderson[3]

[1] Cyprus · [2] Malta · [3] Germany

Original publication: November 5, 2026 | Article 624

Abstract

Thresholds calibrated to one- and five-percent false acceptance did not reproduce those rates in a small held-identity comparison. The two thresholds accepted zero and ten of 1,000 held nonmatching pairs, while accepting 250 and 372 of 900 matching pairs. The more permissive threshold increased true acceptance by 13.56 percentage points, with a ten-block interval of 9.51 to 17.60 points. Calibration used separate identities and fixed pixel distances, not learned face embeddings. These results demonstrate a conditional threshold tradeoff. Zero observed false accepts is not a zero-risk guarantee, and the historical image collection cannot validate a video-identity metric or a deployment security threshold.

An illustrative pixel-similarity verification study with identities separated before threshold selection.

Calibration Targets and Held-Identity Face-Threshold Tradeoffs

An illustrative pixel-similarity verification study with identities separated before threshold selection.

Abstract

Thresholds calibrated to one- and five-percent false acceptance did not reproduce those rates in a small held-identity comparison. The two thresholds accepted zero and ten of 1,000 held nonmatching pairs, while accepting 250 and 372 of 900 matching pairs. The more permissive threshold increased true acceptance by 13.56 percentage points, with a ten-block interval of 9.51 to 17.60 points. Calibration used separate identities and fixed pixel distances, not learned face embeddings. These results demonstrate a conditional threshold tradeoff. Zero observed false accepts is not a zero-risk guarantee, and the historical image collection cannot validate a video-identity metric or a deployment security threshold.

1. A calibration rate names a sample and a rule

A similarity threshold can be chosen to accept a small percentage of nonmatching calibration pairs. Its numerical target is then sometimes used as though it guaranteed the same false-accept rate for new identities. Such transport depends on the score distributions and the population of pairs being evaluated. The threshold itself remains fixed, while the people, image conditions, and composition of nonmatching comparisons can change. A benchmark must identify both the calibration population and the later evaluation population.

We ask how empirical one- and five-percent calibration targets transfer to an identity-disjoint evaluation in a small public face collection. We also measure the change in true acceptance when the threshold is relaxed. The comparison is deliberately elementary: centered and normalized grayscale vectors supply the score, and no learned face representation is fitted. Its purpose is to expose what a threshold designation means, rather than to achieve competitive verification accuracy.

The subject-preservation setting represented by Meng et al., 2026 motivates clarity about identity scores and their decision thresholds. We run no generation model and inspect no generated video. A static pair-verification result cannot establish preservation of a subject across synthesized frames. It can nevertheless illustrate why a threshold should be accompanied by its calibration sample, realized later counts, and retained true-match rate before it becomes evidence for a broader identity claim.

2. Images and identity separation

The source in AT&T Laboratories Cambridge, 2026 contains four hundred historical grayscale photographs of forty released identities, ten images per identity. We use the modified MATLAB mirror whose checksum matches the maintained scikit-learn loader and retain AT&T Laboratories Cambridge attribution. The study context is associated with Samaria and Harter, 1994. Source ordinal identity labels provide the matching reference; they do not provide demographic annotations or a representative sample of contemporary device users.

With random seed 20261107 we permute the forty identity labels and assign twenty to calibration and twenty to evaluation. Every photograph of an identity remains on one side. Thus threshold selection cannot see a later evaluated person's images through a different photo in the calibration set. This design differs from splitting images within every identity, where the same people occur on both sides. Neither design alone establishes transfer to another dataset or camera environment.

The constrained historical images differ from the unconstrained recognition setting represented by Huang et al., 2008. We include no other photographic dataset in the executed calculation. Expressions, lighting, framing, and source acquisition can influence distances, and the split creates only one realization of those distributions. We do not infer fairness across protected groups from ordinal identities or describe source labels as a complete account of how an individual would appear in a new setting.

3. Fixed distances and two empirical thresholds

We flatten each image, subtract its own mean, and divide by the Euclidean norm of its centered vector. Every norm is positive. Squared Euclidean distance between these vectors is the pair score, with smaller distance indicating greater similarity under this representation. The same operation applies to calibration and held images. There is no training of features, projection dimension, probability mapping, or identity classifier.

The projection approaches in Turk and Pentland, 1991 and Belhumeur et al., 1997 provide historical context, but neither is implemented. Modern learned methods represented by Schroff et al., 2015, Deng et al., 2018, Wang et al., 2018, Kim et al., 2022, and Meng et al., 2021 define different representations and objectives. Their reported behavior cannot be assigned to centered pixels. The intentionally simple score makes the present calibration transparent while limiting its relevance as a recognition system.

Among the twenty calibration identities we enumerate every unordered cross-identity pair of identities and all one-hundred image pairs for that identity pair. This yields 19,000 nonmatching image pairs. For each nominal target we take the corresponding empirical distance quantile using the lower-order convention and accept distances at or below the selected value. The thresholds are 0.477244974 and 0.601268849, realizing exactly one and five percent acceptance in the calibration nonmatches.

Only calibration nonmatches choose the thresholds. Held matching or nonmatching outcomes do not adjust them. The nominal rates are illustrative design choices, not security requirements or performance standards. Calibration concerns in Guo et al., 2017 are pertinent because selecting a score quantile does not create a calibrated probability that two photographs belong to the same person. The numerical threshold remains a property of this score and its particular calibration distribution.

4. Ten disjoint identity blocks define the held comparison

The twenty held identities are paired in their seeded order to form ten blocks, each containing two identities and twenty photographs. Every block supplies one hundred cross-identity nonmatches. It also supplies all forty-five within-identity pairs for each of its two identities, giving ninety matching pairs. Across blocks there are 1,000 nonmatches and 900 matches. No identity or image occurs in more than one block.

We intentionally omit other held cross-identity pairs. Including all of them would create many comparisons that share identities across evaluation groups, complicating uncertainty accounting. The ten-block design preserves the full matching-pair set while using a smaller, explicitly defined nonmatch set. Its composition is part of the estimand; 1,000 pairs should not be presented as all possible held nonmatches or as one thousand independent draws of people.

Decision-cost perspectives in Elkan, 2001 and selective-decision work in Geifman and El-Yaniv, 2017 motivate retaining both error and acceptance information. A lower false-accept count can be accompanied by a loss of true matches, and no application cost is supplied here to identify an optimal threshold. We do not choose a winner by an unstated balance between those outcomes. The experiment reports the observed tradeoff at the two previously selected values.

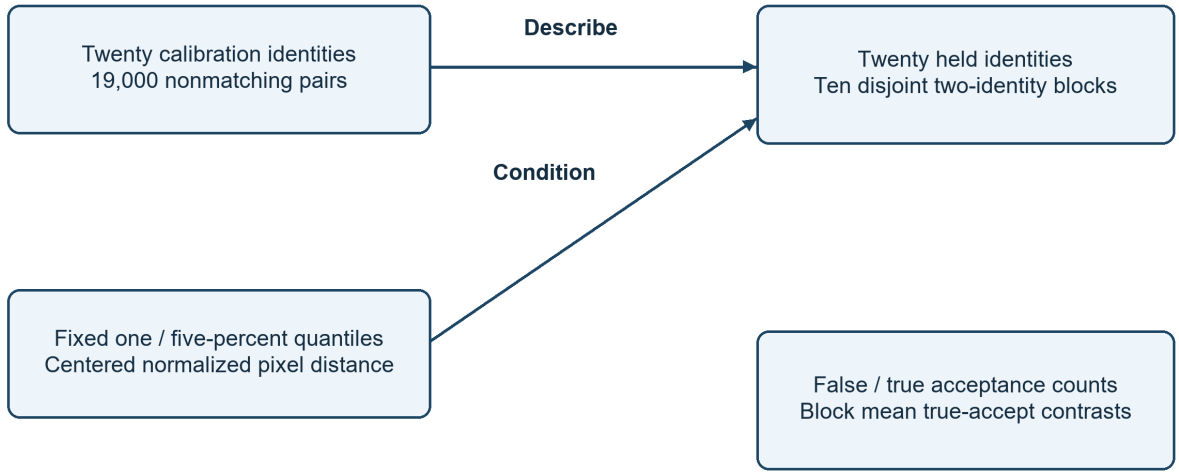

Figure 1. Separate identities supply calibration quantiles and ten held pair blocks; the thresholds remain fixed while false and true acceptance counts are recorded.

5. Counts first, with a boundary issue kept visible

At the stricter threshold, none of the 1,000 held nonmatches is accepted. At the more permissive threshold, ten are accepted, giving one percent. These observed rates are below the respective calibration targets. That outcome is not evidence that quantile calibration guarantees conservative transport. It describes one set of held identity pairings under one fixed split. Another population or pairing of identities could change the result.

All ten blocks have zero false accepts at the stricter threshold. An ordinary variance estimate across those zero block rates would therefore be zero, apart from numerical arithmetic. Reporting the resulting near-zero-width interval as a population-risk guarantee would be misleading. We retain the count descriptively and do not present that degenerate interval. Nor do we apply an independent-pair binomial interval to all one thousand repeated-image pairs as if they represented independent people.

The more permissive threshold accepts 372 matching pairs, compared with 250 at the stricter threshold. Their true-accept rates are 0.413333333 and 0.277777778. A low false-accept rate therefore does not establish useful recognition coverage in this baseline: many same-identity pairs remain unaccepted. An accepted pair means agreement with the released identity labels, rather than a reader's assessment of resemblance, identity preservation in a generated frame, or demographic fairness.

Metrics discussed in Fawcett, 2006, Saito and Rehmsmeier, 2015, and Matthews, 1975 distinguish other possible summaries. We report no full ROC curve, precision-recall curve, or correlation coefficient. Predictive-scoring principles in Gneiting and Raftery, 2007 likewise do not turn a distance threshold into a probability forecast. The retained counts and denominators make the two operating points inspectable without claiming a comprehensive assessment of this representation.

6. Block-level inference for true acceptance

Each block has ninety matching pairs, so the average block true-accept rate equals the pooled matching-pair rate. We fit intercept-only ordinary least-squares models for each threshold's block rate and for their within-block difference. The inferential sample size is ten blocks, not nine hundred matching pairs. We use the ordinary variance of the block mean and a t reference with nine degrees of freedom. These are approximate intervals conditional on the fixed calibration thresholds and split.

Table 1. True acceptance and its paired increase under the more permissive threshold; coefficients are proportions over matching pairs.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.2778 (0.0350)	[0.1986, 0.3569]	10
M1: fit	R-squared = 0.0000	-	-	10
M2	Mean outcome/contrast	0.4133 (0.0454)	[0.3107, 0.5160]	10
M2: fit	R-squared = 0.0000	-	-	10
M3	Mean outcome/contrast	0.1356 (0.0179)	[0.0951, 0.1760]	10
M3: fit	R-squared = 0.0000	-	-	10

Model labels: M1 = True acceptance at one-percent calibration threshold; M2 = True acceptance at five-percent calibration threshold; M3 = True-accept gain at more permissive threshold.

The stricter mean has standard error 0.034979908 and interval 0.198647728 to 0.356907827. The more permissive mean has standard error 0.045385107 and interval 0.310665088 to 0.516001579. Their paired gain is 0.135555556, standard error 0.017900809, with interval 0.095061113 to 0.176049999. This corresponds to 13.56 percentage points of additional true acceptance in the observed blocks.

The dependence reasoning in Cameron and Miller, 2015 explains why image pairs sharing identities cannot simply be counted as independent. The heteroskedasticity-consistent framework in White, 1980 would address a different variance issue. Our grouping design removes cross-block identity sharing but cannot remove shared historical acquisition or guarantee independent subject sampling. The ordinary block intervals remain approximations, and their R-squared values are zero because the models estimate means only.

Calibration-sample uncertainty is also excluded. The two thresholds are themselves estimated from a particular collection of identities. Repeating that split could move them and alter the held rates. The current intervals should therefore not be read as unconditional uncertainty for the whole calibration-and-evaluation procedure. A more extensive design could repeat identity-level splits and preserve the corresponding threshold variation, rather than treating one successful nominal calibration as exact knowledge.

7. Monotonic acceptance is not the empirical discovery

Because the permissive distance threshold is larger, it cannot reject a pair accepted by the stricter threshold. The direction of the acceptance change is mathematical. What the experiment measures is its size on these photographs, the accompanying nonmatch counts, and the mismatch between nominal and realized evaluation rates. We do not present the existence of a positive true-accept gain as a newly discovered law of face similarity.

Agreement concepts in Lin, 1989 and reliability considerations in Koo and Li, 2016 also distinguish repeated-score stability from verifying the intended identity construct. Neither statistic is estimated here. Measurement-error reasoning in Carroll et al., 2006 is relevant because pixel representation, centering, crop, and capture conditions affect the score. The experiment does not isolate which of those factors explains a missed match or establish that a particular threshold measures identity independently of them.

Influence concerns in Cook, 1977 could motivate examining blocks that contribute most to the tradeoff. We retain all ten. Multiplicity considerations in Benjamini and Hochberg, 1995 limit treating several operating points as an unrestricted threshold search. Interpretation under Wasserstein and Lazar, 2016 emphasizes the observed counts and the limited inferential units, rather than using a small tail probability for the paired gain to imply suitability for a real authentication or generation-evaluation system.

8. Identity transfer and scientific scope

The held identities differ from those used for calibration, which is a useful local separation. Validation issues in Roberts et al., 2017 and selection bias in Varma and Simon, 2006 still apply if the current outcomes are used to redesign the representation or choose a final operating point. Another evaluation would then be needed. This one split does not establish how thresholds transfer across devices, age changes, pose, lighting, video frames, or newly generated identities.

The source collection's restricted setting is also important. Photographs are historical, grayscale, and limited to forty released identities. The simple pixel transform has no trained semantic representation, and the true-accept rates are modest at the selected operating points. We do not attach the performance of a modern learned embedding to this baseline or use its results to recommend a production face-verification policy. The target generation framework remains outside the executed experiment.

Creator credit and source-specific access conditions remain part of the research record. The numeric archive is publicly accessible through a maintained mirror, but unrestricted photo redistribution has not been established by that availability alone. The analysis uses the released ordinal identities and reports aggregate quantities without displaying or inventing personal attributes. These provenance limits should be preserved in any later packaging of the raw data.

9. Report what a threshold retains as well as what it excludes

The completed comparison illustrates a threshold's dependence on its calibration and evaluation populations. Nominal calibration false-accept targets do not become guaranteed held-identity rates. The more permissive point retains more true matches and also more nonmatches, while the strict point's zero observed nonmatches does not establish zero future risk. Counts, sampling units, and the retained matching-pair fraction are necessary companions to a threshold label.

Documentation under Gebru et al., 2021 records the exact identity split, score transform, quantile convention, inclusive comparison, disjoint block construction, and omitted cross-block pair population. Preservation following Wilkinson et al., 2016 retains the distances and block outcomes alongside the source archive checksum. These records support reproduction of the stated operating-point comparison. A validated identity metric, generated-video assessment, or deployment threshold requires additional independent evidence and scientific review beyond this small historical-image experiment.

References

1. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.

2. Meng, Qiang, et al. "MagFace: A Universal Representation for Face Recognition and Quality Assessment." *arXiv preprint arXiv:2103.06627*, 2021. https://arxiv.org/abs/2103.06627.

3. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

4. Wang, Hao, et al. "CosFace: Large Margin Cosine Loss for Deep Face Recognition" *arXiv preprint arXiv:1801.09414*, 2018. https://arxiv.org/abs/1801.09414.

5. Turk, M.A.; Pentland, A.P.. (1991). Face recognition using eigenfaces. Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.1991.139758.

6. Huang, Gary B., Marwan Mattar, Tamara Berg, and Erik Learned-Miller. "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments." *ECCV Workshop on Faces in Real-Life Images*, 2008. https://cs.umass.edu/~elm/papers/Huang_eccv2008-lfw.pdf.

7. Fawcett, Tom. (2006). An introduction to ROC analysis. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2005.10.010.

8. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; ’t Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.

9. Saito, Takaya and Marc Rehmsmeier. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.” *PLOS ONE*, vol. 10, no. 3, 2015, pp. e0118432. https://doi.org/10.1371/journal.pone.0118432.

10. Schroff, Florian, et al. "FaceNet: A Unified Embedding for Face Recognition and Clustering" *arXiv preprint arXiv:1503.03832*, 2015. https://arxiv.org/abs/1503.03832.

11. Meng, Zijie; Liu, Jiwen; Liu, Yufei; Tong, Chengzhuo; Liu, Xiaoqiang; Zhang, Yuanxing; Xu, Yulong; Wan, Pengfei. (2026). ARGUS: Stacked Multi-View Identity Mosaic Injection for Subject-Preserving Video Generation. arXiv:2606.11670. https://arxiv.org/abs/2606.11670.

12. Deng, Jiankang, et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition" *arXiv preprint arXiv:1801.07698*, 2018. https://arxiv.org/abs/1801.07698.

13. Kim, Minchul, et al. "AdaFace: Quality Adaptive Margin for Face Recognition." *arXiv preprint arXiv:2204.00964*, 2022. https://arxiv.org/abs/2204.00964.

14. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

15. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.

16. Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

17. Elkan, Charles. “The Foundations of Cost-Sensitive Learning.” *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973-978. https://cseweb.ucsd.edu/~elkan/rescale.pdf.

18. Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

19. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.

20. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.

21. AT&T Laboratories Cambridge. (2026). Modified Olivetti faces: Sam Roweis MATLAB version [Data archive; mirror accessed 7 November 2026]. File identity and row-label layout verified against the scikit-learn maintainer loader. https://ndownloader.figshare.com/files/5976027.

22. Samaria, F.S.; Harter, A.C.. (1994). Parameterisation of a stochastic model for human face identification. Proceedings of 1994 IEEE Workshop on Applications of Computer Vision. https://doi.org/10.1109/acv.1994.341300.

23. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

24. Matthews, B.W.. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. https://doi.org/10.1016/0005-2795(75)90109-9.

25. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

26. Belhumeur, P.N.; Hespanha, J.P.; Kriegman, D.J.. (1997). Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/34.598228.

27. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.

28. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

29. Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

30. Geifman, Yonatan, and Ran El-Yaniv. “Selective Classification for Deep Neural Networks.” *Advances in Neural Information Processing Systems*, vol. 30, 2017. https://arxiv.org/abs/1705.08500.